

## PROJECT 2: LOCAL INFERENCES FROM AQUATIC STUDIES 2.D. ABSTRACT

**2.1. Sorting Code:** 2001–STAR–D1

**2.2. Title:** Local inferences from aquatic studies

**2.3. Investigator:** F. Jay Breidt, Richard Davis, Jennifer Hoeting, Alix Gitelman

**2.4. Institution:** Colorado State University, Colorado State University, Colorado State University, Oregon State University

**2.5. Project Period:** October 1, 2001–September 30, 2005

**2.6. Project Cost:** \$133,492 in year one; \$691,226 total

### 2.7. Overall Summary

**Objectives:** The objectives of this study are to use hierarchical spatio-temporal models (developed in other parts of the overall proposal) for *local inferences* about aquatic resources. These local inferences include three major applications: small area estimation, deconvolution, and causal inference.

Combining probability survey data and auxiliary information for small area estimation and spatially-explicit prediction is the first major objective of this study.

Deconvolution is the estimation of the cumulative distribution function (cdf) of a variable given noisy measurements of that variable and distributional information about the measurement noise. In an ecological context, spatial cdf's are often of interest. Assessing the effect of spatial dependence on semiparametric deconvolution estimators is the second major objective of this study.

The estimation of cause-effect relationships is of paramount importance in scientific investigations, but is extremely difficult in the non-experimental settings typical of ecological monitoring. Adapting existing causal inference methods to recognize spatial information at different scales is the third major objective of this study.

**Approach:** Spatio-temporal models will be used to assess a nonparametric model-assisted small area estimator and to borrow strength across space and/or time in the construction of maps of ecological resources at fine scales. The effect of spatial dependence on semiparametric deconvolution estimators will be assessed analytically and via simulation using hierarchical spatio-temporal models fitted to EMAP data. Causal inference for aquatic resources will be addressed by adapting existing methodology, with an emphasis on Bayesian belief networks.

**Expected Results:** Spatio-temporal models developed in other parts of this overall program will be used in the present research to improve local inferences for aquatic resources. Expected results include defensible small area estimates or maps which combine probability survey data with auxiliary information; semiparametric deconvolution methods which account for spatially dependent data; and Bayesian belief networks which describe causal paths in an ecological context.

**2.8. Supplemental Keywords:** path analysis, spline estimators.

## 2.E. OVERALL DESCRIPTION

### 2.1 OBJECTIVES

#### Objectives of the proposed research

The objectives of this study are to use hierarchical spatio-temporal models (developed in other parts of the overall proposal) for *local inferences* about aquatic resources. These local inferences include three major applications: small area estimation, deconvolution, and causal inference.

Small area estimation is the use of probability survey data and auxiliary information to compute estimates at a resolution finer than that supported by the sampling design. The design may support precise estimates for elements of a table, while users often like to see a map of local inferences. Construction of defensible maps by combining probability survey data with available auxiliary information is the first major objective of this study.

The deconvolution problem is the estimation of the cumulative distribution function (cdf) of a variable given noisy measurements of that variable and distributional information about the measurement noise. In an ecological context, spatial cdf's are often of interest. Local inferences for such spatial cdf's need to recognize the presence of spatial dependence in sample measurements. Existing nonparametric and semiparametric estimators of the cdf of a variable measured with error will be adapted to the spatial case in the second major objective of this study.

Causal inference, or the estimation of cause-effect relationships, is of paramount importance in scientific investigations, but is extremely difficult in the non-experimental settings typical of ecological monitoring. Local inferences about causation need to recognize spatial information at different scales. Existing methodology will be adapted to the ecological context in the third major objective of this study.

Specifically, we propose to pursue the following research objectives:

- Objective 2.1: Small area estimation and spatially-explicit prediction
  - 2.1.1. Spatially-explicit prediction for continuous variables
  - 2.1.2. Spatially-explicit prediction for presence/absence variables
  - 2.1.3. Assessing nonparametric model-assisted synthetic estimation for small areas
  - 2.1.4. Assessing uncertainty in spatially explicit predictions
- Objective 2.2: Deconvolution
  - 2.2.1. Semiparametric estimation of the distribution function of a variable measured with error

- 2.2.2. Comparison of nonparametric and semiparametric deconvolution methods
- Objective 2.3: Causal inferences in ecological data
  - 2.3.1. Specify conditional independence relationships
  - 2.3.2. Implement Bayesian belief network methodology in an ecological context
  - 2.3.3. Assess software systems for Bayesian belief networks in an ecological context

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

### **Importance of the proposed research**

The proposed research is important because it will use spatio-temporal models developed in other parts of this overall program to improve local inference for aquatic resources. These improved local inferences will include more capabilities for producing small area estimates and maps, combining probability survey data with auxiliary information; better tools for semiparametric deconvolution, accounting for spatially dependent data; and better descriptions of causal paths in aquatic systems, using Bayesian belief networks. All these methods will be developed and tested using data sets from EMAP and EMAP-designed studies, so that results will be immediately useful in practice.

