

**LINKAGE OF MULTISCALE MULTISOURCE MULTI-TIER DATA FOR
THE PURPOSES OF REGIONAL ASSESSMENTS AND MONITORING
A Research and Outreach Prospectus of Advanced Mathematical, Statistical, and
Computational Approaches**

G. P. Patil

Center for Statistical Ecology and Environmental Statistics
Department of Statistics
The Pennsylvania State University
University Park, PA 16802
<http://www.stat.psu.edu/~gpp>

Research and Outreach Team

G. P. Patil, Mathematical and Environmental Geospatial Statistics and Analysis

Department of Statistics, The Pennsylvania State University

W. L. Myers, Remote Sensing, Natural Resources, and Software

School of Forest Resources, The Pennsylvania State University

C. Taillie, Computational Statistics and Stochastics

Department of Statistics, The Pennsylvania State University

Stephen Rathbun, Department of Statistics, The Pennsylvania State University

Jaeyong Lee, Department of Statistics, The Pennsylvania State University

G. S. Biging, Resource Information Technology and Landscape Ecometrics

College of Natural Resources, University of California, Berkeley

Ben Kedem, Mathematics Department, University of Maryland

Eric Slud, Mathematics Department, University of Maryland

Sudip Bose, Department of Statistics, George Washington University

December 1, 2001

Project 1: Interleaved Stochastic Modeling for Environmental Inference

Introduction

Often there are multiple sources of empirical information on a particular environmental phenomenon, as for example EPA-MAHA, Chesapeake Bay Program, USGS-NAWQA, and state water resource agency activities overlapping in Pennsylvania. Usually each of the sources is limited in spatial and/or temporal coverage and contains sections of observations missing (not necessarily at random). The sources tend to be measured on disparate supports (in the sense of geostatistics) and on overlapping, but not identical variables. Separately each of the sources provides limited capability to describe the phenomenon. Combining the information sources will considerably enhance capabilities to arrive at sound environmental decisions.

We introduce the term “interleaved” in a spatial-temporal modeling context to represent such situations and to distinguish them from the multivariate spatial time-series situation. In the latter, information is collected on the same variables at the same spatial sites at regular temporal intervals (See, for example, Brown, Le and Zidek, 1994). In the interleaved setting the selection of data sources may be achieved adaptively, where the difficulty of incorporation is weighed against the marginal increase in information (both inferred from the stochastic model). Interleaved thus implies at least partial lack of spatial alignment for collection sites on different occasions, and also admits differences in suites of variables observed. The unaligned aspect is novel relative to conventional geostatistics.

Application Contexts

As indicated above, many prospective applications call for combining information from different surveys that are interleaved over space and time. Such an applications spectrum, however, does not exhaust the possibilities.

A rather different type of context with regard to application is a spatial extension of sampling on multiple occasions with partial replacement, which has an extensive background in forestry literature predating the rise to prominence of spatial statistics. Related settings that are more recent and consider spatial issues explicitly have been concerned with optimally configuring networks of pollution monitoring stations. In the partial replacement setting, an initial spread of sampling points is replaced in part on subsequent occasions.

The remeasured stations provide a longitudinal component of information, while repositioned stations provide new information on spatial variability. Although sophisticated and expensive monitoring stations for atmospheric deposition may be difficult to reposition, there are several types of ecological sampling which do have the requisite mobility. In the spatially motivated version, kriging variance on a prior occasion would guide degree of replacement and determine new positions. Progressive coupling of

variogram models across occasions would drive the kriging estimators and provide for generating maps which reflect spatial organization in the characteristic of interest.

Yet another context might focus on detection of change in spatial structure between occasions for (perhaps only partially) unaligned surveys, possibly with only partial interpenetration of coverages. This might well be the case for epidemiological settings such as tracking of forest insect infestations or diebacks over time. Such settings are likely to be complicated by higher intensity sampling in known or suspected “hotspots”. Bias in the latter situation could be controlled by segregating “supplemental” samples of hotspots to avoid spatial “contamination” arising from purposive placement of the supplemental samples. Mappings of the estimated response surface take place in the usual manner. Opportunities for inference appear to lie primarily in the realm of introducing and relaxing constraints during fitting of variogram and cross-variogram models.

Variogram Modeling

The foregoing sampler of application settings is only suggestive of the considerable scope for comparative spatial analysis of datasets differing temporally and locationally but reflecting a common phenomenological random field. The emphasis in all of these situations should be on the incisiveness of fitting variogram and cross-variogram models under specific assumptions which are consonant with scientific understanding of the phenomena being considered. Greater working latitude is needed in modeling stochastic processes.

One possibility for achieving greater latitude lies in further investigation of continuous non-Gaussian random fields that are mixtures of Gaussian ones. Methods of estimating parameters of such random fields are well known. Optimal statistical decision rules in such mixture models are nonlinear and are considerably more effective than the linear ones. We would propose to pursue development and application of such nonlinear methods to handling of non-Gaussian data.

Approach

We propose to model the phenomena as a (possibly non-Gaussian) spatial-temporal multivariate random field. Spatial-temporal field approaches have received increasing attention in recent years, especially in relation to their application to meteorological and air pollution monitoring networks. In these situations, however, the information is usually assumed to be collected relatively uniformly over time at a fixed number of spatial locations. The situations we are interested in will have information collected at different sites, during different time periods and in potentially different variables. One could view this situation as a special case of the usual situation where there is very large scale missing data. However, one would still need to model the missing data mechanisms explicitly as they will often be dependent on the sampling designs and directly related to the variables of interest (Little and Rubin, 1987).

We propose to take into account the missing and sampling data mechanisms in the modeling process. Conversely, the efficacy of our results can be benchmarked in Monte Carlo fashion by deleting components of multivariate observation vectors that are otherwise appropriate for cokriging. Performance of models for deleted data can then be compared directly to results of cokriging for full datasets.

The random field approach will require modeling of the mean functions and the cross-variograms. Clark *et al.* (1987) have developed a modified cross-variogram specifically intended for non-aligned data. Myers (1991) terms this the *pseudo cross-variogram* and examines its modeling. Additional modeling is required for the mixture parameters if the non-Gaussian formulation is used. The existing homogeneous models for the (cross) variograms often do not capture the complexities of environmental data (Le and Zidek, 1992). The development of realistic parametric heterogeneous models is still in its infancy. Vecchia (1985) and Jones and Vecchia (1993) consider classes of stochastic partial differential equations that result in spatial autoregressive moving average (ARMA) models. See also Loader and Switzer (1992). The most common modification in practice is to use a simple geometric anisotropy parameter with an existing homogeneous model (See Cressie, 1991). We will develop a class of models that generalize the (homogeneous) Matern class (Matern, 1986; Handcock and Wallis, 1994) in the same direction as the (heterogeneous) Vecchia (1985) class of models.

The structure of the models is context specific as the cross-variograms will need to take into account “change of support” issues for each of the data sources and variables.

One major advantage of the stochastic modeling approaches is that they allow the prediction uncertainty to incorporate the various sources of uncertainty that will affect the quality of the predictions. The sources of uncertainty in the prediction are sampling variability, model uncertainty, and model misspecification. Sampling variability is the variation in the prediction given that the specified model is correct. The model uncertainty is the uncertainty about the correct model within the specified class of models given the available data. The model misspecification is the degree to which the specified class of models does not represent the spatial-temporal variation of the phenomena. The current available statistical approaches typically only take into account the first source when they assess the overall uncertainty. In general, complicated (e.g., non-parametric) model classes have smaller model misspecification while simple (e.g., parametric) model classes have smaller model uncertainty.

The best available statistical approaches to estimation in both the spatial and spatial-temporal situations are described in the review paper by Guttorp and Sampson (1994). In particular, Sampson and Guttorp (1992) have developed a nice approach that allows the heterogeneous estimation of the covariance between any two spatial locations, whether monitored or not. See also Mardia and Goodall (1993). The drawback of these approaches is that they as yet only provide a point estimate of the spatial covariance, and do not prove a natural way for the model uncertainty to be incorporated in any later inference (Handcock, 1996). In many circumstances this uncertainty will be the major source of uncertainty in the underlying model. Other methods with a Bayesian flavor have also

been developed to improve on the traditional method (See Kitanidis, 1986; Handcock, 1989; Gotway and Cressie, 1993). We will further develop the Bayesian approach of Handcock and Stein (1993) and Handcock and Wallis (1994) to take into account the multivariate nature of the present setting.

