

PROJECT 1: COMBINING ENVIRONMENTAL DATA SETS

1.D. ABSTRACT

1.1 Sorting Code: 2001–STAR–D1; responding to Statistical Research Area 2

1.2 Title: Project 1: Combining Environmental Data Sets

1.3 Investigators: Jennifer A. Hoeting (Principal Investigator, CSU), Richard A. Davis (CSU), F. Jay Breidt (CSU), Robin Reich (CSU), Don L. Stevens, Jr (OSU), and Steven B. Weisberg (SCCWRP).

1.4 Institution: Colorado State University, Oregon State University, and Southern California Coastal Water Research Project (SCCWRP).

1.5 Project Period: October 1, 2001–September 30, 2005

1.6 Project Cost: \$252,731 in year one; \$971,177 total

1.7 Overall Summary

Objectives: The objectives of this research are to develop approaches for spatio-temporal design and modeling in order to further our understanding of aquatic resources. We will extend current design and analysis methodology in order to incorporate data from different scales and different sources including data from EMAP’s probability based sampling design, intensive site studies, and remote sensing platforms. Our research will focus on three major objectives: development of spatio-temporal models for a continuous response, development of spatio-temporal models for count and/or categorical data, and development of design and analysis methods for data collected at different scales. In meeting these objectives we will focus on specific EMAP needs.

The first objective of the proposed research is to develop spatio-temporal models for a continuous response to model EMAP data. Considerable work has been done in the area of modeling spatially correlated data, but many problems still need to be addressed. In particular, we propose to extend Bayesian hierarchical models to provide better predictions for spatially and possibly temporally correlated data. We also propose to address the issue of providing accurate assessments of uncertainty about these predictions via methodology called Bayesian model averaging. Finally, we propose to develop better methodology for model selection for these models. All of these new developments should lead to more accurate predictions and improved maps of aquatic based resources.

In the EMAP context, one may wish to model species counts, species presence/absence, or contaminant levels above or below a regulatory threshold. The second objective of the proposed research addresses these types of problems. The goal here is to extend and develop spatio-temporal models for count and/or categorical responses. The proposed methodology will extend existing methodology based on a model for binary response data called the autologistic model. We also propose to develop new methods to analyze

spatially and temporally correlated discrete response data including development of a spatio-temporal autologistic model and models for latent processes. These models can be used to produce maps such as a map showing probability that a threshold for a particular contaminant has been exceeded over an area of interest. The methods also can be used to assess which indicators are relevant for predicting whether or not the threshold level has been exceeded.

The final objective of the proposed research is to address an on-going challenge in the modeling of EMAP and other aquatic resource monitoring data: how do we combine data at different scales? The proposed research will develop new guidelines for designing ecological monitoring studies aimed at understanding and assessing key aspects of aquatic systems. In addition to improving design of such studies, we also propose to develop new methodology to analyze data collected at different scales. The goal here will be to both understand and assess key processes in aquatic systems as well as to be able to produce estimates for all scales of interest.

Expected Results: The main emphasis of this program is to develop cutting-edge statistical methods to address environmental problems, with specific emphasis on EMAP problems. Expected results include better interpolated maps of aquatic resources via Bayesian kriging, more honest assessment of predictive uncertainty via Bayesian model averaging, improved maps showing probability of presence/absence via the autologistic model, and guidelines and analysis methods for multi-tiered monitoring studies with the goal of identifying key scales of inquiry. All methodology will be developed and tested using EMAP data and EMAP-designed studies, so that the results will be immediately useful in practice.

A second benefit of this program will be the training of new statisticians in the area of environmental statistics. There is currently a severe shortage of statisticians with extensive knowledge and experience in solving environmental problems and analyzing environmental data. One goal of this project is to increase the number of statisticians who meet such a need. Graduate students and post-doctoral researchers will participate in all research and dissemination activities proposed in this grant, gaining valuable experience in environmental statistical problems. Our project will thus integrate research and education and partly address the chronic shortage of statisticians with expertise in environmental statistics.

1.8 Supplemental Keywords: Latent processes, Matern covariance function, model selection, remote sensing, sampling design.

1.E. OVERALL DESCRIPTION

1.1 *OBJECTIVES*

Objectives of the proposed research

The objectives of this research are to develop approaches for spatio-temporal design and modeling in order to further our understanding of aquatic resources. We will extend current design and analysis methodology in order to incorporate data from different scales and different sources including data from EMAP's probability based sampling design, intensive site studies, and remote sensing platforms. Our research will focus on three major objectives: development of spatio-temporal models for a continuous response, development of spatio-temporal models for count and/or categorical data, and development of design and analysis methods for data collected at different scales. In meeting these objectives we will focus on specific EMAP needs.

The first objective of the proposed research is to develop spatio-temporal models for a continuous response to model EMAP data. Considerable work has been done in the area of modeling spatially correlated data, but many problems still need to be addressed. In particular, we propose to extend Bayesian hierarchical models to provide better predictions for spatially and possibly temporally correlated data. We also propose to address the issue of providing accurate assessments of uncertainty about these predictions via methodology called Bayesian model averaging. Finally, we propose to develop better methodology for model selection for these models. All of these new developments should lead to more accurate predictions and improved maps of aquatic based resources.

In the EMAP context, one may wish to model species counts, species presence/absence, or contaminant levels above or below a regulatory threshold. The second objective of the proposed research addresses these types of problems. The goal here is to extend and develop spatio-temporal models for count and/or categorical responses. The proposed methodology will extend existing methodology based on a model for binary response data called the autologistic model. We also propose to develop new methods to analyze spatially and temporally correlated discrete response data including development of a spatio-temporal autologistic model and models for latent processes. These models can be used to produce maps such as a map showing probability that a threshold for a particular contaminant has been exceeded over an area of interest. The methods also can be used to assess which indicators are relevant for predicting whether or not the threshold level has been exceeded.