## *2.2. APPROACH*

### **Objective 2.1. Small area estimation and spatially-explicit prediction**

In nearly any survey, some users will ask for estimates at a resolution finer than that supported by the design. A survey might be designed to make reliable inferences for large watersheds, but a user may be interested in smaller watersheds. The small and variable sample sizes within these small watersheds may lead to insufficiently precise design-based estimators. A related problem is prediction of ecological conditions at unsampled locations.

When the precision of design-based estimators is not sufficient for inference about subdomains, small area estimation techniques are often used to replace or supplement the direct estimation, through the use of synthetic and composite estimators (Ghosh and Rao, 1994). These techniques typically rely on parametric distributional assumptions about the relationship between the variables of interest and auxiliary information available for the subareas. Much of the small-area estimation literature such as that reviewed in Ghosh and Rao (1994) assumes the availability of auxiliary information only at the level of the small area. In many EMAP applications, complete coverage auxiliary information from a remotely-sensed image or landscape

model will be available. Consequently, it is possible to consider models with site-specific auxiliary information rather than areal averages. Standard spatio-temporal modeling techniques can then be used, though complexities of the design such as rotating panels may require special consideration in the model structure.

***Objective 2.1.1. Spatially-explicit prediction for continuous variables***

A useful class of parametric statistical models for small area estimation and spatially-explicit prediction of continuous variables in the EMAP context is the class of spatial linear models. For a given location  $s$  in a study region  $D$ , an observation is generated according to the linear model

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

where  $\boldsymbol{\beta}$  is a vector of unknown coefficients,  $\mathbf{X}(s) = (X_1(s), \dots, X_p(s))'$  is a vector of known explanatory variables, and  $\delta(s)$  is the unobserved error at location  $s \in D$ . A sample of observations  $\mathbf{Z} = (Z(s_1), \dots, Z(s_n))'$ , is collected. Given  $\mathbf{Z}$ , which is a partial realization of the random field  $\{Z(s), s \in D\}$ , it is desired to predict  $Z(s)$  at unsampled locations,  $s \notin \{s_1, \dots, s_n\}$ , or at a different spatial scale, such as  $|A|^{-1} \int_{ACD} Z(s) ds$ . The former can be thought of as a mapping problem while the latter is a small area estimation problem. The presence of the explanatory variables and the spatial dependence make solution of this problem feasible.

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  $\sigma^2$  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. We focus on the Matern autocorrelation (e.g., Stein, 1999) function as a flexible alternative to standard autocorrelation functions. In fact, both the exponential and Gaussian autocorrelation functions are members of the Matern class of functions.

We will adopt a Bayesian approach for inference about the process (1) at unobserved sites or over partially-observed regions. See Robert and Casella (1999) for an overview of recent numerical methods for approximate Bayesian inference. Thompson (2001) adopts proper prior distributions in the Bayesian kriging context 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. We will apply these methods in the EMAP context and extend them to allow for anisotropy and non-stationarity. Non-Gaussianity will also be considered.

### **Objective 2.1.2. Spatially-explicit prediction for presence/absence variables**

A particularly dramatic and interesting case of non-Gaussianity is that of binary data, such as presence/absence of an ecological condition. Presence/absence models will be used to estimate the probability of an ecological condition at all locations. One choice for modeling spatially correlated binary response data is the autologistic model. 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. This type of model has been applied and extended to agricultural research (Gumpertz, Graham, and Ristaino, 1997), forestry (Preisler, 1993), archaeology (Besag, York, and Mollié, 1991), and biological range mapping (Heikkinen and Högmänder, 1994; Högmänder and Moller, 1995). For many EMAP applications, however, there is more information available than is utilized by the basic autologistic model, such as covariates available as a GIS coverage.

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

$$\text{pr}(z_i = 1 \mid \mathbf{x}_i, \mathbf{z}_{-i}, \boldsymbol{\theta}, \beta) = \frac{\exp\{\mathbf{x}_i^T \boldsymbol{\theta} + \beta s_i\}}{1 + \exp\{\mathbf{x}_i^T \boldsymbol{\theta} + \beta s_i\}}, \quad (2)$$

where  $\mathbf{z}_i$  is a vector of covariates for the  $i$ th site with the first element equal to 1,  $\boldsymbol{\theta}$  is a vector of parameters for the covariates, and  $\beta$  is the parameter associated with the spatial covariate,  $s_i$ . The spatial covariate for site  $i$ ,  $s_i$ , is equal to the total number of sites where the ecological condition 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 hexagonal grid structure and the sampling plan. In the hexagonal grid, all neighbors have a common edge (unlike the second order square lattice where some neighbors share only a corner), and each grid cell is completely surrounded by neighbors (unlike the first order square lattice, where corners are missed).