Project 2: Quantitative Characterization of Spatial Patterns and Algorithmic Detection of Pattern Transitions at Varying Scales

Introduction

Understanding ecological processes and monitoring their dynamics requires characterization of spatial patterns and objective detection of pattern transitions at scales ranging from landscape to regional. Conception and computation of landscape “metrics” is commonly practiced among contemporary landscape ecologists (O'Neill, et al., 1988; Milne, 1988; Riitters, et al., 1995; Kepner, et al., 1995). This work, however, has not generated more substantial pattern resolving power. The typical mode of operation is to begin with a subjective choice of extent, over which a global computation of chosen landscape. A global landscape metric will predictably meld any inconsistencies of pattern that are internal to the predetermined extent. What the landscape ecologist sees in his/her computed metric is, in part, a reflection of subjectivity in choice of extent. This circularity can only be broken by objectivity in treatment of extent. In effect, the general idea underlying a variogram must be incorporated in landscape pattern analysis.

Digital sampling of continuous signals versus categorical elements, and categories derived from classifications of continuous signals are fundamental distinctions. Remote sensing and geographic information systems have provided a rich repository of all these sources of landscape information which is, in turn, traditionally expressed as maps (for example, see Riitters and Wickham, no date). However, such maps again tend to be treated subjectively.

Pattern itself remains illusive, and the human eye is as yet unequalled in ability to detect shifts in spatial frequency of signals that can be presented as visual stimuli. One approach to the pattern issue is to exploit rather than reject this visual capability. Images of a given type and resolution can be examined visually to select areas of perceptual uniformity that also have integrity relative to current landscape understanding. Such prototype landscape elements can then be modeled parametrically as random fields with secondary calibration in terms of a suite of landscape metrics. This corresponds to the training set concept of supervised image analysis and training of neural networks. The probabilities of juxtaposed pattern elements and spans of transition between the several types of elements can then be determined from visual analysis of random transects on the image. One can then develop empirical Bayesian expectation of co-occurrence for different pattern elements in larger spatial extents, as well as model the composite properties of such mixtures. This, in turn, provides a basis for calibrating moving (multi-) window detectors of pattern transitions.

To complement the modeling approach, one can progressively degrade image and other spatial data sets from the floor resolution using low pass filters. The degraded data can then be compared to the mixture models for first order validation. Different filters, however, have different effects. Therefore, several filters with different effects must be investigated. It is likewise necessary to make comparisons to other data sets having coarser native resolution.

Proposed Research

Dimensional Analysis for Determining Scaling Domains

As pointed out by Levin (1992), concern should not lie with determining an appropriate measurement scale, but rather with performing analyses at multiple scales. Part of multiscale analysis then becomes the determination of scaling ranges that reveal statistically self-similar patterns, which generally implies that the shape or spatial distribution of an entity measured at one scale is similar to that of another measurement scale. This is akin to defining "domains of scale," which Wiens (1995) proclaims should be central to the development of a theory of scaling in ecology. Scaling domains are indeed central to hierarchy theory (see O'Neill, Johnson and King, 1989).

Ecologists who desire tools for doing multiscale analysis have been attracted to fractal-based methods (Hastings and Sugihara, 1993). For a recent review of concepts and applications of fractals in ecology, with emphasis on multiscale landscape analysis, see Johnson, Tempelman and Patil (1995). Thus far, ecological scaling ranges have been identified subjectively from breaks in the linearity of log-log plots (Bradbury, et al, 1984) or in a series of rolling regression estimates over different scales (Krummel, et al, 1987). Objectifying this approach will require formalized hypothesis testing.

Numerical Data and the Information Dimension

When the value within each pixel of a tessellated image is numerical, an estimate of the information dimension has been proposed by Loehle and Wein (1994) as the finite difference approximation between any two scales. This approach is intended to be more robust than estimating dimension via linear regression (Loehle and Li, 1995). Application to a forested landscape indicated that diversity was very high at small measurement scales (pixel sizes), while the larger measurement scales revealed lower diversity that corresponds to greater fragmentation. Between these scaling extremes, the estimated information dimension revealed visual evidence of statistical self-similarity over a large spatial range.

Categorical Data and the Stochastic Similarity Dimension

When each pixel in a landscape image is labeled as one of M distinct states, such as land cover types, a method based on the dimension of self-similar stochastic processes may be applicable. If a self-similar random field can be viewed as a stationary Markov Chain, then arguments similar to one used in Pesin and Tempelman (1995) may allow us to

compute the correlation dimension of the associated shift dynamical system. A preliminary study (Tempelman, in progress) shows that this dimension is closely related to the conditional entropy as one moves from a "mother" cell to a "daughter cell" which is nested within the mother cell. This *stochastic similarity dimension* can be considered as a generalization of the Mandelbrot (1983) similarity dimension.

This new notion of dimension reveals how the entropy of a set of categories, such as land uses, is scaled across different measurement unit sizes. The stochastic similarity dimension has the added feature of quantifying the dependence of entropy at one scale on the entropy of the next larger scale.

Approach

The information dimension and the stochastic similarity dimension discussed above would be applied to detecting changes in pattern at different measurement scales. In order to establish formal tests of self-similarity, we would develop the statistical properties of these dimension statistics. Progress has been made in this direction for the fractal dimension (Hall and Wood, 1993; Ogata and Katsura, 1991; Taylor and Taylor, 1991) and the correlation dimension (Olofsen, Degoede and Heijungs, 1992; Theiler, 1990) when data are realizations of a point process or are line graphs. More broadly, the aim of the proposed research is the development of statistical tools for detecting and characterizing process change across a spatially nested hierarchy of regular tessellations. For reasons of estimability, the process would be taken to be spatially homogeneous at each level in the hierarchy, in a sense explained below.

Two cases distinguish themselves depending upon whether process response is numerical (and spatially additive) or categorical. In the latter case, the relationship between the process at two different levels (scales) in the hierarchy will be modeled by a spatially homogeneous set of transition probabilities $\{p_{st}\}$ where p_{st} is the probability that a daughter cell is in state t when its mother cell is in state s . Homogeneity requires that p_{st} be the same for all mother cells at a given scale, although our approach allows this parameter to depend upon the location of the daughter cell within the mother cell.

For a numerical response, spatial additivity is exploited to model the inter-scale relationship by a spatially homogeneous set of 'splitting fractions' $\mathbf{p}_i = \{p_{ij}\}$, with p_{ij} as the fraction of the response from mother cell i that is allocated to daughter cell j . For a deterministic process, the p_{ij} are fixed numbers and homogeneity means that p_{ij} does not depend upon the mother cell i . More realistically, the p_{ij} would be random but with $\mathbf{p}_{i1} = \{p_{i1j}\}$ and $\mathbf{p}_{i2} = \{p_{i2j}\}$ identically distributed realizations from some common 'splitting distribution.' The Dirichlet would be a natural candidate for consideration as the splitting distribution. Splitting models of this type occur also in representing hierarchical species diversity as well as in rock fragmentation and particle size issues. But the present context has two simplifying features. First, the number of daughter cells per mother cell is constant and fixed ahead of time rather than random. Second, the labelling of daughter

cells (subscript j) has an intrinsic significance determined by the daughter cell's location within the mother cell. Intrinsic labels allow the use of the Dirichlet distribution, for example, rather than the more complicated ordered Dirichlet that would be needed in a species diversity study.

At each level in the hierarchy, the scale dependence of the process can thus be characterized either by a set of transition probabilities (categorical response) or by a splitting distribution (continuous response). The null hypothesis states that these characterizations are constant across the scaling hierarchy. Tests for this hypothesis would be developed. Initially, the various entropy-based dimensions would be studied as candidate test statistics with development of the needed distribution theory and computational procedures. While the various fractal dimensions appear to be of independent interest in ecological work, they are powerful as test statistics against only a limited set of scaling transitions. Tests with broader diagnostic capability would directly examine equality of the transition probabilities or splitting distributions across the scaling hierarchy. Likelihood ratio tests would be employed for this purpose, using multinomial distributions for categorical responses and Dirichlet distributions for continuous responses.

Spatial homogeneity is crucial for parameter estimation in the above approach. The consequences of using spatially averaged parameters would be studied via simulation. In addition, localized or moving window versions of the tests would be developed for handling high resolution data with a broad spatial extent.