The final objective of the proposed research is to address an on-going challenge in the modeling of EMAP and other aquatic resource monitoring data: how do we combine data at different scales? The proposed research will develop new guidelines for designing ecological monitoring studies aimed at understanding and assessing key aspects of aquatic

systems. In addition to improving design of such studies, we also propose to develop new methodology to analyze data collected at different scales. The goal here will be to both understand and assess key processes in aquatic systems as well as to be able to produce estimates for all scales of interest.

The specific objectives are as follows:

Objective 1.1: Develop spatio-temporal models for a continuous response to model EMAP and related data.

1.1.1: Extend Bayesian methods for spatio-temporal prediction.

1.1.2: Extend Bayesian model averaging methodology for spatial prediction.

1.1.3: Improve model selection methodology for spatial models.

Objective 1.2: Extend spatio-temporal models for count and/or categorical-valued data to model EMAP and related data.

1.2.1: Extend autologistic model to account for multivariate response.

1.2.2: Develop a spatio-temporal autologistic model.

1.2.3: Apply autologistic model to maximize detection in EMAP sampling design.

1.2.4: Extend and apply models with latent processes.

Objective 1.3: Develop design and analysis methods for data collected at different scales.

1.3.1: Develop guidelines for designing multi-tiered monitoring studies with the goal of identifying key scales of inquiry.

1.3.2: Develop methodology for analyzing data collected at different scales.

Details on the approach used in pursuing these objectives are provided below.

Importance of the proposed research

The underlying goal of the proposed research is to provide a framework for improving our understanding of spatial and temporal structures of aquatic systems. The proposed methods will address problems that are encountered in the analysis of data on aquatic systems such as honest accounting for uncertainty in predictions for spatially correlated data and methodology for analyzing data collected at different scales. All methodology will be developed and tested using EMAP data and EMAP-designed studies, so that the results will be immediately useful in practice.

1.2 APPROACH

Objective 1.1: Develop spatio-temporal models for a continuous response to model EMAP and related data

Much of the data collected to monitor and further understand aquatic systems involves a continuous response such as contamination levels of fish tissue or water chemistry. We provide here a brief overview of standard spatial models in the context of EMAP applications and suggest areas for future research.

Review of existing methodology

Consider the class of linear models for spatially-explicit prediction of a continuous response. Let $\mathbf{Z} = (Z(s_1), \dots, Z(s_n))'$ be a partial realization of a random field $\mathbf{Z}(s)$, where $s \in D$ and D is a fixed finite area under study, such as an EMAP study area. Assume that p explanatory variables $\mathbf{X}(s) = (X_1(s), \dots, X_p(s))'$ were also observed at each location. A model for the random field at any location s is given by

$$Z(s) = \sum_{j=0}^p X_j(s)\beta_j + \delta(s) = \mathbf{X}'(s)\boldsymbol{\beta} + \delta(s), \quad (1)$$

where $\boldsymbol{\beta}$ is a vector of unknown coefficients and $\delta(s)$ is the unobserved error at location s .

It is typically assumed that the error is a weakly stationary, isotropic Gaussian process with mean zero and covariance function $\text{Cov}(Z(s), Z(t)) = \sigma^2\rho(|s - t|)$, where σ^2 is the variance of the process and $\rho(\cdot)$ is the autocorrelation function. Weak stationarity means that the mean of the spatial process is location invariant and the autocorrelation between two values of the spatial process depends only on the distance and direction between locations. A weakly stationary process is also isotropic if the autocorrelation depends only on the Euclidean distance between two sample locations.

There are several widely used parametric forms for isotropic autocorrelation functions such as the exponential, Gaussian, and spherical autocorrelation functions (Haining, 1990). While these autocorrelation functions may be useful in a variety of applications, they are somewhat limited in the types of behavior that they can be used to model. The Matern autocorrelation (e.g., Stein 1999) function is a flexible alternative to standard autocorrelation functions. In fact, both the exponential and Gaussian autocorrelation functions are members of the Matern class of functions. The Matern autocorrelation function is featured in Stein's monograph on kriging (Stein, 1999) and has been used successfully in a variety of applications (Handcock and Stein, 1993; Handcock and Wallis, 1994; Ecker and Gelfand, 1997; Ecker and Gelfand, 1999; Thompson, 2001).

For non-isotropic, or anisotropic, processes, parameter estimation becomes more complex. The simplest approaches to modeling anisotropic processes involve transforming a non-isotropic process so that the resulting process is isotropic (see, for example, Smith 2000). A more complex, and arguably, more realistic alternative is a deformation approach

proposed by Sampson and Guttorp (1992). Here it is assumed that for some nonlinear function $g(s)$, the process $Z(g(s))$ is a stationary isotropic process. This approach applies to non-stationary as well as anisotropic processes.

One goal in the analysis of spatially correlated data may be to predict the response at unobserved locations, such as the status of the coho salmon population off of the Oregon coast. Kriging is a standard approach to spatial prediction with the goal of minimizing the mean square prediction error (Cressie, 1993). In the estimation of the kriging prediction variance it is standard practice to use estimates of the parameters in the autocorrelation function $\rho(\cdot)$. One drawback with this approach is that it ignores the uncertainty in estimation of the autocorrelation function and thus prediction intervals may be optimistically narrow. In the prediction of contaminant levels, for example, predictions intervals that are too narrow could lead to inappropriate policy decisions. Bayesian approaches to prediction offer one solution to this problem.