The likelihood function for the autologistic model is analytically intractable, except in trivial cases, so alternative estimation methods are necessary. We will adopt the Bayesian methods used in Hoeting *et al.* (2000) to estimate the parameters of the model. These methods and extensions of the autologistic model described elsewhere in this overall program will be applied to produce parameter estimates as well as maps showing the posterior probability of presence over the area of interest. Such predictions can be averaged over subregions for small area estimation.

### ***Objective 2.1.3. Assessing nonparametric small area estimators***

Because small area estimation techniques rely on parametric distributional assumptions about the relationship between the variables of interest and auxiliary information available for the subareas, the resulting small area estimates are subject to potential model bias if these relationships are incorrectly specified.

Nonparametric specification of regression relationships between study variables and auxiliary information has the potential to reduce model misspecification bias and make it possible to provide a more robust synthetic estimator for use in small area estimation. The synthetic estimator could be used alone or in a composite estimator, combined with a noisy but unbiased direct estimator. Note that the bias is often the key component of the mean squared error for a synthetic estimator, since the design variance of the synthetic estimator is typically very small (e.g., Särndal, Swensson, and Wretman, 1992, Ch. 10). Nonparametric regression has the potential to trade off some increase in design variance for large reductions in bias in some circumstances.

We are particularly interested in the case in which it is of interest to produce internally consistent small area estimates for a large number of different study variables, where the “small” areas have moderate but not extremely small sample sizes. If the nonparametric methods work here, it may provide a prescription for fairly routine small area estimation at a resolution somewhat finer than the design resolution, without a great deal of modeling effort. (For extremely small areas, it will be necessary to use stronger modeling assumptions and more sophisticated spatio-temporal models.) We will compare nonparametric regression synthetic estimators to parametric regression synthetic estimators via simulation. Simulations will be based on hierarchical spatio-temporal models fitted to EMAP data sets from surveys of lakes, streams, and estuaries.

### ***Objective 2.1.4. Assessing uncertainty in spatially explicit predictions***

While fitted spatio-temporal models can yield prediction mean square errors (mse’s) as well as predictions, these prediction mse’s typically ignore uncertainty due to model selection. Model selection usually 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 data-generating model. This approach can lead to underestimation of the uncertainty of the inferences (Regal and Hook, 1991; Draper, 1995; Madigan and York, 1995; Kass and Raftery, 1995; Raftery, 1996). Underestimation of uncertainty can lead to misinterpretation of maps and tables of small area estimates.

Bayesian model averaging (BMA) is one solution to this problem. Posterior distributions are computed under each model in a candidate set of models, and weighted by the corresponding posterior model probability to obtain the final posterior. Hoeting *et al.* (2000) provide an overview of the currently available methods for implementing BMA in a variety of statistical models. Thompson (2001) extended BMA methods to spatial linear models. BMA is a promising approach for assessing

uncertainty of small area estimators and spatially-explicit predictors in EMAP data sets. Relationships among environmental variables are often poorly understood, so different candidate models may yield plausible predictions. BMA is a disciplined approach to combining such predictors and assessing the overall uncertainty.

## Objective 2.2. Deconvolution

In many studies of aquatic resources, it is of interest to estimate the cumulative distribution function (cdf)  $F_X$  for a study variable  $X$  given noisy measurements of the form  $Y = X + \epsilon$ , where  $\epsilon$  represents measurement error. For example, it may be of interest to characterize the distribution  $F_X$  of average dissolved oxygen,  $X$ , in a population of small stream segments. A stream segment is sampled from the population and a water sample is drawn from the segment. The measurement of dissolved oxygen for the water sample, denoted  $Y$ , is a noisy version of the actual average dissolved oxygen in that stream segment,  $X$ .

We assume that  $\epsilon$  has a continuous density,  $f_\epsilon$ , and  $X$  has a continuous density,  $f_X$ . For identifiability, information about  $f_\epsilon$  is required. It is often the case in practice that  $f_\epsilon(x)$  is estimated from repeated measurements known to have common  $X$ -values. We will not focus on the estimation of  $f_\epsilon$ , but instead assume that it is known.

The cdf of  $Y$  is given by

$$F_Y(x) = \int F_X(x - s)f_\epsilon(s) ds = F_X * f_\epsilon(x),$$

where  $*$  denotes convolution. A naive estimator of this cdf would be the empirical cdf of  $Y$ , given by

$$\hat{F}_Y(x) = n^{-1} \sum_{i=1}^n \mathbf{1}_{\{Y_i \leq x\}},$$

where  $\mathbf{1}_{\{Y_i \leq x\}} = 1$  if  $Y_i \leq x$  and 0 otherwise. This estimator is biased for  $F_X(x)$ , however, because