A limitation of the above approach to modeling the relationship between responses at two scales is that the transitions depend only on the mother cell and not on neighbors of the mother cell. Although neighborhood dependencies could be incorporated at the model level, estimation and testing becomes problematic because there is no formal vehicle for incorporating the dependencies into the inferential procedures. Instead, a joint random field model of the responses at the two scales appears to be needed. Gibbs random fields, described below, provide an apparatus for such models.

Gibbs Random Fields

Following publication of the seminal article of Geman and Geman (1984), Gibbs fields have provided a unifying framework for the probabilistic approach to the problems of image analysis (Winkler, 1995). Perhaps because the emphasis has been on the image itself rather than on the fitted process model and its parametric structure, Gibbs modeling has been little applied to environmental assessment issues. Our approach would be to operationalize the modeling and inference for Gibbs random fields, bringing these methods into the environmental scientist's toolbox. The effort would be directed toward random fields having a categorical response defined over a rectangular lattice.

Let the configuration $x = \{x_i\}$ be a realization of such a field where i ranges over the lattice and x_i is the response at lattice point i . Under the Gibbs model, configuration x has its probability given in terms of parameters θ by

$$\pi(x) = \frac{\exp[-H(x; \theta)]}{Z(\theta)}, \quad (1)$$

where $Z(\theta)$ is the normalizer. The utility of the Gibbs model results from the fact that the function H , known as the Gibbs measure, has the same parametric form as an indicator regression function, namely,

$$H(x; \theta) = \sum_{i,j} \sum_{s,t} \theta_{ij}^{st} \cdot I_{ij}^{st}(x). \quad (2)$$

Here, the first sum ranges over all pairs i, j of lattice points and the second sum ranges over all pairs s, t of possible response values (categorical). The indicator $I_{ij}^{st}(x)$ vanishes unless $x_i = s$ and $x_j = t$ when it takes the value 1.

The number of parameters in the model (2) can be enormous and parsimony is generally achieved by (i) putting $\theta_{ij}^{st} = \theta_{i'j'}^{st}$ when i', j' is a translate of i, j (homogeneity) and (ii) putting $\theta_{ij}^{st} = 0$ unless i and j are near (often, nearest) neighbors in the lattice.

The parameters θ_{ij}^{st} are known as the interaction potentials and specify the spatial dependence in the Gibbs model just as the variogram specifies spatial dependence in the kriging model. Compared with variogram modeling, however, the large number of parameters entering linearly in (2) becomes advantageous for objective hypothesis formulation and testing. As an example, the null hypothesis of isotropy in model (2) is expressed formally as $\theta_{ij}^{st} = \theta_{i'j'}^{st}$ whenever i', j' is a rotation of i, j and, subject to computational limitations, is easily tested by a likelihood ratio test. By contrast, krigers typically assess isotropy by subjectively eyeballing various directional variograms.

Multi-resolution Gibbs models incorporate interaction cross-potential parameters relating responses at two different scales. These are functionally similar to image degradation models specifying joint probability of true and degraded images in image analysis (Guyon, 1995, section 6.7.2).

The probability function (1) is of the exponential family type, for which there is a well-established statistical theory (Lehmann, 1986) modulo availability of an analytic expression for the normalizer $Z(\theta)$, $Z(\theta) = \sum_x \exp[-H(x; \theta)]$. Unfortunately this sum, ranging over all possible configurations x , has an astronomical number of summands, making explicit computation impossible. The day is saved because all but a tiny fraction of these configurations occur with infinitesimal probability. While the remaining configurations cannot be precisely delineated, they can be effectively sampled by Markov Chain Monte Carlo (MCMC) techniques, including the Gibbs sampler (Geman and Geman, 1984; Smith and Roberts, 1993), thereby enabling computation of means, higher moments, and similar characteristics.

There is extensive literature on point estimation for Gibbs fields (Guyon, 1995). Younes (1988) has developed a stochastic gradient algorithm, driven by the Gibbs sampler, that converges to the maximum likelihood estimate. The conditional pseudolikelihood estimator (Besag, 1974) is easier to compute but is statistically less efficient and is less convenient for likelihood ratio tests. Possolo (1986) describes a logistic estimator for the case of binary responses.

Hypothesis testing has received less attention. Prum and Fort (1991) examine some specialized hypotheses using asymptotic normality of the test statistics. Guyon (1995) establishes the general theory for the likelihood ratio (LR) test, showing that the reference distribution is chi square when maximum likelihood estimates are used but is a linear combination of chi square variates for the less efficient parameter estimation methods.

Guyon (1995) does not discuss actual computation of the LR test statistic. The difficult part of the computation is the difference, $\ln Z(\hat{\theta}_1) - \ln Z(\hat{\theta}_0)$, where $\hat{\theta}_0$ and $\hat{\theta}_1$ are the parameter estimates under the null and the alternative hypotheses, respectively. We propose to calculate this difference from the secant approximation,

$$\frac{\partial \ln Z}{\partial \theta} [\hat{\theta}_1 - \hat{\theta}_0],$$

where the derivative is evaluated at some point θ on the ray joining $\hat{\theta}_0$ and $\hat{\theta}_1$. (A multi-step secant approximation will actually be used to provide needed accuracy.) The partial derivatives will need to be computed but, from the general theory of exponential families, $\partial(\ln Z)/\partial \theta$ is an expected value of a linear combination of the indicator functions in (2) and this expectation can be obtained by MCMC methods.

Small, artificially generated, data sets would be used for algorithm testing and debugging, and this experience will provide empirical guidance in selecting data structures and data access protocols for fitting large data sets and complex models.

Data

The Office of Remote Sensing of Earth Resources (ORSER) at Penn State University is currently producing a very detailed land cover database for Pennsylvania. Starting with six frequency bands from LANDSAT images, pixel responses will be classified into 256 categories. The resulting image will consist of 7000 by 7000 pixels, each pixel being 30 meters on a side. A vector format will be developed simultaneously, whereby polygons will be delineated according to similar pixels. These data, which are expected to become available this Spring (1996), will be managed through the ARC/INFO and ATLAS GIS packages along with the ER-MAPPER image analysis software.

Project 3: Combining Data From Probability Sampling with Data From Hot Spot Directed Sampling

It can be beneficial to focus a portion of the sampling intensity upon the right hand tail of a probability distribution where the response of interest is expected to be most pronounced. Examples include the use of catch data in fisheries management and the contaminant sampling protocol adopted in the REMAP program for the New York bight. Hot spot sampling may occur simply because the large response values are of primary scientific interest, but statistical justification for this type of sampling can also be given in terms Neyman's theorem on optimal sample allocation across strata which says that sampling should be most intense where response variability is highest. Since variability often grows with response, e.g., constant coefficient of variation, Neyman's result then implies that sampling efficiency is enhanced when large response values are sampled more intensively than they occur in nature. Further, the benefit is greatest for very skew distributions.

Sound inference cannot be based upon hot spot data alone since the latter is a biased sample and extrapolation to the full population would require and would be inordinately sensitive to unverifiable modeling assumptions (Patil and Taillie, 1991). But the extrapolation can be largely finessed when hot spot data is combined with sample data from the full population, and the proposed research would develop a methodology for accomplishing this. Cox and Piegorsch (1994) give a broad discussion of the need for combining environmental information.

Clearly, inference from the combined data requires a link between the two types of samples. We propose to study parametric links that can be represented in terms of a weight function $w(x) = w(x; \theta, \beta)$ with

$$f_{\text{HS}}(x; \theta, \beta) \propto w(x; \theta, \beta) f(x; \theta),$$

where f and f_{HS} are the density functions for the probability sampling and the hot spot sampling, respectively, and where θ and β are vectors of parameters. Suitably normalized, the weight function $w(x)$ can be interpreted as the probability that a value x is associated with a hot spot and becomes a candidate to enter the hot spot data set through the hot spot density f_{HS} . Typically, then, one would expect $w(x)$ to be an increasing function of x . In general, the parametric form for the weight function must involve the parameters θ of f , if only to reflect the scaling of the data. Additional parameters β are needed to regulate the shape of the weight function, determining, for example, how quickly w grows with increasing x and a 'threshold' value x_0 above which the weight function is effectively equal to unity. In a simple but idealized case, the weight function takes the values 1 or 0 depending upon whether x is above or below a threshold x_0 . Here, f_{HS} is the truncation of f to the interval (x_0, ∞) . The case of general weight function might be described as smooth or 'fuzzy' truncation.

