

# STATISTICAL METHODS FOR ECOLOGICAL ASSESSMENT OF RIVERINE SYSTEMS BY COMBINING INFORMATION FROM MULTIPLE SOURCES

Mark S. Handcock (UW), Joseph Sedransk (Case Western Res. Uni.), Anthony Olsen (EPA)  
Mark S. Handcock, Department of Statistics, University of Washington, Seattle, WA 98195-4322

**KEY WORDS:** spatial statistics, GIS, hierarchical models, relative distribution, EMAP

## Abstract

The assessment of environmental risk and the evaluation of environmental policies increasingly require accurate and relevant information about the environment. For policy makers and stakeholders to evaluate possible policy changes, an understanding of the cause-and-effect relationship between the alternative policies and their environmental and social outcomes is essential. Addressing these questions requires information from diverse sources to be collected, organized, and combined. Evolutionary improvements in Geographic Information Systems (GIS) now routinely allow the management and mapping of spatial-temporal information. However there is a dearth of statistical methodology, not only to represent the complexities of the information, but also to allow the uncertainty of the resulting inference to be quantified. The development of statistical models to combine information of different types and spatial support is of vital importance to environmental social science.

The objective of the paper is to improve understanding of the biological integrity of stream and river systems in the United States Mid-Atlantic Region by combining information from separate monitoring surveys, available contextual information on hydrologic units and remote sensing information.

An essential initial step in this process is the compilation of a geographic information system combining standardized data sets available on the WWW in a form that facilitates modeling. The research is based in part on information underlying the Landscape Atlas of the mid-Atlantic region produced by the US Environmental Monitoring and Assessment Program (EMAP). The construction of the Atlas is an extraordinary achievement. We also combine information from two overlapping separate monitoring surveys, the EMAP Stream and River Survey and

the Maryland Biological Streams Survey.

The heart of the research program outlined in this paper is to complement the mapping presented in the Atlas with new hierarchical spatial statistical models for environmental indicators on the streams and rivers that capture the spatial variation in the measures. These models are used to estimate the indicators through the riverine system based on the information from multiple sources and aggregate scales. We also quantify the uncertainty in the estimates and develop methods to visualize the resulting estimates and uncertainties.

In environmental research, regional differences or temporal changes are a common focus of study. While means and variances are typically the basis for statistical methods used in this research, the underlying environmental questions involve properties of spatial distributions that are not well captured by these summary measures. Examples include the current questions of environmental justice, risk assessment, and the impact of environmental change on survival and health. The distributional differences that animate these debates are complex. They comprise the usual mean-shifts and changes in variance, but also more detailed comparisons of changes in the upper and lower tails of the distributions. Environmental data contains a wealth of this distributional information, but the traditional methods of analysis leave most of it untapped.

The third objective of the paper is to present a general framework for comparative distributional analysis. The methods are based on the *relative spatial distribution*. The spatial models developed are used to predict spatial distributions and relative spatial distributions. These methods are then used to combine county-level social science data with the different sources of environmental data. This makes it possible to investigate questions of environmental justice in a systematic and rigorous way.

## 1. Introduction

Policy decisions by governmental and industrial organizations increasingly require accurate information about the environment. Information is needed on the status of, and trends in, basic environmental conditions in order to develop environmentally appropriate policies. In addition, proper planning requires an understanding of the interactions between social and physical environmental processes. For policy makers and stakeholders to evaluate the range of options, a framework for evaluating the potential impact of alternative policies on environmental and social outcomes is essential.

While much of the initial interest in the population-environment interaction focused on the impact of humans on their environment, over the last decade there has been increasing interest in the reverse question: the impact of environmental degradation on the health and wellbeing of populations. This interest has been motivated by the recognition that certain types of environmental degradation, such as incinerators or polluting industries, tend to have a disproportionate impact on the local population. The spatial distribution of such environmentally undesirable activities has therefore become an important public health issue. To the extent that these activities are spatially concentrated in economically disadvantaged communities, this raises questions of *environmental justice*: “the fair treatment for people of all races, cultures, and incomes, regarding the development of environmental laws, regulations, and policies” (EPA 1993).

Evaluating questions of environmental justice requires information from diverse sources to be collected, organized, and combined. For example, consider investigating the relationship between the level of pollution of streams and the economic status of the surrounding residents. The pollution level can be assessed based on an environmental monitoring survey on selected stream sites. These will need to be adjusted for biophysical cofactors, such as soils and land cover, that effect pollution levels, but are not necessarily evidence of environmental injustice. Such information may be available in separate monitoring surveys, contextual spatial databases and remote sensing sources. The characteristics of the human population usually need to be determined by additional sources of data, such as social surveys or census information.