$$E[\hat{F}_Y(x)] = F_X * f_\epsilon(x) \neq F_X(x).$$

Alternatives to this naive estimator include parametric estimators, which assume a known parametric form for  $f_X$ ; nonparametric estimators, which assume smoothness and other regularity conditions on  $f_X$ ; and semiparametric estimators, which assume  $f_X$  belongs to a parametric class of densities with an unknown number of parameters. Parametric estimators seem less useful in the context of EMAP, since many different variables are of interest, most of which have no *a priori* known parametric distributional form. Nonparametric estimators in the literature include nonparametric maximum likelihood estimators (Eggermont and LaRiccia, 1997), which are not smooth, and kernel estimators (Diggle and Hall, 1993; Goutis, 1997; Liu and Taylor, 1989; and Stefanski and Carroll, 1990). Semiparametric estimators include mixture estimators (Mendelsohn and Rice, 1982; Cordy and Thomas, 1997; West, 1997) and spline estimators (Chen, 1999; Chen, Fuller, and Breidt, 1999). We focus on spline-type estimators in this proposal, and compare results with optimal kernel estimators and mixture estimators.

***Objective 2.2.1. Semiparametric estimation of the distribution function of a variable measured with error***

Following Chen (1999) and Chen, Fuller, and Breidt (1999) (henceforth CFB), we consider semiparametric estimation of the distribution function of a variable measured with error. The first proposed semiparametric estimator is derived in two major steps: weight estimation and normal transformation. In the weight estimation step, an estimator of the empirical cdf of  $X$  is obtained by acting as if the observations were generated by a set of  $n$   $\tilde{X}$ -values. The  $\tilde{X}$ -values are defined as a transformation of the original  $Y$ -values such that the sample mean and variance of the  $\tilde{X}$  are consistent estimators of the mean and variance of  $X$ . The weight of each  $\tilde{X}$ -value in the estimation of the cdf of  $X$  is calculated by post-stratification of the  $Y$ -values and by using the cdf of  $Y$  conditional on  $X$ . In the normal transformation step, the weighted empirical cdf of  $X$  is smoothed by mapping the approximate quantiles of  $X$  to the quantiles of a standard normal using a cubic spline. The spline coefficients are obtained by solving a standard quadratic programming problem and the number of join points of the spline is selected by AIC (Akaike's Information Criterion). In the simulation study reported in CFB (1999), this semiparametric estimator dominated an optimal kernel estimator (Diggle and Hall, 1993) and a normal mixture estimator (Cordy and Thomas, 1997) for a wide class of densities. The densities considered included a variety of symmetric, skewed, and bimodal alternatives.

The other semiparametric estimation procedures we will consider estimate the parameters of the spline function by maximum likelihood; that is, assuming the transformed spline cdf is the true data-generating mechanism. These methods use the quantile regression spline estimator described above as an initial estimator. In all cases, the number of parameters of the spline function is determined by the data with a simple criterion, such as AIC. The proposed spline estimators performed better than a normal mixture estimator (Cordy and Thomas, 1997) and better than an optimal kernel estimator (Diggle and Hall, 1993) in simulation studies reported in Chen (1999). These estimators will be adapted as necessary to the spatial setting.

***Objective 2.2.2. Comparison of nonparametric and semiparametric deconvolution methods***

The simulation studies of Chen (1999) will be extended to consider cases in which the true underlying variable is spatially dependent. Measurement errors will initially be assumed to have no spatial dependence, as would be the case when they represent random sampling error. The spline estimators of the cdf of a variable measured with noise will be compared to mixture estimators and optimal kernel estimators, for a variety of target distributions and noise levels. Analytic comparisons will be developed as well.

### Objective 2.3. Causal Inferences in Ecological Data

Causal inference in non-experimental settings is generally quite difficult, even though questions of cause and effect remain of paramount interest in scientific inquiry. Estimating cause-effect relationships is additionally difficult when only one condition of the causal agent in question—for example, only the presence of environmental stressors, not their absence—is observed. In these situations, the *strength of correlations* between the stressors and the ecological condition of the lake, stream or estuary where the stressors are observed can be estimated, but it must often be left to subject area arguments to establish and/or defend any *causal* significance. What we lack in these situations is a basis for comparison: what would the ecological condition look like if the stressors were absent?

We will therefore consider the non-experimental situations in which some subset of the sampled lakes, streams or estuaries can be characterized by the presence of an environmental stressor (or some collection of stressors), while the remaining subset is characterized by its (their) absence. In this situation, where “treatment” conditions are observed, but not assigned, several related statistical models have been used in efforts to tease out causal connections among the variables of interest. These models include structural equation models (SEM), path analysis (PA) and Bayesian belief networks (BBN). Our proposal is to examine these models in the context of ecological data, with our primary focus upon BBN. We will also consider methods using propensity score matching (Rubin & Rosenbaum, 1983). Propensity scores are scalar summaries of (possibly highly dimensioned) observed covariates, and they can be used to match sampled units (for example, a stream with stressors present would be matched with a stream with stressors absent) according to a wide array of observed quantities that characterize the units.