Alternatively, the link might be modeled by letting f_{HS} be the density for the largest order statistic in a hypothetical sample of size M drawn from f . The integer M then becomes a parameter of f_{HS} . From the well known expression for the density of order statistics, we obtain

$$f_{\text{HS}}(x) \propto [F(x; \theta)]^{M-1} f(x; \theta),$$

where F is the cumulative distribution function corresponding to the density f . We see that this is a special case of the above weighted representation with

$$w(x; \theta, \beta) = [F(x; \theta)]^{M-1}$$

and $\beta = M$. Although M is an integer in the order statistics motivation, mathematically it can take any real value greater than unity---an important feature for applicability of likelihood methods. This model for the weight function has the advantage that the dependence upon θ is already built in. On the other hand, there is only the single external parameter M which may not provide sufficient flexibility for regulating the shape of the weight function. Linear combinations of weight functions with different values of M would provide additional model flexibility. When the weight function is not required to be monotonic, then other order statistics can be used to model the weight function, e.g., second largest out of M , etc.

The research would include modeling of the weight function with corresponding inferential and goodness of fit procedures, as well as examination of the sensitivity of conclusions to modeling assumptions. An important issue here is the minimum size needed for the (less efficient) probability sample in order to get a handle on the form of the weight function. Our approach would emphasize likelihood methods since the likelihood function for the combined data sets takes the simple form

$$L(\theta, \beta) = \left[\prod_{x \in \text{HS}} w(x; \theta; \beta) \right] \left[\prod_x f(x; \theta) \right],$$

where the first product is over the values in the hot spot data set and the second product is over the values in the combined data set. The research would also document the improvement in parameter estimation attainable from inclusion of hot spot data with a selection of weight functions. As indicated above, the best improvement is expected for skew distributions which is exactly where improvement of standard methods is most needed.

Our focus would be parametric, but the issue of combining data can also be addressed through nonparametric density estimation. The best known of such estimators are of the kernel type (Silverman, 1986). Letting \hat{f} and \hat{f}_{HS} be kernel estimators of the densities

f and f_{HS} , respectively, a nonparametric estimate of the weight function is given by the ratio

$$\hat{w}(x) = \frac{\hat{f}_{\text{HS}}(x)}{\hat{f}(x)}.$$

We see these nonparametric estimates as preliminary graphical aids to be used in identifying appropriate models for f and for the weight function.

Realistic environmental assessments generally involve multivariate data. Multivariate extensions of the foregoing are necessarily application specific depending upon the question to be addressed as well as the interpretation of hotspot. In a multi-species fishery, for example, hotspot is determined by an individual fisherman's target species and this information might not be recorded in the dataset. Again, hotspot might be determined by distance from the multivariate mean vector. In the simplest setting, hotspot is determined by a single component of the multivariate observation.

Methods would be illustrated using the NY Bight data set and multivariate extensions would be developed specifically for these data. Availability of other datasets, including for the Chesapeake Bay and the Mid-Atlantic Highlands, will be explored.

Data

Several data sets have been gathered to describe New York Harbor area sediments and their effects upon biota. Probability samples were selected for one data set (REMAP, gathered in 1993--94), which sampled the entire Harbor, western Long Island Sound, and the New York Bight Apex. Several additional data sets are non-probabilistic and from subareas of the Harbor region. There are two motivations for combined analyses and interpretations of the data sets (Joel O'Connor, EPA Region II, personal communication): (i) to gain more precise descriptions of sediments and biotic effects in specific areas, providing a better understanding of the causes of the effects and (ii) to describe sediments within channels which are expected to be dredged, providing quantitative bases for more efficient, inexpensive sampling designs for sediments to be dredged.

Approach to Synthesis of Multi-Tier Data by Spatial Accord/Discord, Conveyance, and Reconciliation, and Related Statistical Techniques for Regional Ecological Assessment

To generalize the scope of inference for proposed research, the multi-tier data terminology is taken to mean multiple spatially referenced datasets which differ between datasets with respect to observational detail and spatial inclusiveness. The number of such datasets can and usually will change over time in any given applications setting.

Generically stated, the analytical need is to achieve a synthesis of the information contained in the several datasets whereby the synthesis speaks to a particular environmental question or set of questions. The question of interest must have direct expression among one or more of the datasets, or else be expressed by a preliminary knowledge-based mapping for at least a portion of the region to be assessed. Since it would be extremely rare for all datasets to be collected at exactly the same time, a temporal component of variation is intrinsic. While it is possible to handle the temporal component(s) implicitly for purposes of assessment at a point in time, it will generally be more informative to treat the possibility of temporal change explicitly. There thus occurs some blurring of usual distinctions between assessment and monitoring as a consequence of temporal considerations in linkage of tiers.

The subset of tiers containing variables or themes that directly address the current question(s) of interest can serve to decode relevant information contained in other tiers. In effect, this primary (sub)set constitutes first-hand evidence with the remaining tiers comprising circumstantial evidence. A first concern thus becomes one of determining whether the several sources (tiers) of circumstantial evidence are telling the same story relative to the question of interest. More formally but still generically stated, the first analytical issue is one of accord/discord among the tiers of circumstantial evidence. With maps as circumstantial tiers, for instance, there might be two kinds of informational overtones among the suite of available maps. One subset of maps might reflect physiographic influences such as geology, soils, and terrain. The other subset of maps might reflect human influences on the landscapes/resources. Signals from physiographic variables might exhibit considerable accord over space, and so also for the human influences. The physiographic and human components, however, could show substantial discord. Thus some or many sites having favorable physiographic influences on ecology could be severely degraded by human influence. This same issue of informational accord can penetrate within a tier when there is multivariate acquisition of data for the tier. There are also regional and local sides to this issue. Local sectors of discord may occur between tiers that are otherwise in accord. A systematic approach to identifying such anomalies is therefore needed.

Multivariate statistical strategies such as clustering, principal components, ordination, canonical correlation, multidimensional scaling, and factor analysis provide a strategic framework for segregating tiers and variables into subsets of mutual accord. However, these broad strategies require tactical adaptation to account for and exploit the spatial nature of environmental data. Despite the long history of cartography, there still remains considerable scope to formalize expressions of pairwise discord or disparity between maps and/or mappable datasets, and to conduct analysis of comparative advantage among alternatives. The persistence of this challenge is due in part to the computationally problematic nature for earlier computer generations of otherwise appealing possibilities. The phenomenal rate of growth in computing power and prospective involvement of high performance computing facilities at UNM make the computational aspects much less daunting than before. Tactical adaptation of such statistical strategies is an ongoing thrust of current research at Penn State focused on multi-scale statistical analysis for critical areas in watersheds and landscapes under joint NSF/EPA sponsorship

(Johnson *et al.*, 1996, 1997a,b,c,d; Patil, 1997; Patil *et al.*, 1997; and Patil and Taillie, 1998a,b)

Two recent outgrowths of the Penn State research effort have particular relevance to linking tiers of data over broad areas. One of these is a compressive approach to multivariate spatial data using progressive/recursive hyperclustering in a manner that has been given the acronym PHASE for "Pixel Hyperclusters Approximating Spatial Ensembles" (Myers, 1997; Myers, Patil and Taillie, 1997a,b,c). PHASE analysis extracts the dominant "messages" from a multivariate spatial dataset such as a multiband remotely sensed image and creates a spatial "quilt" consisting of patch types for which all blocks in a type have substantial informational concordance to the degree of being a first order surrogate for the entire dataset. The spatial structure of the block pattern is subject to ready analysis by methods as simple as tabulation of edge apportionment and comparative scrutiny relative to random expectation. Patch types exhibiting pronounced edge affinity constitute landscape complexes which lend spatial support to classificatory work that is conventionally otherwise limited to variates alone. The spatial distribution of such landscape complexes can likewise be analytically investigated at a higher level of abstraction in spatial pattern. Locally variant areas can be determined by differencing the patch type means from the original data vectors on a cell-by-cell basis and then studying the residual vectors as second order data. PHASING the residual vectors yields a second order approximation, and the process can be carried to still higher orders until the residuals approach spatial white noise.

The second development called "echelons" decomposes a quantitative spatial variable of at least ordinal strength into topological hillform objects that comprise a spatial dendrogram (Myers, Patil and Joly, 1997). Among the many potential uses of this methodology is determination of particularly variant sectors of a tier according to length of residual vectors for a PHASE approximation.