Objective 1.1.1: Extend Bayesian methods for spatio-temporal prediction

Bayesian methods offer promising opportunities to understand the complex spatial and temporal structures of aquatic systems. As computational methods for Bayesian techniques are being developed, Bayesian methods are being considered by a ever increasing number of authors. For a recent overview of Monte Carlo computational techniques for Bayesian inference, see Robert and Casella, 1999.

Kitanidis (1986) and Le and Zidek were early proponents of Bayesian methodology in the context of spatial prediction using kriging. Handcock and Stein (1993) adopt the Matern covariance function and propose a model for “Bayesian kriging.” Berger *et al.* (2000) focus on objective Bayesian analysis of Gaussian random fields. They show that common choices of prior distributions to model Gaussian random fields typically produce improper posterior distributions. In her recent Ph.D. thesis (advised by project principal investigators Hoeting and Davis), Thompson (2001) adopts proper prior distributions to further develop models for Bayesian kriging and shows that the resulting predictions with the Matern class of autocorrelation function leads to improved predictive performance as compared to traditional methods of spatial predictions using standard autocorrelation functions.

Opportunities for extending the work on Bayesian kriging abound. Extensions to apply Bayesian kriging methods to large data sets would be useful in the context of some EMAP studies where a large number of observations are collected for some sites. It is also important to consider measurement error in the analysis of EMAP data for spatial prediction. Accounting for missing data in the explanatory variables may be important as measurements for some indicators may be missing at some sites. Finally, accounting for temporal as well as spatial correlation between observations may provide new understanding of environmental processes.

Objective 1.1.2: Extend Bayesian model averaging methodology for spatial prediction

Standard statistical practice ignores the uncertainty in selecting models. Model selection typically involves selecting a class of models, selecting a single model from this class, and then estimating parameters and making predictions as if the selected model were the model that generated the data. As has been shown by Regal and Hook (1991), Draper (1995), Madigan and York (1995), Kass and Raftery (1995), and Raftery (1996), this approach can lead to underestimation of the uncertainty in the inferences. This could be a severe problem in the modeling of environmental monitoring data. Often regulatory thresholds for environmental samples are conveyed in terms of upper confidence limits, or performance goals are specified using the standard error. Underestimation of uncertainty could be under-protective of the environment.

The approach commonly referred to as Bayesian model averaging (BMA) provides one solution to this problem. Let M_1, \dots, M_K be the set of models under consideration. If Δ is the quantity of interest, such the predicted level of contaminant or the utility of an environmental clean-up effort, then the posterior distribution of Δ given data D is:

$$\text{pr}(\Delta | D) = \sum_{k=1}^K \text{pr}(\Delta | M_k, D)\text{pr}(M_k | D). \quad (2)$$

This is an average of the posterior distribution under each of the models considered, weighted by the corresponding posterior model probability. The posterior probability for model M_k is given by

$$\text{pr}(M_k | D) = \frac{\text{pr}(D | M_k)\text{pr}(M_k)}{\sum_{l=1}^K \text{pr}(D | M_l)\text{pr}(M_l)}, \quad (3)$$

where

$$\text{pr}(D | M_k) = \int \text{pr}(D | \theta_k, M_k)\text{pr}(\theta_k | M_k)d\theta_k \quad (4)$$

is the integrated likelihood of model M_k , θ_k is the vector of parameters of model M_k (e.g., for regression models $\theta = (\beta, \sigma^2)$), $\text{pr}(\theta_k | M_k)$ is the prior density of θ_k under model M_k , $\text{pr}(D | \theta_k, M_k)$ is the likelihood, and $\text{pr}(M_k)$ is the prior probability that M_k is the true model. All probabilities are implicitly conditional on \mathcal{M} , the set of all models being considered.

There are many challenges involved in the implementation of BMA, including the computation of (2) for a very large number of models, the evaluation of the integrals implicit in (2) which do not typically exist in closed form, and the specification of the prior model probabilities $\text{pr}(M_k)$. Hoeting *et al.* (2000) provide an overview of the currently available methodology to implement BMA for linear regression models, generalized linear models, survival analysis, and graphical models.

Thompson (2001) extended BMA methods to approximate (2) for the class of spatial linear models. Thompson showed that BMA over all possible subsets of explanatory variables leads to more accurate predictive coverage and smaller predictive errors as compared

to the standard practice of basing predictions on a single model. BMA is a promising approach to the analysis of EMAP data, since a whole suite of indicators are measured and the relationships among environmental variables are often not well understood.

There are many potential extensions to BMA for the class of spatial linear models. Averaging over a broader class of autocorrelation functions or averaging over all possible transformations of the response and the predictors may lead to additional improvements in predictive performance. In addition, computational methodology to apply BMA to data sets with a large number of observations and/or a large number of possible models will need to be developed to be fully applicable for a broader class of spatial analysis.

Objective 1.1.3: Improve model selection methodology for spatial models

For EMAP data, a key component is to investigate relationships among variables. Therefore, it is important that the selection of explanatory variables in a spatial model is done with considerable care and rigor.

For spatially correlated data, standard practice for model selection is to first choose a set of explanatory variables assuming independent errors and then to model the residuals using a standard autocorrelation function such as the exponential, Gaussian, or spherical autocorrelation function. Thompson (2001) shows via simulation studies that this approach can lead to the selection of explanatory variables that did not generate the data and thus to incorrect inferences about the explanatory variables. Thompson proposes an alternative methodology for choosing the explanatory variables in the context of a spatial model. She demonstrates that her approach of simultaneously selecting explanatory variables and the form of the autocorrelation function improves the mean square prediction error and predictive coverage as compared to standard methods for model selection for spatially correlated data. Another important finding of this work is that this approach correctly identifies the “true” explanatory variables much more readily than the standard methods. For EMAP data, the simultaneous approach may lead to the selection of key explanatory variables to aid in predicting a response of interest. This may lead to identifying impacted areas, assessing whether all indicators need to be measured, and furthering the understanding of the relationships between indicators such as sediment chemistry, sediment toxicity, and benthic health indicators.