Perhaps the most popular of the three aforementioned statistical models for causal modeling is SEM, popularized in the econometrics literature (Heckman and Robb, 1986; Morgan, 1990), though these models have also made inroads into educational and sociological research (Holland, 1988). PA is a related model that has seen primary use in educational and social science research (Holland, 1988). With both of these methods, there are *a priori* notions of cause and effect relationships. In SEM, the *structural equations* relate cause and effect variables, and in PA, *paths* between variables specify cause and effect relationships.

In both SEM and PA, one goal is to estimate the strength of postulated effects. In certain circumstances, however, an actual causal effect can be estimated; this requires the presence of a viable *instrumental variable* (IV) among the collection of observed variables. A variable  $T$  is an instrumental variable if the distribution of another variable,  $Y$ , depends on  $T$  only through a third variable,  $X$ . Ideally,  $T$  and  $X$  should be correlated, and  $T$  and  $Y$  uncorrelated conditionally on  $X$ . Then  $T$  can be used as a surrogate or instrument for  $X$ . An instrumental variable, then, is a variable that is partly defined by a conditional independence relationship, known as the *exclusion restriction assumption*, or *IV assumption*.

The third statistical formulation, related to both SEM and PA, is BBN, intro-

duced first by Pearl (1988; see also Madigan et al., 1997). A BBN consists of a system of nodes and directed (acyclic) arcs between them, and in this way is quite similar to a path diagram. The arcs represent causality (by way of conditional independence relationships), and the acyclic property incorporates the notion that if A is a cause of B, B cannot simultaneously be a cause of A. Each node in the BBN represents a variable in the model, and each has a prior probability distribution, conditional on inputs to that node. Updates to the prior distributions are made as data are added or incorporated into the network. Though IV's are not explicitly defined in BBN, the conditional independence relationships between nodes can essentially capture their functionality.

***Objective 2.3.1. Specify conditional independence relationships***

Finding reasonable instrumental variables (or specifying conditional independencies) in ecological datasets may be quite difficult. The conditional independence relationship (i.e., the IV assumption) is specifically required, and with complicated, spatially correlated ecological data, such independence may simply not exist among observed variables. Were such an IV to exist in this context, it would have to be something correlated with the environmental stressors, but uncorrelated with ecological condition once environmental stressors are accounted for. Further complicating matters, Bound et al. (1996) caution that in practice, not only may instruments be difficult to find, but also even slight correlation between  $Y$  and  $T$  (in our setting ecological condition and the IV, respectively), after accounting for  $X$  (the stressors), can substantially affect the causal estimates.

One component of our proposed research is to investigate several datasets at our disposal for IV candidates. If legitimate IV's are identified, we will use them to help tease out causal effects between environmental stressors and ecological conditions. In the absence of defensible IV's, however, all three models (SEM, PA and BBN) reduce to models in which only the strength of postulated cause and effect relationships are estimable. In these cases, we will look to incorporating propensity scores in an effort to, at a minimum, account for observable covariates to explain presence or absence of environmental stressors.

***Objective 2.3.2. Implement Bayesian belief network methodology in an ecological context***

Due to the similarities between SEM, PA and BBN, and for the purpose of coordinating with the other modeling components of the broader statistical modeling in this ecological data project, we will focus on BBN. This construct is, from the Bayesian viewpoint, the most natural for incorporating subject area knowledge into both the prior distributions at each node in the network, and the possible causal linkages. Furthermore, since Bayesian hierarchical modeling will be used in other components of this project, it seems appropriate to adopt the BBN framework for this component, in the hope of eventually linking the models. Lee (2000) uses a BBN

to assess the effect of land-use differences on bull trout populations. His model incorporates watershed information as well as land-use and bull trout population trend information.

As an illustration of the possible implementation of BBN methodology to ecological data, consider the hypothetical BBN in Figure 1. This network represents the systems under study in the EMAP-Oregon dataset—all streams and rivers in Oregon are sampled, and information on fish, fish tissue contaminants, macroinvertebrates, water chemistry, habitat, watershed stressors and temperature is obtained. In this network, we postulate that stressors effect the ecosystem at two different levels—at the watershed (landscape) level and at the stream level. This framework allows us to incorporate data regarding both the stressors (watershed stressors and water chemistry information) and the ecological condition of the ecosystem at different levels of the model. We will estimate direct effects of stressors at each level, recognizing that different stressors may have different effects at each level, and estimate indirect effect of the stressor below the level at which it is introduced to the system. For instance, a watershed stressor may effect stream-level condition (e.g., as measured by water chemistry), and that stream-level condition may in turn effect the health of inhabitants of the stream.