The combination of PHASE and echelon methodologies is prospectively a very powerful approach to linkage of tiers and broad area change detection. A NASA-sponsored research effort for the latter purpose is just getting underway. Two PHASE-formulated tiers can be linked on the basis of comparative spatial structure in a manner that was not heretofore possible. A PHASE formulation of each tier is first done individually, and reciprocal cross-compilation is then done for each tier on the basis of spatial structure in the other tier. If there is accord in the spatial structure of the two tiers, then the cross compiled version should closely approximate the internally compiled version. Where there is discord in the spatial structure of the two tiers, the informational distance between the respective centroids should be correspondingly large. The direction relative to informational axes for the difference vector between centroids of the two versions will indicate the nature of the local discord between tiers. An echelon analysis of lengths for difference vectors of the two versions enables computer-intelligent determination for the placement of discord as well as its magnitude. It is especially important to appreciate that this strategy entirely avoids direct comparative analysis of variables for one tier with those of the other. The comparison is always between two different compressions of the same tier. In one version the compression is intrinsic to the tier. In the other version the

compression of the tier is conditioned by the spatial structure of the other tier. We are currently in the process of demonstrating the considerable flexibility of this approach by analyzing the guild-wise species richness of breeding birds in a manner that treats the richness of each guild as an analog for a band in a remote sensing device.

The concept of conveyance is particularly important for linking locally intensive but not spatially expansive site data to less intensive but more expansive tiers. The essence of conveyance is that the more expansive tier must provide a sort of "carrier signal" that serves as an extrapolator. PHASEs are again particularly promising in this regard (Myers, Patil and Taillie, 1997a,b,c). Reasonable presumption is that influence extends into similar environments, with perhaps some decay of influence over distance. PHASEs organize the extent of their tier into collections of patches having similarity with respect to the (circumstantial) variables. Each collection of patches will likewise have varying degrees of similarity to other collections of patches. The similarity of one collection to another is bolstered by companion spatial affinity relative to juxtaposition (edge apportionment). A fuzzy membership function for site likeness can be formulated in terms of cluster similarity to the cluster(s) occurring at or in the immediate vicinity of the intensive site. Distance decay in the fuzzy membership can be incorporated if desired. If variables measured for the intensive site and extensive tier are essentially the same, then the evidence is direct rather than circumstantial. A single extensive tier of direct evidence may be considered an adequate basis for extrapolation. If site and extensive tiers do not have common variables, then evidence is essentially circumstantial. When evidence is circumstantial, it is well to have additional extensive tiers (or layers) that provide "second opinion" for site likeness. The different indications of site likeness must then be weighted. A conservative approach would use minimum indication of likeness, but more liberal weighting rules may be preferred according to context of assessment. Several mappings with different weightings can also be prepared and compared for discord. This again places emphasis on having robust and flexible metrics of accord/discord between mappings and/or mappables.

When working with tiers of plot-based survey data, it becomes necessary to interpolate data from plot locations over space in the survey tier before proceeding with linkage of survey tier to site tier. Geostatistical strategy with its several kriging tactics for interpolation is reasonably well-suited for this kind of intra-tier conveyance. Adequate agency precedent for incorporating this approach into assessments has been established by the USFS with its FIA (Forest Inventory/Analysis) data in the northeast, whereby both intensity of effects and distributions of species have been speculatively extrapolated across the intervening space between plot locations by geostatistical methods (Hershey, 1996).

Given some further development, informed application of accord/discord and conveyance to multi-tier data will produce competing syntheses that require formal reconciliation. The earlier supposition of competing physiographic influences versus human influences can again be illustrative. Considered over the extent of a region there will be varying degrees of contention among competing syntheses derived from (sub)sets of tiers. The word "contention" here has a basic sense of discord as discussed above, but also involves

differing strengths of evidence for the competing syntheses. In certain sectors of a region one synthesis may bear strongly (perhaps high fuzzy relevance) whereas another bears only weakly. A broader analogy is that of need for systematic negotiation protocols among competing syntheses with provision for expression of minority opinion. It is of paramount importance to remain cognizant that the evidence in some of the tiers may be highly circumstantial. There must be adequate safeguard against "rushing to judgment" on the basis of circumstantial evidence. Thus the reconciliation protocol must map not only the apparent compromise, but also the subregions in which particular sources of evidence call the compromise into question. It will then be the responsibility of scientists and experts to resolve the contention either by human consensus or by mounting supplementary data acquisition targeting the sectors of uncertainty.

The proposed research has two complementary contexts for demonstrative application, both with substantial real relevance to environmental issues and management. One context is a humid temperate setting in the eastern deciduous forests of PA. The other is a more arid zone of transition between biomes in the Sevilleta LTER area of central NM. Ecotype mappings of PA's extensive state and national forests will play a major role in future ecosystem-oriented management of the natural resources, particularly for shaping the new generation of state forest management plans for the planning period beginning in the year 2000. The transitional ecotones between biomes in Sevilleta constitute ecological tension zones with moisture and temperate being major factors which should respond in a barometer fashion to prospective global change.

Ecomapping protocols and initial maps for representative sectors of each forest in PA are currently being developed, with ERRI and School of Forest Resources at Penn State helping with strategy, technology, and facilitation. The ERRI also houses a large repository of spatial data for PA ranging from satellite image data to floodplain information and watershed mappings relative to pollution loadings derived from hydrologic models. The FIA unit in the Northeastern Forest Experiment Station of USFS has a strong program of analysis for PA FIA forest survey data that complements the repository at ERRI. The Bureau of Forestry in PA DCNR is among the national leaders in certification of public forest lands for sustainable forest management. The demonstrative applications context for PA is to treat the initial ecotype maps as spatial representations of expert opinion which may be subject to refinement through synthesis of the several tier of spatial data that are available.

The Sevilleta LTER conducts intensive data collection for the biome transitional zones along with more extensive surveys and also maintains a large repository of more synoptic data such as satellite imagery, physiography, climatic records, and so on for which metadata catalogs are available on the Worldwide Web. The demonstrative applications context for Sevilleta is to determine the spatial representativeness of intensive data collection sites, and to use the determinations for strengthening synthesis of tiers across the entirety of the large LTER for improved tracking of biome transition zone dynamics. The existence of the High Performance Computing Center at UNM with a focus on computational biology serves to complement the LTER and lends feasibility to some linkage investigations that would otherwise not be computationally tractable.

Traditional minimum distance classifiers do not explicitly take spatial proximity into account. Two approaches will be investigated for incorporating spatial coherence into thematic classifications. The first approach adds a penalty term to the distance to be minimized, where the penalty measures spatial incoherence or roughness (Luo, 1998). A constant of proportionality in the penalty term (traditionally denoted by λ) is adjustable and determines the relative contribution of data response versus spatial coherence to the final classification. Discord among different data tiers can also be included into the penalty term; minimization then achieves inter-tier concurrence.

The second approach develops a categorical analog of the kriging error model for continuous responses (Cressie, 1991, p.128). In this approach, the results of available classification are acknowledged to be imperfect with category specific error rates (λ). Application of Bayes theorem leads to a revised empirical variogram and revised indicator kriging reflective of intra-tier coherence (Filipponi, Patil, and Taillie, 1998a,b). The method would be extended to achieve inter-tier concurrence with indicator cokriging. Results to date have been with a single thematic map, and with λ spatially constant and chosen subjectively by visual inspection of the revised maps. Availability of several competing thematic maps opens the way to choosing λ as spatially variable with small values in regions of map concordance and with large values in regions of map discord. The method is then similar to fuzzy kriging (Piotrowski *et al.*, 1994, 1996).

Approach to Synthesis of Multi-Tier Data by Adaptive Fuzzy Data Modeling Frameworks and Related Statistical Techniques for Regional Ecological Assessment

A growing focus on ecological assessment at regional scales has been changing the traditional view in ecological risk assessment. Characterization, quantification, estimation and prediction of ecological risks at multiple scales are typically very difficult. Valuation of ecosystem components and scaling from intensive investigations of single sites or relatively small geographic areas to regional landscape level are also ill-defined. Some uncertainty is unavoidable in ecologists' assessment and prediction about ecological systems. In these situations, unexpected risks and/or environmental changes may result from decisions that must be made. The probabilistic approach alone cannot represent uncertainties attached to systems where some deterministic dynamical characteristics are unknown or deliberately ignored as well as uncertainties attached to their mathematical model. The challenge now is to introduce/develop new methods to address these concerns. In this proposal, we propose a fuzzy statistical and modeling approach to multiscale ecological risk assessment under uncertainty to improve risk-based decision-making. This new method integrates utilization of knowledge or judgment of experts together with statistics of the available vague data and imprecise information so as to provide the modeling basis for a predictive ecological risk assessment at regional scale. Our intent is to begin the process of bringing the concepts and principles of fuzzy logic and systems into main stream ecological assessments.