Combining all of these different sources of information has been facilitated by dramatic improvements in Geographic Information Systems (GISs). These now routinely allow the management, dis-

play and mapping of spatial-temporal information. More importantly they allow the “spatial indexing” of multiple types of information over the study region. To a large extent, however, these new facilities are descriptive, rather than analytical. The methodology for analyzing the interrelationships between these multiple sources of information remains underdeveloped.

To move beyond descriptive analysis, statistical methods are necessary. Statistical models make it possible to investigate hypotheses regarding the complex relationships within and between the different levels of analysis. They also provide the framework for evaluating the findings. This enables researchers to draw inferences about characteristics of the phenomena most directly relevant to the environmental social science questions, and to quantify the uncertainty of the resulting inference. The development of statistical models to combine information from the different sources of spatial data is of vital importance to environmental social science.

With respect to environmental data, the current situation is ironically one of both wealth and poverty. The wealth arises from the many forms of remote sensing and spatial extant data that provide coverage of the regions of interest. The poverty arises from the lack of longitudinal data with spatial extent and representativeness. Most environmental monitoring programs are subject to scientific, political, ethical and cost considerations, and these have resulted in an unwieldy patchwork of spatial-temporal information.

A notable exception is provided by the US Environmental Monitoring and Assessment Program (EMAP). EMAP is designed to address questions about the current status, changes, and trends in indicators of ecological condition of the nation’s ecosystems (Messer *et al.* 1991). The EMAP Landscape Atlas of the mid-Atlantic region (Jones *et al.* 1997) represents one of the first regional-scale ecological assessments that incorporates multiple sources of non-EMAP data. It is an extraordinary achievement in combining multiple data sources, and will clearly be the basis for the development of environmental statistics in the future.

The research plan here is one of the first to take advantage of the opportunities that EMAP Landscape Atlas of the mid-Atlantic region provides. As the first part of our research plan, we seek to complement the mapping presented in the Atlas with new statistical models for combining information from multiple sources to gain insight into the complex relationship between environmental factors.

An excellent review on the state of statistical

methods for combining environmental data is given by Cox and Piegorsch (1996), Piegorsch and Cox (1996), and Cox (1998). For a discussion of issues for combining information from different agricultural experiments, see Besag and Higdon (1999). Handcock, Huovilainen and Rendall (2000) develop statistical models to combine social survey data with population-level census information. They show how likelihood-based inference for models based on survey data can be extended to include census and registry information. The paper also reviews methods used by social scientists for combining data of this type.

As the second part of our research plan, we develop preliminary approaches for integrating social science data into these environmental models. Here we draw on methods developed for measuring the spatial distribution of environmental indicators (Handcock 1999), relative distribution methods for measuring economic inequality (Handcock and Morris, 1999), and spatial-temporal models for community economic status (McLaughlin and Handcock, 1999).

This paper is an attempt to answer the call by Cox (1998) for "... the development of a theoretical framework for integrating spatial, and [probability]-sample methods for environmental assessment, new methods and extensions of existing methods for combining spatial data collected at different aggregate scales, ..., and hierarchical methods that enable combination and intercomparison of different environmental studies."

## 2. Specific Evaluation of Ecological Indicators of Streams

In this section we describe the region of study, and the component sources of information that will be combined. Streams form a continuous network embedded in the watersheds they drain. The conditions of the watersheds and ecoregion through which the streams run is reflected in the quality of the ecological indicators of the streams (Herlihy *et al.*). Any modeling approach must respect this fundamental tenet of limnology – that these conditions depend on the network structure of the streams, and the fact that water moves continuously downstream (Vannote *et al.* 1980).

### 2.1. The Study Area: The United States Mid-Atlantic Region

The study area is the mid-Atlantic region of the eastern United States and its watersheds. This region is defined by the EPA to be the land and

near coastal area that includes all of EPA Region III and parts of Regions II and IV. The region extends from southern New York into northeastern North Carolina. The region includes EPA Region III (i.e., Pennsylvania, West Virginia, Maryland, Delaware, and Virginia); the Susquehanna and Allegheny River basins, which extend into New York; the Delaware River basin, which extends into New Jersey; and the Chowan-Roanoke and Neuse-Pamlico basins, which extend into North Carolina. The mid-Atlantic region encompasses the area from the mid-Appalachian highlands to the estuaries.