***Objective 2.3.3. Assess software systems for Bayesian belief networks in an ecological context***

The BBN in Figure 1 by no means exhausts the possibilities for associations and connections between the nodes. Our objectives include fitting several different BBNs, and assessing their relative suitability. Several software systems have been developed to analyze BBN (these include Pearl, 1988; Lauritzen and Spiegelhalter, 1988; Anderson et al., 1989; as well as the Hugin system [<http://www.hugin.dk>]; Netica [<http://www.norsys.com>]). We will evaluate some of these systems for their applicability to ecological data and its inherent hierarchical and spatially correlated nature. Indeed, the BBN described in Figure 1 underscores the need to incorporate hierarchical modeling into the network, since data are collected at two levels—watershed and stream (river) levels.

Our research into statistical methods for causal inferences from ecological data will begin with an examination of existing methods developed for substantially different contexts. We will also proceed with analysis of several datasets with an eye towards identifying conditional independence relationships that will ultimately inform our modeling of causal associations among stressors and ecological condition of sampled waterways. Our modeling perspective will be primarily Bayesian, as we view the inclusion of expert subject area knowledge as essential to the modeling in general, and particularly to the modeling of causal pathways.

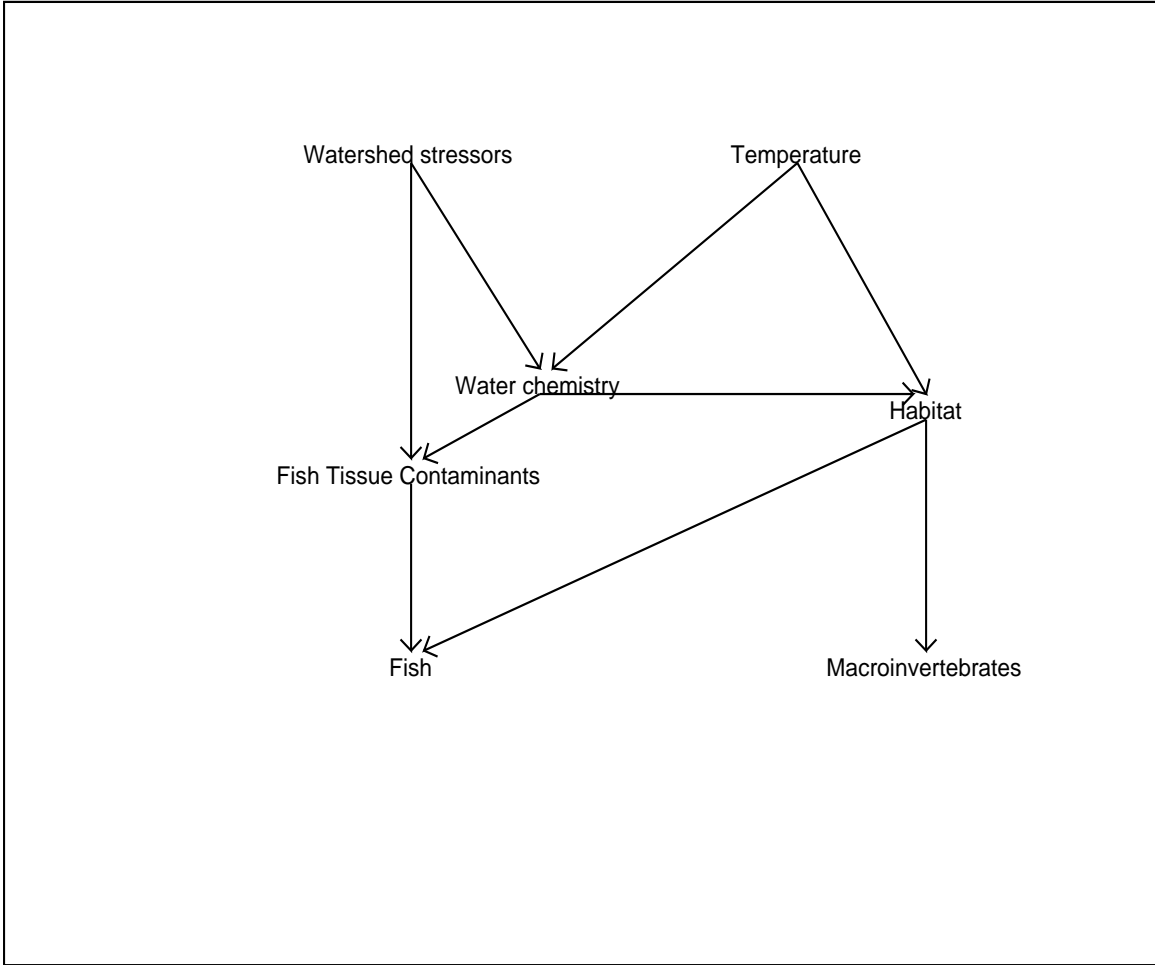


Figure 1: *Hypothetical Bayesian belief network representing the systems under study in the EMAP-Oregon dataset. All streams and rivers in Oregon are sampled, and information on fish, fish tissue contaminants, macroinvertebrates, water chemistry, habitat, watershed stressors and temperature is obtained.*

### *2.3. EXPECTED RESULTS*

The EPA's solicitation calls for research in hierarchical modeling as a means of combining information and capturing spatial and temporal structure in ecological resource data. In response to this solicitation, hierarchical spatio-temporal models will be developed in other parts of this overall program. These developed models will be used in the present research to address other parts of the solicitation.