Although multi-valued logic was developed early in this century, the development of fuzzy set theory by Zadeh in 1965 marks the turning point of its development from an academic backwater to modern application. The central idea is that members of a set may have only partial membership; that is, there is gray between the black and white of true and false. Zadeh referred to this gray area as "fuzzy," which was an inauspicious choice of terms for science and engineering. Despite much thought, no better term has yet emerged. But, once we realize that fuzzy logic and set theory are quantitative ways of characterizing intrinsic ambiguity rather than substituting imprecision for precision, the shorthand term fuzzy becomes acceptable in science and engineering. Already the application of this theory to the development of fuzzy systems for commercial electronic products and system controllers in Japan has been remarkable. Fuzzy mathematics itself is not fuzzy. It is a mathematical method for dealing with fuzzy phenomena in nature.

Allen and Hoekstra (1992) have described the conceptual approaches and potential of fuzzy systems analysis for applied ecology. Li (1996) recently edited a special issue on Fuzzy Modeling in Ecology for the international journal *Ecological Modeling*, and has contributed an invited chapter for Springer-Verlag's ecological assessment guidebook to begin to explore the horizons of this new approach for ecological assessment applications (Li, 1998a).

Fuzzy sets as formulated by Zadeh (1965) are based on the simple idea of quantifying the degree of "belongingness" of an element to a subset. Assume that the symbol S represents the entire set, in other words, a universe of discourse. In classical set theory, each element $x \in S$ either belongs to a given subset A of S , or does not belong to A . The subset A is accordingly represented by a function $f_A : S \rightarrow \{0,1\}$ where

$$f_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

The function f_A is called the characteristic function of subset A .

Fuzzy sets are introduced by generalizing the characteristic function f_A . Let μ_B be a function defined on S whose values are in the unit interval $[0,1]$, namely, $\mu_B \rightarrow [0,1]$, where B is a label identifying the function. We call the label B a fuzzy set when we have a certain interpretation of this label. Let us see an example.

Let S be the set of all nonnegative integers. We interpret S to be the set that includes landscape vegetation cover under grazing and dry conditions (Wu *et al.*, 1996). Let us consider if we can define a subset that implies "good vegetation cover." No doubt cover above 80 percent is good, whereas cover of about 60 percent is not so good---but not too bad either. There is no definite criterion that separates good and not good vegetation covers. Now, we define a function that corresponds to the concept of "goodness of vegetation cover." That is, for $x \in S$, μ_B is defined to show degree of relevance of the cover percentage x to the concept "good." Thus, $\mu_B(80) = 1$ to indicate that 80 percent

cover is good, while $\mu_B(30) = 0$ indicating that 30 percent cover is not good. For other cover percentages, we might use

$$\mu_B(x) = \begin{cases} 1 & \text{if } x \geq 70 \\ (x - 50)/20 & \text{if } 50 < x < 70 \\ 0 & \text{if } x \leq 50. \end{cases}$$

Thus, a fuzzy set B is a grade of relevance of elements to a concept represented by the label B . In terms of the function $\mu_B(x)$, it is a generalization of the characteristic function $f_A(x)$ for an ordinary (crisp) subset.

Mathematically, B is nothing but a label attached to the function μ_B , but an interpretation as in the above example enables us to call it a fuzzy set, a subset without a clearly defined boundary. The function μ_B is the membership function for the fuzzy set B . If $\mu_B(x) = 1$, then x certainly belongs to the fuzzy set B ; if $\mu_B(x) = 0$, then x does not belong to B at all; if $0 < \mu_B(x) < 1$, then belongingness of x to B is ambiguous. If $\mu_B(x_1) > \mu_B(x_2)$, then the relevance of x_1 to the concept represented by B is greater than the relevance of x_2 .

Every crisp set can be regarded as a fuzzy set, since the characteristic function of a crisp set is regarded as its membership function. In particular, the entire set S is the fuzzy set whose membership function is $\mu_S(x) = 1$ for all $x \in S$. In the same way, the empty set \emptyset is the fuzzy set with $\mu_\emptyset(x) = 0$ for all $x \in S$.

There are many ways to define fuzzy set operations (Klir and Folger, 1988; Kaufmann and Gupta, 1991; Li, 1998b). Here we discuss only the basic fuzzy set operations. Consider a finite universe $S = \{x_1, x_2, \dots, x_n\}$ and let $A \subset S$ be a fuzzy set. Its membership values are expressed by a simplified notation:

$$A = \mu_A(x_1)/x_1 + \mu_A(x_2)/x_2 + \dots + \mu_A(x_n)/x_n.$$

Note that here the symbol “+” does not refer to ordinary addition.

For example, let $S = \{1, 2, 3, 4, 5\}$ and consider a fuzzy set given by

$\mu_A(1) = 0$, $\mu_A(2) = 0.5$, $\mu_A(3) = 0.8$, $\mu_A(4) = 1$, and $\mu_A(5) = 0.2$. Using the above notation, A would be specified by

$$A = 0/1 + 0.5/2 + 0.8/3 + 1/4 + 0.2/5.$$

Equality of two fuzzy sets is defined by the equality of their membership functions. That is, for two fuzzy sets $A, B \subset S$

$$A = B \Leftrightarrow \mu_A(x) = \mu_B(x) \quad \text{for all } x \in S.$$

Inclusion of fuzzy sets is defined by the ordering of their membership functions:

$$A \subset B \Leftrightarrow \mu_A(x) \leq \mu_B(x) \quad \text{for all } x \in S.$$

The union and intersection of two fuzzy sets are defined by maximum and minimum, respectively:

$$\begin{aligned} \text{Union:} \quad & \mu_{A \cup B}(x) = \max[\mu_A(x), \mu_B(x)], \\ \text{Intersection:} \quad & \mu_{A \cap B}(x) = \min[\mu_A(x), \mu_B(x)]. \end{aligned}$$

The complement of a fuzzy set A which is denoted by A^c , is defined as follows

$$\text{Complement:} \quad \mu_{A^c}(x) = 1 - \mu_A(x)$$

For example, let $S = \{1,2,3,4,5\}$ again and assume that fuzzy sets A and B are given by

$$\begin{aligned} A &= 0/1 + 0.5/2 + 0.8/3 + 1/4 + 0.2/5, \\ B &= 0.9/1 + 0.4/2 + 0.3/3 + 0.1/4 + 0/5, \\ A \cup B &= 0.9/1 + 0.5/2 + 0.8/3 + 1/4 + 0.2/5, \\ A \cap B &= 0/1 + 0.4/2 + 0.3/3 + 0.1/4 + 0/5, \\ A^c &= 1/1 + 0.5/2 + 0.2/3 + 0/4 + 0.8/5. \end{aligned}$$

Many extensions to the preceding fuzzy set operations can be found in Kaufmann and Gupta (1991). These fuzzy set operations can be used to aggregate different ecological indicators during ecological risk analysis. We propose to study different possible concepts and methods for using such fuzzy set operations to integrate different source/scale information.

We propose to use an adaptive fuzzy system modeling framework as a new approach to the characterization of ecological landtype zonation in terms of multi-tier data. This system will involve several different data analysis methods (Figure 1a) including a pure fuzzy logic system (IF-THEN rules) and fuzzy classification linking a training (or learning) algorithm (back-propagation algorithm) that have been developed recently in fuzzy research community (Bandemer and Nather, 1992; Ichihashi and Turksen, 1995). This system combines linguistic and numerical information from expert knowledge, field sampling and intensive study data, and remote sensing. In data preprocessing and algorithmic data analysis, we use traditional statistical data exploration methods, fractals and geostatistics, and fuzzy clustering to sort and analyze any available numerical information that we are developing during recent years. The IF-THEN rule based data analysis we use to quantify expert knowledge and experience. In this approach, we apply Sugeno and Yasukawa's modeling procedure (Sugeno and Yasukawa, 1993) for our multi-input and single-output system. Neuro-fuzzy data analysis is an important step for our adaptive fuzzy modeling approach to characterizing ecological landtype zonations that helps us to understand ecological mechanisms, extracting knowledge, and generating rules in fuzzy systems based on incomplete knowledge and uncertain sampling data (see Figure 1b). Although fuzzy classifications have been used in vegetation (Banyikwa *et al.*, 1990) and Landsat TM data (Lindsey *et al.*, 1992) classification, downscaling (Bardossy,

1994) and GIS (Saint-Joan and Desachy, 1995), a new hierarchical fuzzy c-means clustering algorithm and its generalization (the ISODATA methods) will be developed under this adaptive fuzzy modeling approach framework for addressing ecological landtype zonations with spatial orders by using multi-tier data, quantitative or qualitative.