This region was chosen for a number of scientific and practical reasons. The mid-Atlantic region has been extensively studied by the EPA and other scientific groups. The region is one of the most data-rich areas in the country, in part because of its dense population and proximity to Washington, D.C.

Most of the component surveys, especially those addressing water-related concerns, further partition the region into the USGA defined hydrologic accounting units. Roughly speaking, these units follow watershed boundaries - areas of land that is drained by a single stream, river, lake or other body of water. Hence watersheds are the natural units for the environmental analysis based on riverine systems. We note that the hydrologic units are not, strictly speaking, watersheds in the sense of topographically-defined catchment areas. Following the usage in the component surveys, we shall use these as the basic unit of analysis and for simplicity refer to them as watersheds. The methods developed here can equally be applied to other partitions of the region - indeed in Section 6 we consider counties.

One of the problems indicated by the hydrography of the region is that of using naturally-defined units such as watersheds to assess environmental conditions over politically-defined units such as counties or states. Individual watersheds can lie in two or more states or counties. This issue of *misalignment* is fundamental one in environmental statistics. Excellent progress has been made on these issues by Mugglin, Carlin and Gelfand (2000), Mugglin and Carlin (1998) and Mugglin *et al.* (1999). These papers are mainly concerned with variables that are aggregated over differing sets of units. To the extent that the models proposed here are used to change units, they can be regarded as variants of the modeling approach of the above papers. The models described below focus on combining multiple sources of information and are based on explicitly modeling the underlying riverine systems.

## 2.2. Sources of Information on the Mid-Atlantic Region Combined

The paper is the first part of what is to be a case-study for the combination of information from multiple sources. The accessibility of the case-study will be improved by basing it on readily available, compatible and mature data sets. Most of the data are available in formats readily adapted to standard GIS and statistical analysis packages (e.g., ARC/INFO, SAS and S-PLUS).

The project uses as its foundation the work of the following two interrelated initiatives:

1. The EPA/ORD Mid-Atlantic Integrated Assessment (MAIA)

This multi-year project undertakes a ecosystem-based evaluation of the mid-Atlantic region and its watersheds. MAIA incorporates numerous state, regional, and national environmental monitoring programs into an assessment process specifically targeted to the management needs of EPA Region III. Examples of programs with which there are specific cooperative efforts include the Environmental Monitoring and Assessment Program (EMAP), the mid Atlantic Highlands Assessment, the National Biological Service's Gap Analysis Program, the Chesapeake Bay Program, the Delaware Estuary Program, the Virginia Coastal Bays Program, the U.S. Geological Survey's National Water Quality Assessment Program, the Forest Service's Forest Inventory and Analysis Program, and the National Oceanic and Atmospheric Administration's Coastal Change Analysis Program. For an description of MAIA see <http://www.epa.gov/emap/maia>.

2. An Ecological Assessment of the United States Mid-Atlantic Region: A Landscape Atlas

As described above the Atlas is an assessment of relative ecological conditions across the mid-Atlantic region, and was published in April 1998. The Atlas identifies, with never-before achieved detail and comparability, patterns of land cover and land use across the region.

The report is based on data from satellite imagery and spatial databases on biophysical features such as soils, elevation, and human population patterns. It compares nine landscape indicators on a watershed-by-watershed basis for the lower 48 states (at a relatively coarse-scale resolution of 1 km), placing the mid-Atlantic region in the context of the rest of the country. Using finer-scale spatial resolution (e.g., 30-90

meters), the report then analyzes and interprets environmental conditions of the 125 watersheds in the mid-Atlantic region based on 33 landscape indicators. Results are presented relative to four general themes identified by stakeholders in the region: (1) people (potential human impacts), (2) water resources, (3) forests (forest habitat), and (4) landscape change.

The data underlying this Atlas is publicly available and initial analysis indicates that it proves invaluable for the project. For an description of the Atlas, see [http://www.epa.gov/emap/html/ma\\_atlas.html](http://www.epa.gov/emap/html/ma_atlas.html).

The specific components surveys that we use are:

- a **EMAP Mid-Atlantic Integrated Assessment (MAIA) Survey**(Larsen and Christie 1993).

We use the Stream and River Survey that has data on 100-200 sites from 1993-96. Some of the sites are repeat visits. For a description of the EMAP Surface Waters Mid-Atlantic Streams 1993-96 data set see <http://www.epa.gov/emap/html/dataI/surfwatr/data/mastreams>.