We expect to develop new methods for spatial prediction and small area estimation, together with appropriate measures of uncertainty. Spatio-temporal models developed for continuous variables will be used for spatial prediction and small area estimation, effectively combining probability survey data with auxiliary information from GIS coverages. The autologistic model with covariates will be adapted and extended as necessary for predicting probabilities of an ecological condition at all sites in a study region. In both the continuous and discrete case, Bayesian inference and Bayesian model averaging will account for uncertainty due to unobserved values of the process, uncertainty due to unknown parameter values, and uncertainty due to model selection. Spatio-temporal models will also be used to assess by simulation a simple nonparametric model-assisted synthetic estimator. This estimator may be useful for routine estimation in subregions smaller than those supported by the design, but not extremely small.

Spatio-temporal models will also be used to give new understanding of the behavior of nonparametric and semiparametric deconvolution methods, using spatially correlated data. Spline estimators of the distribution function of a variable measured with error will be studied and compared to optimal kernel estimators and semiparametric mixture estimators for data with spatial dependence.

Bayesian belief networks and other techniques for causal inference will be adapted to the ecological context and applied to EMAP data sets. Specification of appropriate conditional independence relationships will be investigated using available EMAP data sets. Propensity scoring methods will also be considered. Finally, the utility of existing software for causal inference in an ecological context will be assessed.

### *2.4. MANAGEMENT PLAN AND MILESTONES*

The following is a rough timeline for the proposed study. Year one will involve start-up: assessing local inference needs, obtaining appropriate data sets, and developing methods for the simplest cases. (While the components of this project rely to a large extent on models developed elsewhere in the overall program, there are components which can be worked on without a model or with only a simple model.) In years two, three, and four, we will use spatio-temporal models for local inference as they become available from other parts of the overall program. In year four, however, we will emphasize dissemination and outreach more than research and methodological development.

We now describe this rough timeline in more detail.

In year one, nonparametric small area estimators will be developed and applied to some EMAP data sets, as described in Objective 2.1.3. These estimators are simple, weighted estimators which assume only that the mean and variance of the study variable are smooth functions of the available covariates. These estimators ignore any residual spatial dependence in the study variable after removing the effect of the covariates. Results for the nonparametric estimator will provide a benchmark for comparison with estimators that rely more heavily on parametric modeling assumptions about the regression relationship between study variables and available covariates, and about the residual spatial dependence structure. These comparisons will be conducted on a case-by-case basis throughout the study period, as new models become available through Objectives 2.1.1 and 2.1.2. Model uncertainty, addressed in Objective 2.1.4, will be continually assessed throughout the study period.

In Objective 2.2.1, the spline-based deconvolution estimators will be studied analytically for simple cases of spatial or temporal dependence in the  $X$ -variables in year one. The extensive simulations of Objective 2.2.2, which compare spline estimators to kernel methods and mixture estimators using fitted spatio-temporal models, will be completed in year two. Further analytic comparisons will follow in year three.

Objective 2.3.1, specification of appropriate conditional independence relationships will be investigated using available EMAP data sets in years one and two. Propensity scoring methods will also be considered. Adaptation of Bayesian belief networks and other techniques for causal inference to the ecological context will be addressed in years two and three. Finally, the utility of existing software for causal inference in an ecological context will be assessed throughout the study period.

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. Students and/or post-doctoral research associates will participate in all of these research and dissemination activities.

## 2.5. GENERAL INFORMATION

Hoeting and Davis have jointly advised students in spatio-temporal modeling research. Breidt and Davis have collaborated extensively on time series modeling. We anticipate further successful collaborations among Hoeting, Davis, and Breidt on various aspects of spatio-temporal modeling, as described in Project 1 of the proposed program. Extension by the investigators of the developed models for applications in local inferences from aquatic studies is expected to be seamless.

Breidt was co-advisor on the dissertation of Chen (1999), which developed spline estimators for the cdf of a variable measured with error. Breidt has access to simulation code used for that research, which will facilitate adaptation of these methods to the spatial context.

Gitelman's dissertation research was on causal inference. Her background in hierarchical modeling and Bayesian inference will interface naturally with the inter-

ests of the other investigators, so that successful collaboration is anticipated. Travel funds have been requested to allow Gitelman to interact extensively with the group at CSU.

## *2.6. IMPORTANT ATTACHMENTS*

None.