Integrated ecological assessment at multiple scales is complex. Uncertainty is a major issue in all application of risk assessment, but it presents a particular problem for ecological risk assessment due to the inherent variability of biological and ecological systems. Errors in the information used for assessments can arise from errors in measurements (e.g., toxicology index, environmental concentrations); extrapolation from species to species or from one set of conditions to another, and other ways in which expert judgment is used; and the assumptions that are made and the arbitrary thresholds that are set at different stages. There is nothing to be gained from ignoring the potential inaccuracies these errors may cause or the substantial risks of misjudging the scale of problems if they are not examined, even imperfectly. Increasing awareness of ecological issues has emphasized the need for improved ecological risk assessment methodology. In this proposal, we are proposing an adaptive fuzzy statistical and modeling approach as a promising tool for integrated ecological risk assessment. This approach may shed new light on dealing with uncertainty explicitly in risk assessment. Fuzzy mathematical and statistical theory will help exploit these different kinds of ecological realities under conditions of uncertain and incomplete information.

References:

Allen, T. F. H. and Hoekstra, T. W. (1992). *Toward a Unified Ecology*. Columbia University Press, New York.

Bandemer, H. and Nather, W. (1992). *Fuzzy Data Analysis*. Kluwer, Dordrecht.

Banyikwa, F. F., Feoli, E. and Zuccarello, V. (1990). Fuzzy set ordination and classification of Serengeti short grasslands, Tanzania. *J. Vegetation Sciences*, **1**, 97--104.

Bardossy, A., (1994). Downscaling from GCMs to local climate through stochastic linkages. In: G. Paoli (ed.), *Climate Change, Uncertainty, and Decision-Making*, IRR and IGBP, pp. 33--46.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**, 192-236.

Bradbury, R. H., Reichlet, R. E. and Green, D. G. (1984). Fractals in ecology: methods and interpretation. *J Marine Ecol. Prog. Ser.* **14**, 295-296.

Brown, P. J., Le, N. D. and Zidek, J. V. (1994). Multivariate spatial interpolation and exposure to air pollutants. *J Canadian J. of Statistics*. **22**, 489-516.

Clark, I., Basinger, K. L., and Harper, W. V. (1987). MUCK---A novel approach to cokriging. In *Geostatistical, Sensitivity, and Uncertainty Methods for Ground-Water Flow and Radionuclide Transport Modeling*, B. E. Buxton, ed. Battelle Press, Columbus Ohio. pp. 473--493.

Cox, L. H. and Piegorsch, W. W. (1994). Combining environmental information, open areas of environmental research in ecosystem monitoring, epidemiology, and data reporting. Manuscript.

Cressie, N. (1991). *Statistics for Spatial Data*. Wiley, New York.

Filipponi, D., Patil, G. P., and Taillie, C. (1998a). Use of indicator kriging to improve spatial coherence of thematic maps. Technical Report 98-0103, Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA.

Filipponi, D., Patil, G. P., and Taillie, C. (1998b). Spatial and spectral classification of compressed image data for landscape analysis. Technical Report 98-0104, Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and*, **6**, 721-741.

Gotway, C. A. and Cressie, N. (1993). Improved multivariate prediction under a general linear model. *J. Multivariate Analysis*, **45**, 56-72.

Guttorp, P. and Sampson, P. D. (1994). Methods for estimating heterogeneous spatial covariance functions with environmental applications. In *Handbook of Statistics*, Vol. 12, G. P. Patil and C. R. Rao, eds. Elsevier. pp. 661-689.

Guyon, X. (1995). *Random Fields on a Network*. Springer-Verlag, New York.

Hall, P. and Wood, A. (1993). On the performance of box-counting estimators of fractal dimension. *J Biometrika*, **80**, 246-252.

Handcock, M. S. (1989). *Inference for spatial Gaussian random fields when the objective is prediction*. Ph.D. thesis, Department of Statistics, University of Chicago.

Handcock, M. S. (1995). Discussion of "Multivariate Imputations" in "Cross Sectional Analysis of Health Effects Associated with Air Pollution." *Environmental and Ecological Statistics*, **2(3)**, 206-210.

- Handcock, M. S., and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, **35**, 403-410.
- Handcock, M. S. and Wallis, J. (1994). An approach to statistical spatial-temporal-modeling of meteorological fields (with discussion). *Journal of American Statistical Association*, **89**, 363-378.
- Hastings, H. M. and Sugihara, G. (1993). *Fractals: A User's Guide for the Natural Sciences*. Oxford University Press, Oxford.
- Hershey, R. R. (1996). Understanding the spatial distribution of tree species in Pennsylvania. In: H. T. Mowrer, R. L. Czaplewski, and R. H. Hambre (eds.) *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium*.} General Technical Report RM-GTR-277. U.S. Dept. of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado. pp. 73--82.
- Ichihashi, H. and Turksen, I. B. (1995). Neuro-fuzzy data analysis and its future directions. In: Proceedings of 1995 IEEE Int. Conf. On uzzy Systems, Yokohama, Japan, IEEE Press, pp. 1919--1925.
- Johnson, G. D., Myers, W. L., Patil, G. P., and Walrath, D. (1996). Multiscale analysis of the spatial distribution of breeding bird species richness using the echelon approach. Technical Report 98-1101, Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA.
- Johnson, G. D. and Patil, G. P. (1997a). Quantitative multiresolution characterization of landscape patterns for assessing the status of ecosystem health in watershed management areas. *Ecosystem Health* (invited paper submitted).
- Johnson, G. D., Myers, W. L., and Patil, G. P. (1997b). Stochastic generating models for simulating hierarchically structured multi-cover landscapes. *Landscape Ecology* (invited paper submitted).
- Johnson, G. D., Myers, W. L., Patil, G. P., and Taillie, C. (1997c). Quantitative characterization of hierarchically scaled landscape patterns. *J. Statistical Planning and Inference* (invited paper submitted).
- Johnson, G. D., Myers, W. L., Patil, G. P., and Taillie, C. (1997d). Multiresolution fragmentation profiles for assessing hierarchically structured landscape patterns. *Ecological Modeling* (submitted).
- Johnson, G. D., Tempelman, A. K. and Patil, G. P. (1995). Fractal based methods in ecology: a review for analysis at multiple spatial scales. *Coenosis*, **10**, 123-131.

- Jones, R. H. and Vecchia, A. V. (1993). Fitting continuous ARMA Models to unequally spaced spatial data. *Journal of American Statistical Association*, **88**, 947-954.
- Kaufmann, A. and Gupta, M. M. (1991). *Introduction to Fuzzy Arithmetic, Theory and Applications*. Van Nostrand, New York.
- Kepner, W. G., Jones, K. B., Chaloud, D. J., Wickham, J. D., Riitters, K. H. and O'Neill, R. V. (1995). Mid-Atlantic Landscape Indicators Project Plan, Environmental Monitoring and Assessment Program. EPA 620/R--95/003.
- Kitanidis, P. K. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research*, **22**, 499-507.
- Klir, G. J. and Folger, T. (1988). *Fuzzy sets, Uncertainty, and Information*. Prentice-Hall, Englewood Cliffs, NJ.
- Krummel, J. R., Gardner, R. H., Sugihara, G. , O'Neill, R. V. and Coleman, P. R. (1987). Landscape patterns in a disturbed environment. *Oikos*, **48**, 321-324.
- Le, N. D. and Zidek, J. V. (1992). Interpolation with uncertain spatial covariances: a Bayesian alternative to Kriging. *J. Multivariate Analysis*, **43**, 35-374.
- Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, second edition. Wiley, New York.
- Levin, S. (1992). The problem of pattern and scale in ecology. *Ecology*, **73**, 1943-1967.
- Li, B. L. (ed.). (1996). Fuzzy Modeling in Ecology. Special Issue of *Ecological Modeling*, **90**, 109--186.
- Li, B. L. (1998a). The use of fuzzy-set theory for ecological assessments. In: P. Bourgeron, M. Jensen, and G. Lessard (eds.), *Integrated Ecological Assessment Protocols Guidebook*. Springer-Verlag (in press).
- Li, B. L. (1998b). *Mathematical Modeling in Ecology*. Springer-Verlag (in preparation).
- Li, B. L. and Yeung, A. T. (1994). An adaptive fuzzy modeling framework for characterization of subsurface contamination. In: Proceedings of 1994 1st Int. Joint Conf. of North American Fuzzy Information Processing Society Biannual Conf., the Industrial Fuzzy Control and Intelligent Systems Conf., and NASA Joint Technology

Workshop on Neural Network and Fuzzy Logic, San Antonio, TX, IEEE Press. pp. 194-195.