In addition to using Thompson’s (2001) methodology to select predictors for EMAP applications, we propose several extensions of this work. Thompson’s methodology is based on Akaike’s Information Corrected Criteria (AICC) (Hurvich and Tsai, 1989). We propose to compare results obtained using AICC to results obtained using more modern techniques such as Rissanen’s two-part minimum description length (MDL) principle (Lee, 2001). MDL may be particularly useful for the large data sets possible with EMAP. It may also be fruitful to investigate the performance of the Thompson’s simultaneous approach using other autocorrelation functions and other spatial models. Specific EMAP applications may suggest other fruitful extensions to this work.

Objective 1.2: Extend spatio-temporal models for count and/or categorical-valued data to model EMAP and related data

For the prediction and modeling of species counts, species presence/absence or contaminant levels above or below a regulatory threshold, we must consider models for a discrete response. For data collected on a grid, such the hexagon structure of the EMAP design, we might consider the class of autologistic models for binary data or auto-Poisson models for count data (Cressie, 1993). We consider here extensions to the autologistic model. Similar approaches are also possible for count data modeled using the auto-Poisson model.

Review of existing methodology

For a lattice of spatially correlated responses, the basic autologistic model (Besag, 1972) predicts the response at site i using a logistic function on the responses for the neighboring sites. Applications of this type of model have been extended to agricultural research (Gumpertz *et al.*, 1997), forestry (Preisler, 1993), archaeology (Besag *et al.*, 1991), and biological range mapping (Heikkinen and Högmänder, 1994; Högmänder and Møller, 1995). For many applications, however, there is more information available than is utilized by the basic autologistic model, such as covariates that are related to the response of interest. In addition, there is often only partial information available about the response of interest. This is the case when a sample of sites is collected over an area of interest, such as in EMAP applications.

Hoeting *et al.* (2000) and Leecaster (1999) develop methodology to estimate parameters in the autologistic model with covariates. This model allows the user to fully utilize data from a sparsely sampled area of interest for prediction and inference over the entire area. To improve predictions, the model incorporates other covariates measured over the area of interest. The autologistic model with covariates is defined

$$\text{pr}(z_i = 1 \mid \underline{x}_i, \underline{z}_{-i}, \underline{\theta}, \beta) = \frac{\exp\{\underline{x}_i^T \underline{\theta} + \beta s_i\}}{1 + \exp\{\underline{x}_i^T \underline{\theta} + \beta s_i\}}, \quad (5)$$

where \underline{x}_i is a vector of covariates for the i th site with the first element equal to 1, $\underline{\theta}$ is a vector of parameters for the covariates, and β is the parameter associated with the spatial covariate, s_i . The spatial covariate for site i , $s(z_i)$, is equal to the total number of sites where the species was present in the neighborhood of site i . A neighborhood might be defined as first order, which is the set of pixels directly north, south, east and west of the pixel of interest or second order, which also includes the sites diagonal from the site of interest.

Data from EMAP designed studies is ideal for applying the autologistic model with covariates for sample data due to the grid structure and the sampling plan. The hexagon structure of the lattice allows for a more appealing neighborhood structure since all neighbors have a common edge (unlike the second order square lattice where some neighbors share only a corner), and the area is completely surrounded by neighbors (unlike the first

order square lattice, where corners are missed). Often, due to subsampling or stratification, there are samples collected in only a subset of the hexagons, but prediction is often required over the entire area.

The likelihood function for the autologistic model is analytically intractable, except in trivial cases, so alternative estimation methods are necessary. Hoeting *et al.* (2000) adopt a Bayesian approach to estimate the parameters of the model. Their methodology produces parameter estimates as well as maps showing the posterior probability of presence over the area of interest. They showed that the inclusion of covariates in an autologistic model improved predictions as compared to the standard logistic and autologistic models. Software to implement the methodology of Hoeting *et al.* (2000) is available at no cost via the internet. It is anticipated that any extensions to this methodology will also be made available via similar means.

Objective 1.2.1: Extend autologistic model to account for multivariate response

Just as the inclusion of covariates in an autologistic model improved predictions, it is reasonable to assume that using two or more related responses, such as species, in a multivariate response autologistic model will likely improve prediction accuracy for both species. The resulting prediction map would indicate areas of probable presence for each species and a combined map for all of the species that are considered. For example, since many benthic species have similar habitat requirements, this approach could improve prediction for seldom-seen species. Another example would be applying the multivariate autologistic model to benthic species and sediment contamination to improve prediction, since different species respond differently to specific levels of various contaminants.

Objective 1.2.2: Develop a spatio-temporal autologistic model

Once an autologistic model has been used to model the spatial interaction between sites, it is natural to extend the model to account for temporal as well as spatial correlation between observations. For example, to monitor rare (or even abundant) species, it is important to find out how the species distribution is changing over time. To monitor fish tissue contaminants, one may be interested in the change in areal extent of the problem. To address these types of questions, the autologistic model needs to be extended to account for correlation between observations that are observed over time as well as space.