## 2.7. REFERENCES

- Anderson, S.K., Oleson, K.G., Jensen, F.V., and Jensen, F. (1989) “Hugin—a shell for building Bayesian belief universes for expert systems,” in *Proceedings of the Eleventh International Congress on Artificial Intelligence*, 1080–1085. Reprinted in Shafer, G. and Pearl, J. (1990), *Readings in Uncertainty*, San Mateo, CA: Morgan Kaufmann.
- Besag, J. (1972), “Nearest-neighbor systems and the auto-logistic model for binary data,” *Journal of the Royal Statistical 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.
- Bound, J., Jaeger, D.A. and Baker, R.M. (1995), “Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak,” *Journal of the American Statistical Association*, 90, 443–450.
- Chen, C. (1999), *Spline estimators of the distribution function of a variable measured with error*, unpublished Ph.D. dissertation, Iowa State University, Ames, IA.
- Chen, C., Fuller, W.A., and Breidt, F.J. (1999), “A semiparametric estimator of the distribution function of a variable measured with error,” *Communications in Statistics*, to appear.
- Cordy, C.B. and Thomas, D.R. (1997), “Deconvolution of a distribution function,” *Journal of the American Statistical Association*, 92, 1459–1465.
- Diggle, P.J. and Hall, P. (1993), “A Fourier approach to nonparametric deconvolution of a density estimate,” *Journal of the Royal Statistical Society, Series B*, 55, 523–531.
- Draper, D. (1995), “Assessment and propagation of model uncertainty,” *Journal of the Royal Statistical Society, Series B*, 57, 45–97.
- Eggermont, P.P.B. and LaRiccia, V.N. (1997), “Nonlinearly smoothed EM density estimation with automated smoothing parameter selection for nonparametric deconvolution problems,” *Journal of the American Statistical Association*, 92, 1451–1458.
- Ghosh, M. and Rao, J.N.K. (1994), “Small area estimation: an appraisal” (with discussion), *Statistical Science*, 9, 55–93.
- Goutis, C. (1997), “Nonparametric estimation of a mixing density via the kernel method,” *Journal of the American Statistical Association*, 92, 1445–1450.
- Gumpertz, M. L., Graham, J.M., and Ristaino, J.B. (1997), “Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence,” *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 131–156.
- Heckman, J. and Robb, R. (1985), “Alternative methods for evaluating the impact of interventions,” in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, New York: Cambridge University Press.
- 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.
- Högmander, H. and Moller, J. (1995), "Estimating distribution maps from atlas data using methods of statistical image analysis," *Biometrics*, 51, 393–404.
- Holland, P. (1988), "Causal inference, path analysis, and recursive structural equation models," *Sociological Methodology*, 18, 449–484.
- Kass, R.E. and Raftery, A.E. (1995), "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.
- Lauritzen, S.L. and Spiegelhalter, D.J. (1988), "Local computation with probabilities on graphical structures and their application to expert systems" (with discussion), *Journal of the Royal Statistical Society, Series B*, 50, 205–247.
- Lee, D. (2000), "Assessing land-use impacts on bull trout using Bayesian belief networks," in Ferson and Burgman (eds.), *Quantitative Methods for Conservation Biology*, New York: Springer.
- Leecaster, M. (1999), *The autologistic model with covariates for sample data and robust sampling designs using predicted probability of presence*, unpublished Ph.D. dissertation, Colorado State University, Fort Collins, CO.
- Liu, M. and Taylor, R.L. (1989), "A consistent nonparametric density estimator for deconvolution problem," *Canadian Journal of Statistics*, 17, 427–438.
- Madigan, D., Krzysztof, M. and Almond, R.G. (1997), "Graphical explanation in belief networks," *Journal of Computational and Graphical Statistics*, 6, 160–181.
- Madigan, D. and York, J. (1995), "Bayesian graphical models for discrete data," *International Statistical Review*, 63, 215–232.
- Mendelsohn, J. and Rice, J. (1982), "Deconvolution of microfluorometric histograms with B-splines," *Journal of the American Statistical Association*, 77, 748–753.
- Morgan, M. (1990), *The History of Econometric Ideas*, Cambridge: Cambridge University Press.
- Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA: Morgan Kaufmann.
- 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.
- Rosenbaum, P.R. and Rubin, D.B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.
- Rubin, D.B. (1974), "Estimating causal effects of treatments in randomized and non-randomized studies," *Journal of Experimental Psychology*, 66, 688–701.
- Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.
- Stefanski, L.A. and Carroll, R.J. (1990), "Deconvoluting kernel density estimators," *Statistics*, 21, 249–259.

- Thompson, S.E. (2001), *Bayesian model averaging and spatial prediction*, unpublished Ph.D. dissertation, Colorado State University, Fort Collins, CO.
- West, M. (1997), "Studies of neurological transmission analysis using hierarchical Bayesian mixture models," *Journal of the American Statistical Association*, 92, 587–606.