- b **Maryland Biological Streams Survey (MBSS)**

(Heimbuch, Seibel, Wilson and Kazyak 1998)

To provide much needed information about the ecological consequences of acid deposition and other human-related impacts, the Maryland Department of Natural Resources designed the MBSS. The MBSS is a long-term monitoring program designed to describe the current status of aquatic biota, physical habitat and water quality in first, second and third order non-tidal streams within the state of Maryland. The MBSS was implemented as a three-year study in 1995. Sampling is probability-based, and stratification is based on stream order and drainage basin. Approximately 1000 sites were sampled during 1995-1997. A "State of the Streams" report which summarizes the initial round of the MBSS has recently be completed.

For a description of the MBSS see <http://www.dnr.state.md.us/streams>.

- c **EMAP Mid-Atlantic Landscape Indicators**(Kepner, Jones, Chaloud and Wickham 1995).

This project supplies landscape classification data with a resolution of 30 meters based on

satellite imagery. For a description of the mid-Atlantic Landscape Indicators, see [http://www.epa.gov/emfjulte/html/ma\\_atlas.html](http://www.epa.gov/emfjulte/html/ma_atlas.html).

These surveys are supplemented by:

d **EMAP Streams network database (RF3)** (Dewald and Olsen 1994).

This is the primary database for the locations of the rivers and streams in the mid-Atlantic region. We use the River Reach File Version 3, derived from the U.S. Geologic Survey Digital Line Graph - streams, 1:100,000-scale. The study use all first- through third-order (i.e. wadeable) streams. There are 230,400 kms of wadeable streams in the study region.

e **USDA Natural Resources Conservation Service soils database**(USDA 1994).

This data set is a digital general soil association map developed by the National Cooperative Soil Survey and distributed by the Natural Resources Conservation Service (formerly Soil Conservation Service) of the U.S. Department of Agriculture. It consists of a broad based inventory of soils and nonsoil areas that occur in a repeatable pattern on the landscape and that can be cartographically shown at the scale mapped.

For a description of the soils database, see [http://www.epa.gov/emfjulte/html/ma\\_atlas.html](http://www.epa.gov/emfjulte/html/ma_atlas.html).

Note that there is an overlap between the Maryland monitoring survey and the EMAP monitoring survey, as Maryland is a subset of the mid-Atlantic region. The density of monitoring sites from the Maryland survey is much greater than the EMAP survey in Maryland. This offers an opportunity to calibrate the EMAP information and use the Maryland information to build a better model of the local-scale structure of environmental indicators that can be leveraged over the mid-Atlantic region by combination with the EMAP monitoring survey. In addition the Landscape Atlas provides data on demographic characteristics, potential human impacts and use, water resources, forest habitat, and landscape change at the watershed level. These are also used.

### 2.3. Indicators of Environmental Condition

As detailed in above references, the two monitoring surveys used (EMAP and MBSS), collect an array of ecological indicators at each site. These

include various biotic, chemical, physical, riparian and watershed characteristics (Lazorahak *et al.* 1998). Fish species are particularly effective indicators of the condition of aquatic systems (Fausch *et al.* 1990). Human impact of streams and the environment effect key characteristics of aquatic ecosystems: water quality, habitat structure, hydrologic regime, and biologic interactions (Karr and Dudley 1981). Although the models can be adapted for alternative indicators, we describe models for fish assemblage metrics. Models for water chemistry indicators will also be considered in later work.

### 3. Methods for Combining Information from Multiple Surveys

In this section we propose statistical modeling methods for combining information from the surveys identified in the previous section. The purpose of these models is to create a stochastic representation for the measurement at each location on the riverine system. This representation forms the basis for the further modeling developments proposed in Sections 4, 5 and 6. The underlying modeling approach is hierarchical to allow complex structure to be represented by a hierarchy of relatively simple model specifications. The idea is to model the spatial dependence indirectly through latent stochastic processes. Related work is Besag (1974, 1975), Cressie (1995), Mollié and Richardson (1991), and Bernardinelli and Monotomoli (1992). Further references are given below.

Let  $R \subset \mathbb{R}^2$  be the set of locations on rivers and streams in the mid-Atlantic region. We define  $R$  operationally by those in the River Reach File Version 3 (RF3). Let  $W(x)$  represent the hydrologic unit (watershed) that the location  $x$  belongs, and  $\{W_i : i = 1, \dots, H\}$  represent the set of all hydrologic units. The units form a partition of  $R$ . Let  $Z(x)$  be a measure at each location in  $R$ . We consider a number of indicators of condition and stress related to fish or water chemistry. For simplicity of exposition, we shall consider linear formulation for  $Z(x)$  here. However the approach can be extended to cover much more general forms – we postpone this to Section 7. Throughout we will use as an example the fish index of biotic integrity (IBI) (Karr 1986), although difficulties with the representativeness of this measure may lead us to focus on the component fish assemblage metrics.