There are several recent approaches to model space-time data using a hierarchical Bayesian setup (Wikle, 1996; Royle and Berliner, 1997; Royle *et al.*, 1998; Wikle *et al.*, 1998). Initial investigation into these methods appears quite promising for the autologistic model with covariates. A key component in these methodologies is to appropriately modify the locally dependent Markov random field to allow for estimation in the context of space-time correlation. These authors have used several simplifications in the assumptions to model space-time correlation for Gaussian data. This approach may offer one way to

extend the autologistic model to account for correlation between binary data that are observed over time as well as space.

Objective 1.2.3: Apply autologistic model to maximize detection in EMAP sampling design

Leecaster (1999) uses maps of predicted probability of presence produced using the autologistic model to develop robust sampling designs with the goal of maximizing detection of the sampling unit. If detection is the goal, this work could be extended to incorporate the EMAP sampling design. Predictions of impacted areas could be also be used to define subpopulations for future sampling efforts. The definition of stratum areas is often a very difficult part of the design process and the predicted probabilities of presence for various indicators would be an invaluable resource. This same information could be used outside of EMAP designs to aid researchers monitoring point sources in creating a “footprint of impact” and in designing regulatory monitoring programs. These predictions could also be used to determine hotspots of contamination which may require further investigation.

Objective 1.2.4: Extend and apply models with latent processes

Generalized state models offer a flexible and useful framework for modeling dependent non-Gaussian data such as binary or count data that may arise in a spatial-temporal setting. A generalized state-space model consists of two equations referred to as the observation and state equations. Such models can be loosely characterized as either *observation-driven* or *parameter-driven* (see Section 8.6 of Brockwell and Davis, 1996, and Davis *et al.*, 1999). The observation equation specifies the distribution of the observation given a state variable. Equation (5) of the autologistic model is an example of an observation equation where x_i is the *observation* variable and s_i is the corresponding *state* variable. For an observation-driven model, the state-variables are explicit functions of other observation variables. The autologistic model of (5) is example of an observation-driven model since the state s_i is a function of the neighboring observations at the i^{th} site. On the other hand, the state-equation for a parameter-driven model specifies a stochastic model for the state-variables that does not depend on the observations. In this case, the state-variables are often referred to as a *hidden* or *latent generating process* of the model.

To illustrate the latent process model, consider the binary response variable x_i at site i whose distribution conditional on a state-process s_i and a set of explanatory variables z_i , is given by

$$p_i = \text{pr}(z_i = 1 \mid \underline{x}_i, s_i, \underline{\theta}, \beta) = \frac{\exp\{\underline{x}_i^T \underline{\theta} + s_i\}}{1 + \exp\{\underline{x}_i^T \underline{\theta} + s_i\}} \quad (6)$$

(cf. 5). To complete the specification, a probabilistic model is required for s_i . A common choice is to assume that $\{s_i\}$ is either a Markov random field or a stationary isotropic Gaussian random field. The intuition behind the latent process model, is that the logistic

function of the mean follows the general linear spatial model (1). That is,

$$\ln\left(\frac{p_i}{1-p_i}\right) = \underline{x}_i^T \underline{\theta} + s_i. \quad (7)$$

There are a number of modeling advantages in this formulation of the model over an observation-driven specification such as the one described in Objective 2.

1. *Ease of interpretation.* There is a close analogy between this model and the general linear spatial model (1). In this formulation of the model, the scientist can directly interpret, on the logistic scale, the impact of each explanatory variable. This is generally not true for observation-driven models, since the state-variable s_i involves other values of the observed process which in turn depend on the explanatory variables.
2. *Flexibility.* The noise process is capable of modeling small-scale variability and some large-scale variability that may be unaccounted for by the regression term. That is, the noise can absorb part of the large-scale information that is not contained in the available explanatory variables.
3. *Fundamental properties of model are easier to establish.* The process given by (6) and (7) is consistently specified and, assuming the regression function is constant, properties such as stationarity and ergodicity of the process are inherited directly from those imposed on the latent process $\{s_i\}$. Such behavior is essential for carrying out consistent estimation procedures. For observation-driven models, stability problems related to stationarity and ergodicity are often difficult to establish.
4. *Easily adapted to non-lattice spatial data and spatial-temporal data.* The latent-process formulated model is readily extendable to the case of non-lattice spatial or to spatial-temporal data by taking $\{s_i\}$ to be a random field on the plane or a spatial-temporal Gaussian process. In the latter, one can entertain latent process models such as those proposed by Niu and Tiao (1995) and Stoffer (1986) for the case when the spatial component is observed over a lattice and by Wikle *et al.* (1998) for a spatial-temporal hierarchical model.

One of the primary research goals of this section will be the implementation of these models to EMAP data. Estimation of both the parameters of the regression function and those of the latent process presents one of the major challenges in the use of latent process models. In the time series setting, there have been a number of approaches for estimation. These include the use of estimating equations as in Zeger (1988), the Monte Carlo EM algorithm by Chan and Ledolter (1995), the Monte Carlo Newton-Raphson algorithm by Kuk and Cheng (1997), MC maximum likelihood estimation by Durbin and Koopman (1997; 2000), and GLIM type estimation by Davis *et al.* (2000). Unlike the time series case, the lack of a directionality in space does not allow for a convenient definition of

an *innovation*, which, in turn complicates the adaptation and implementation of these estimation algorithms to the spatial setting.

In addition to estimation and prediction for the latent-process model, the entire modeling paradigm will be explored. This will include the development of diagnostic tools for model identification of a class of models for both the explanatory variables and the latent process itself, for assessment of model adequacy, and for detection of outliers, etc. It is anticipated that much of the research described for Objective 1 will provide the building blocks for extension to the latent process model formulation.