Lindsey, S. D., Gunderson, R. W. and Riley, J. P. (1992). Spatial distribution of point soil moisture estimates using Landsat TM data and fuzzy-c classification. *Water Resources Bulletin*, **28**, 865--875.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis and Missing Data*. Wiley, New York.

Loader, C. and Switzer, P. (1992). Spatial covariance estimation for monitoring data. In *Statistics in Environmental and Earth Sciences*, P. Guttorp and A. Walden, eds. Griffen, London. pp. 52--69.

Loehle, C. and Li, B. (1995). Statistical properties of ecological and geological fractals. *Ecological Modeling*(in press).

Loehle, C. and Wein, G. (1994). Landscape habitat diversity: a multiscale information theory approach. *Ecological Modeling*, **73**, 311-329.

Luo, Z. (1998). A patch-based multiscale image simplification method. Technical Report 98-0102, enter for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA.

Mandelbrot, B. B. (1983). *The Fractal Geometry of Nature*. Freeman, New York.

Mardia, K. V. and Goodall, C. R. (1993). Factorized models in spatial modeling. In *Multivariate Environmental Statistics*, N. K. Bose, G. P. Patil and C. R. Rao, eds. North-Holland, New York. pp. 661--689.

Matern, B. (1986). *Spatial Variation*, 2nd edition. Lecture Notes in Statistics, Number 36. Springer-Verlag, Berlin.

Milne, B. T. (1988). Measuring the fractal geometry of landscapes. *Appl. Math. Comp.*, **27**, 67-79.

Myers, D. E. (1991). Pseudo Cross-variograms, positive definiteness, and cokriging. *Mathematical Geology*, **23**, 805-816.

Myers, W. L. (1997). PHASE approach to remote sensing and quantitative spatial data. Technical Report ER9710, Environmental Resources Research Institute, Pennsylvania State University, University Park, PA.

- Myers, W. L., Patil, G. P., and Joly, K. (1997). Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics*, **4**,131--152.
- Myers, W. L., Patil, G. P., and Taillie, C. (1997a). Exploring landscape scaling properties through constrictive analysis. *J. Statistical Planning and Inference*(invited paper submitted).
- Myers, W. L., Patil, G. P., and Taillie, C. (1997b). PHASE formulation of synoptic multivariate landscape data. *J. Statistical Planning and Inference* (invited paper submitted).
- Myers, W. L., Patil, G. P., and Taillie, C. (1997c). Adapting quantitative multivariate geographic information system data for purposes of sample design: the PHASE approach. In: *Multivariate, Design, and Sampling.*} Marcel Dekker, New York. (invited paper submitted).
- Ogata, Y. and Katsura, K. (1991). Maximum likelihood estimates of the fractal dimension for random spatial patterns. *Biometrika*, **78**, 463-47.|
- Olofsen, E., Degoede, J. and Heijungs, R. (1992). A maximum likelihood approach to correlation dimension and entropy estimation.*Bull. of Mathematical Biology*, **54**, 45-58.
- O'Neill, R. V., Johnson, A. R. and King, A. W. (1989). A hierarchical framework for the analysis of scale. *Landscape Ecology*, **3**, 193-205.
- O'Neill, R. V., Krummel, J. R., and others (1988). Indices of landscape pattern. *Landscape Ecology*, **1**, 153-162.
- Patil, G. P. (1991). Encountered data, statistical ecology, environmental statistics, and weighted distribution methods. *Environmetrics*, **2**, 377-423.
- Patil, G. P. (1997).
Statistical ecology and environmental statistics for cost effective ecological synthesis and environmental analysis. In: *Modern Trends in Ecology and Environment*. Backhuys Publishers, The Netherlands.
- Patil, G. P., Myers, W. L., Luo, Z., Johnson, G. D., and Taillie, C. (1997).
Multiscale assessment of landscapes and watersheds with synoptic multivariate spatial data in environmental and ecological statistics. Special issue of *Mathematical and Computer Modeling on Stochastic Models in Mathematical Biology*. (invited paper submitted).
- Patil, G. P. and Taillie, C. (1998a). Parametric families of transition matrices for hierarchical landscape fragmentation assessment. Technical Report 98-1201,

Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA.

Patil, G. P. and Taillie, C. (1998b). Assessment of hierarchical landscape fragmentation with Kullback-Leibler distance: Issue of statistical significance versus scientific importance. Technical Report 98-0202, Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA.

Patil, G. P. and Taillie C. (1991). Performance of the largest order statistics relative to the sample mean for the purpose of estimating a population mean. *Bull. International Statistical Institute*, **54**, 1-21.

Pesin, Ya and Tempelman, A. (1995). Correlation dimension of measures invariant under group actions. *Random and Computational Dynamic*, **3**, 137-156.

Piotrowski, J. A., Bartels, F., Salski, A., and Schmidt, G. (1994). Fuzzy kriging of imprecise hydrogeological data. In: *Proceedings of IMAG 94 Conference*, Mont Tremblant, Quebec, Canada. pp. 2282-2288.

Piotrowski, J. A., Bartels, F., Salski, A., and Schmidt, G. (1996). Geostatistical regionalization of glacial aquitard thickness in Northwestern Germany based on fuzzy kriging. *Mathematical Geology*, **28**, 437--452.

Possolo, A. (1986). Estimation of binary Markov random fields. Technical Report number 77, Department of Statistics, University of Washington.

Prum, B. and Fort, J. C. (1991). *Stochastic Processes on a Lattice and Gibbs Measures*. Kluwer, Dordrecht.

Riitters, K. H., O'Neill, R. V., Hunsaker, C. T., Wickham, J. D., Yankee, D. H., Timmins, S. P., Jones, K. B. and Jackson, B. L. (1995). A factor analysis of landscape pattern and structure metrics. *Landscape Ecology*, **10**, 23-29.

Riitters, K. H. and Wickham, J. D. (no date). *A Landscape Atlas of the Chesapeake Bay Watershed*. Tennessee Valley Authority, Norris, TN.

Saint-Joan, D. and Desachy, J. (1995). A fuzzy expert system for geographical problems: an agricultural application. In: *Proceedings of 1995 IEEE Int. Conf. on Fuzzy Systems*, Yokohama, Japan. IEEE Press. pp. 469--476.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of American Statistical Association*, **87**, 108-119.

- Schowengerdt, R. A. (1997). *Remote Sensing: Models and Methods for Image Processing*. Academic Press, NY.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.
- Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of Royal Statistical Society, Series B*, **55**, 3-23.
- Sugeno, M. and Yasukawa, T. (1993). A fuzzy-logic-based approach to qualitative modeling. *IEEE Trans. Fuzzy Systems*, **1**, 7--31.
- Taylor, C. C. and Taylor, S. J. (1991). Estimating the dimension of a fractal. *Journal of Royal Statistical Society, Series B*, **53**, 353-364.
- Theiler, J. (1990). Statistical precision of dimension estimators. *Phys. Rev. A*, **41**, 3038-3051.
- Vecchia, A. V. (1985). A general class of models for stationary two-dimensional random processes. *Biometrika*, **72**, 281-291.
- Wiens, J. A. (1995). Landscape mosaics and ecological theory. In *Mosaic Landscapes and Ecological Processes*, L. Hansson, L. Fahrig, and G. Merriam, eds. Chapman and Hall.
- Winkler, G. (1995). *Image Analysis, Random Fields, and Dynamic Monte Carlo Methods*. Springer-Verlag, New York.
- Wu, H., Li, B. L., Stoker, R. and Li, Y. (1996). A semi-arid grazing ecosystem simulation model with probabilistic and fuzzy parameters. *Ecological Modeling*, **90**, 147--160.
- Younes, L. (1988). Estimation and annealing for Gibbsian fields. *Annales de l'Institut Henri Poincare*, **24**, 269-294.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, **8**, 338--353.