Objective 1.3: Develop design and analysis methods for data collected at different scales

The monitoring of coastal waters is one example of the importance of developing design and analysis methods for data collected at different scales. Coastal monitoring is conducted on various spatial scales to address different goals. As a framework for investigation we might concentrate on three scales of interest. On a national scale, the goal is assessment of large regions. An example is EPA's endeavor to achieve national assessment through implementation of Coastal 2000. On a regional scale, there is the need to assess a large region, but also to provide a comparison of conditions among subpopulations such as coastal urban runoff zones and offshore wastewater outfalls. On a local sampling scale, the goal is to monitor the effects of point-source discharges, where there are needs to map gradients away from an individual facility's discharge and to assess trends at sites along these gradients. There is considerable work to be done on developing design guidelines and developing statistical methods for data aggregation in the context of aquatic ecosystems.

Objective 1.3.1: Develop guidelines for designing multi-tiered monitoring studies with the goal of identifying key scales of inquiry

We will consider the problem of developing guidelines for designing multi-tiered monitoring studies with the goal of identifying key scales of inquiry for classes of aquatic ecosystems such as streams, rivers, wetlands, lakes, or coastal systems. We propose to investigate ways to design multi-level surveys that can achieve the specific goals for the data collected at each scale while allowing for the data to be integrated to maximize understanding of the underlying aquatic ecosystem. Ideally, if the surveys are designed with the goal of making inferences across different scales, understanding and precision will increase and costs will decrease without unreasonable added effort.

Combination of data from different levels will require coordination of designs (including sample site selection, commonly measured indicators, and comparable methodology for sampling, analysis, and data management) and development of estimation procedures, discussed below. Although there is a substantial amount of planning and work to be done to coordinate indicators, sampling and analysis techniques, and data management systems, we concentrate here on the sample site selection and estimation procedures.

Coordination of sample site selection will involve two approaches; one to combine two or more EMAP surveys and one to combine EMAP and non-EMAP surveys. Generally national and regional surveys are EMAP designs while local survey sites are subjectively chosen. We propose that national and regional scale surveys could be combined, so that the national sample is a subsample of the regional survey. This would allow sharing of costs as well as data for those two scales. For cases when the timing of surveys does not allow this, we will investigate effects of replacing data points from one design to the other. The coordination of designs for local programs (non-EMAP) with regional and national (EMAP) will require further investigation and analysis. The key will be to design a local survey that is a probability sample so that combination with EMAP data will be more equitable. It may be possible to retain a nearly systematic grid while utilizing an EMAP design, or to randomly select a systematic grid.

Another area of concern is the determination of the scale on which to assess the landscape metrics. Schuft *et al.* (1999) investigated the connection between landscape metrics and water quality for a random sample of stream segments in the Pudding River basin in the northern Willamette Valley, Oregon. They used a plot design whereby the landscape metrics were evaluated in a series of expanding buffers around stream segments. The authors then examined a variety of statistical measures to estimate the buffer size that for which the landscape metrics showed the strongest relationship with measures of stream quality.

This particular study illustrates two key components that a design for assessing key scales should address: (1) a plot design that permits a variety of aggregation scales; and (2) a sampling design that distributes the plots over the target population. The specifics of the implementation of these two components almost surely depend on the indicators being evaluated. For example, Schuft *et al.* (1999) used a plot design with buffers increasing isotropically. It may be that aggregation areas defined by elevation contours would be more suitable for some indicators.

Two of the tiers of a multi-tiered study will normally consist of spatially continuous data and point measurements at a number of sites. If we regard the spatially continuous data as defining a function $f(s)$ defined for every point s in some spatial domain D . If we have point data at s_1, s_2, \dots, s_n , our problem is to determine a kernel function $h(\cdot)$ so that we can maximize the correspondence between the indicators measured at the s_i , and the aggregated continuous data,

$$a_i = \int_{H(s_i)} h(s)f(s)ds,$$

where $H(s_i)$ is some neighborhood of s_i . One approach to the issue would be to maximize the correlation between the point-supported condition indicators and the aggregate values a_i , for a suitable restricted class of kernel functions, e.g., bivariate Gaussian. The particular setting of the problem may suggest a more appropriate class of kernel functions.

Objective 1.3.2: Develop methodology for analyzing data collected at different scales

There are various approaches to analyzing data from different scales. Data from different scales are often considered at each scale and then the most appealing one chosen. This technique does not use all of the data since it treats the scales separately instead of combining them. Overton (1990) and Overton *et al.* (1993) combine EMAP data with subsequently ‘found’ sample data. They develop an approach to determine inclusion probabilities for the found data based on their similarity with EMAP data to provide estimates from the combined data. They warn that the effort involved in combining the information using their approach may be more than what the results are worth. Considering multiple frames (Sarndal *et al.*, 1992) may be a possibility for combining data, especially if the surveys measure some different indicators.

There has also been some research into using Bayesian methods for dealing with different scales. For example, Mugglin *et al.* (2000) develop a Bayesian hierarchical model for spatially misaligned data, or data collected on different scales. In the context of an environmental risk analysis of inhalant exposure to radon, they propose a Bayesian hierarchical model which provides posterior distributions of imputed counts at any level of aggregation. Similar methodology might be used in the EMAP context to impute, say, species counts or, in the continuous context, contamination levels.

1.3 EXPECTED RESULTS OR BENEFITS

The main emphasis of this program is to develop cutting-edge statistical methods to address environmental problems, with specific emphasis on EMAP problems. The main outcome of this research will be new methodology developed to meet the objectives outlined above. All methodology will be developed and tested using EMAP data and EMAP-designed studies, so that the results will be immediately useful in practice.

A second benefit of this program will be the training of new statisticians in the area of environmental statistics. There is currently a severe shortage of statisticians with extensive knowledge and experience in solving environmental problems and analyzing environmental data. One goal of this project is to increase the number of statisticians who meet such a need. Graduate students and post-doctoral researchers will participate in all research and dissemination activities proposed in this grant, gaining valuable experience in environmental statistical problems. Our project will thus integrate research and education and partly address the chronic shortage of statisticians with expertise in environmental statistics.

Potential users at many levels will benefit from this research. Statistical researchers will benefit as new methods will be developed for and applied to spatially and temporally correlated data. A wide range of researchers of all statistical abilities will be able to apply the new methodology because it will be made available in well documented and carefully tested software. Finally, policy makers and other interested parties will be able to use the results produced using the new methodology that is developed. In particular, the new

methodology will allow for better understanding and assessment of the processes that underly our aquatic resources.

While the focus of the program is to develop methodology to address EMAP objectives, the new methods will be applicable to a wide range of other problems in environmental statistics. For example, the proposed research will provide better maps to predict probability of presence/absence for rare species of wildlife, will allow for improved assessment of forest cover, will provide more honest assessment of predictive uncertainty for air quality monitoring, and will provide new ways to assess and predict disease incidence.

1.4 MANAGEMENT PLAN AND MILESTONES

Hoeting will oversee the project as well as work on specific objectives. For design and methodology development, Hoeting will focus on objectives 1.1 and 1.2.1-1.2.3; Davis will work on objectives 1.1 and 1.2.4; Stevens and Weisberg will focus on objective 1.3. The general aim will be to work on methodology development in years 1 and 2 and then to apply the methodology to EMAP and related data sets in years 3 and 4.

Year 1: In year 1 all project participants will work to refine grant objectives to address specific EMAP problems. Key participants in this process will be Urquhart, Stevens, and Weisberg, who have considerable background in the design and analysis of EMAP related data. Specific data sets will be identified for each of the objectives in the grant proposal. Reich will contribute to the project here via his extensive knowledge of applying statistical techniques to address ecological problems. Davis and Hoeting will oversee the project as well as recruit graduate student(s) and post-doctoral researcher(s) to work on the project.

Year 2: In year 2 the main thrust will be on methodology development. All post-docs will be on board and graduate students will be “up to speed” on the project.

Year 3: The focus of the project will shift in year 3 to the application of the newly developed methodology to address EMAP problems and analyze EMAP data. Methodology development will continue as new problems arise in the analysis of the data.

Year 4: A main focus in year 4 will be to finish the production of the software produced under this grant. The software will be well documented and carefully tested before it is made available to the public. Hoeting and Davis will oversee this component of the research with input from Urquhart, Reich, Stevens, Weisberg, and Iyer. In addition to producing software, application of the methodology developed in the preceding years will continue.

We will disseminate results of our research through conference presentations at the Joint Statistical Meetings and other meetings, through the published conference proceedings, and through a series of journal articles. While the focus of this proposal is on methodology and design issues, it is anticipated that software will be developed and made publicly available for all new methodology that is developed under this proposal. This will allow for the dissemination of the research to a broad audience of potential users.

1.5 *GENERAL INFORMATION*

JENNIFER HOETING (CSU) - Principal Investigator, Project 1, is an assistant professor of statistics. Her main research emphasis has been in developing Bayesian methods for improving assessment of predictive uncertainty. Her recent research focus has been on developing methodology for modeling spatially correlated data. She has applied her expertise to a variety of problems, including modeling sandbar size in the Grand Canyon for the National Park Service, predicting presence/absence of rare species for the National Forest Service, and assessing mercury in lakes in Maine, the latter based in REMAP data, and in collaboration with Anthony Olsen of EPA.

RICHARD A. DAVIS (CSU) - Co-Principal Investigator of the Program, is a professor and chair, Department of Statistics. He has substantial training and experience in time series analysis, and has extended those perspectives to spatial statistics. He teaches an advanced course in spatial statistics. The software developed for this course and made available on the web, has generated substantial interests among its users.

F. JAY BREIDT (CSU) - Principal Investigator, Project 2, is an associate professor with expertise in time series and survey sampling. He recently came to CSU from Iowa State University. As part of the Survey Section of Iowa State's Statistical Laboratory, he worked extensively on design and estimation for surveys of ecological conditions and trends, including the National Resources Inventory.

ROBIN REICH (CSU) is an associate professor of forest science with expertise in forest biometry. He has applied spatial statistics to a variety of ecological problems, and has made extensive use of landscape scale information. As a team member he will advise on that project, but his major efforts will be on Project 1 in providing a knowledgeable link between the needs of the potential user community and the research statisticians.

DON L. STEVENS, Jr (OSU) has supervisory and project management experience, both in academia and contract research. While at Eastern Oregon State University, he was a Principal Investigator on a cooperative agreement from EPA to develop the sampling design for the Direct-Delayed Research Project. He has also managed projects on spatial sampling, development of indicators of forest health, ecoregion development, aquatic monitoring, and development of condition indicators for lakes and streams.

STEPHEN B. WEISBERG (SCCWRP) is a biologist who specializes in the design and implementation of environmental monitoring programs. He joined SCCWRP as its Executive Director in 1996. His present research efforts focus on the development of coordinated, integrated, cost-effective regional monitoring in the Southern California Bight. He will provide access to valuable data sets, and be a interface between the academic researchers of the proposed Program, and potential users of the methods to be developed.

1.6 *IMPORTANT ATTACHMENTS* - none

1.7 REFERENCES

References

- Berger, J. O., Oliveira, V. D., and Sanso, B. (2000). Objective Bayesian analysis of spatially correlated data. Technical Report 00-12, Institute of Statistics and decision Sciences, Duke University.
- Besag, J. (1972). Nearest-neighbor systems and the auto-logistic model for binary data. *Journal of the Royal Statistics Society, Series B*, 36:75–83.
- Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–20.
- Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting*. Springer-Verlag: New York.
- Chan, K. S. and Ledolter, J. (1995). Monte Carlo Em estimation for time series models involving counts. *Journal of the American Statistical Association*, 90:242–252.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data, revised edition*. Wiley: New York.
- Davis, R., Dunsmuir, W., and Yang, Y. (1999). Modelling time series of counts. In Ghosh, S., editor, *Asymptotics, Nonparametrics and Time Series*. Marcel Dekker, New York.
- Davis, R. A., Dunsmuir, W. T. M., and Wang, Y. (2000). On autocorrelation in a Poisson regression model. *Biometrika*, 87(3):491–505.
- Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, 57:45–97.
- Durbin, J. and Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-gaussian state space models. *Biometrika*, 84:669–684.
- Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62(1):3–56. With discussion and a reply by the authors.
- Ecker, M. D. and Gelfand, A. E. (1997). Bayesian variogram modeling for an isotropic spatial process. *Journal of Agricultural, Biological, and Environmental Statistics*, 2:347–369.
- Ecker, M. D. and Gelfand, A. E. (1999). Bayesian modeling and inference for geometrically anisotropic spatial data. *Mathematical Geology*, 31:67–83.
- Gumpertz, M. L., Graham, J. M., and Ristaino, J. B. (1997). Autologistic model of spatial pattern of phyophthora epidemic in bell pepper: Effects of soil variables on disease presence. *Journal of Agricultural, Biological, and Environmental Statistics*, 2:131–156.
- Haining, R. (1990). *Spatial data analysis in the social and environmental sciences*. Cambridge University Press: Cambridge.
- 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. *Journal of the American Statistical Association*,

89:368–390.

- Heikkinen, J. and Högmänder, H. (1994). Fully Bayesian approach to image restoration with an application in biogeography. *Applied Statistics*, 43:569–582.
- Hoeting, J. A., Leecaster, M., and Bowden, D. (2000). An improved model for spatially correlated binary responses. *Journal of Agricultural, Biological, and Environmental Statistics*, 5:102–114.
- Högmänder, H. and Møller, J. (1995). Estimating distribution maps from atlas data using methods of statistical image analysis. *Biometrics*, 51:393–404.
- Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.
- Kitanidis, P. K. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research*, 22:499–507.
- Kuk, A. Y. C. and Cheng, Y. W. (1997). The Monte Carlo Newton-raphson algorithm. *Journal of Statistical Computation and Simulation*, 59:233–250.
- Lee, T. C. (2001). An introduction to coding theory and the two-part minimum description length principle. *International Statistical Review*.
- Leecaster, M. (1999). *The Autologistic model with covariates for sample data and robust sampling designs using predicted probability of presence*. PhD thesis, Colorado State University.
- Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.*, 63:215–232.
- Mugglin, A., Carlin, B., and Gelfand, A. (2000). Fully model based approaches for spatially misaligned data. *J. Amer. Statist. Assoc.*, to appear.
- Niu, X. and Tiao, G. C. (1995). Modeling satellite ozone data. *Journal of the American Statistical Association*, 90:969–983.
- Overton, J., Young, T. C., and Overton, W. (1993). Using 'found' to augment a probability sample: procedure and case study. *Environmental Monitoring and Assessment*, 26:65–83.
- Overton, W. S. (1990). A strategy for use of found samples in a rigorous monitoring design. Technical Report 139, Department of Statistics, Oregon State University, Corvallis.
- Preisler, H. K. (1993). Modelling spatial patterns of trees attacked by bark-beetles. *Applied Statistics*, 42:501–514.
- Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, 83:251–266.
- Regal, R. and Hook, E. B. (1991). The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine*, 10:717–721.
- Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer, New York.
- Royle, J. A. and Berliner, L. M. (1997). A hierarchical approach to bivariate spatial prediction. in review, *Journal of Agricultural, Biological and Environmental Statistics*.

- Royle, J. A., Berliner, L. M., Wikle, C. K., and Millif, R. (1998). *A Hierarchical Spatial Model for Constructing Wind Fields from Scatterometer Data in the Labrador Sea*, pages 367–382. Springer Verlag, New York, NY.
- Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.*, 87:108–119.
- Sarndal, C., Swensson, B., and Wretman, J. (1992). *Model assisted Survey Sampling*. Springer-Verlag, New York.
- Schuft, M., Moser, T., Wigington, P., Jr., Stevens, D., Jr., McAllister, L., Chapman, S., and Ernst, T. (1999). Development of landscape metrics for characterizing riparian-stream networks. *Photogrammetric Engineering and Remote Sensing*, 65:1157–1167.
- Smith, R. L. (2000). Environmental statistics. Web Reference <http://www.stat.unc.edu/postscript/rs/envnotes.ps>.
- Stein, M. L. (1999). *Interpolation of Spatial Data*. Springer: New York.
- Stoffer, D. S. (1986). Estimation and identification of space-time Armax models in the presence of missing data. *Journal of the American Statistical Association*, 81:762–772.
- Thompson, S. E. (2001). *Bayesian Model Averaging and Spatial Prediction*. PhD thesis, Colorado State University.
- Wikle, C. (1996). *Spatio-temporal statistical models with applications to atmospheric processes*. PhD thesis, Department of Statistics, Iowa State University.
- Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5:117–154.
- Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, 75:621–629.