First we describe a model for the measure at each location. We write:

$$Z(x) = L(x)\beta_1 + C(x)\beta_2 + S(x)\beta_3 + \eta(W(x); \gamma)$$

$$+\phi(x; \nu) + \epsilon(x; \sigma) \quad (3.1)$$

where the first three terms capture variation due to differences in covariates, the  $\phi$  and  $\eta$  terms capture residual spatial variation and the last term the unexplained variation. The terms are:

- $L(x)$  row vector of location-specific covariates at location  $x$  and are potentially spatially varying in a neighborhood of  $x$ .

These measures are required to be known at each location in  $R$ . Examples, of covariates are latitude, longitude, and elevation. We will also include here indicators for the monitoring survey that provide the measurement. Hence systematic differences between the measurements of the surveys, here EMAP and MBSS, can be identified. These difference could be due to difference in the collection protocol or calibration differences. Clearly if more complicated calibration issues are envisaged they can also be added here or in the stochastic components. This set of covariates is restricted because we also need to know them at each value in  $R$ .

- $C(x)$  row vector of contextual covariates related to location  $x$ .

These measures are required to be known at each location in  $R$ , but can be areal. Examples, of covariates are characteristics of the reaches from RF3 such as stream order and stream level. Variables on demographic characteristics, air pollution, agricultural usage, human use index from the Landscape Atlas database are include here. This set is also restricted because we need to know them at each value in  $R$ .

The effects of political divisions can be investigated using contextual variables to indicate the location is within a given political division. The most direct example is the state (or states) that the watershed resides in. While the watershed does not necessarily respect state boundaries, state and local government regulations may directly influence the environmental condition and human activities. Hence the relative comparison of state-level effects if a very important way of *assessing the role of institutions* at the state level. This approach can be applied to other political division such as labor-market regions and counties (See Section 6).

- $S(x)$  row vector of complete coverage covariates related to location  $x$ .

These measures are assumed to be known at

each location in the region, including those at each location in  $R$ . Examples of covariates are biophysical features such as soil types from the USDA Natural Resources Conservation Service soils database, forest habitat, riparian cover, and human population patterns available from the Landscape Atlas database and other satellite-based landscape indicators.

Note that this division between location-specific, contextual, and complete coverage covariates is not absolute. Indeed similar modeling approaches can be applied to each of these components. The taxonomy is mainly to aid the identification of factors from the component surveys and to group the factors for interpretation. Each of these terms appears in a linear functional form with regression coefficient vectors  $(\beta_1, \beta_2, \beta_3)$ . The functional form of the covariate vectors themselves can be adapted so that this functional form is appropriate. Note the spatial variation terms represent the effects of unadjusted for, or unobserved, covariates as well as the effects of spatial proximity.

In addition to these effects we explicitly model the spatial variation between and within watersheds.

- $\eta(W(x); \gamma)$  latent between watershed effects. Each location within the same watershed receives the same effect. It represents the overall level differences between the units. We will consider two models for  $\{\eta(i; \gamma) : i = 1, \dots, H\}$ . The first represents them as fixed but unknown environmental characteristics (i.e., a classical “fixed effects” specification). This representation is of interest as the watershed are unchanging over the time scale of the study, and the watershed effects are themselves of direct scientific interest. Under the second specification the  $\{\eta(i; \gamma) : i = 1, \dots, H\}$  form a spatial lattice random field. The simplest model has the values independent of each other. We use a neighborhood-based lattice pairwise-difference model (Cressie 1993, Anselin and Florax 1995). Consider a neighborhood system for the watershed based on spatial contiguity, that is, units that share a common boundary are neighbors. We capture this effect with a class of non-stationary Gaussian intrinsic autoregressions (Besag *et al.* 1991, Bernardinelli and Monotomoli 1992). Let  $v_{ij}$  be prescribed non-negative weights, with  $v_{ij} = 0$  unless watersheds  $i$  and  $j$  are neighbors. Let  $\lambda_\gamma$  be a scale parameter to which we give a Gamma distri-

bution. The conditional distribution of  $\eta(i; \gamma)$  given the other effects in the watershed is specified to be Gaussian:

$$\eta(i; \gamma) \mid \eta(j; \gamma), i \neq j; \gamma \\ \sim N \left( \sum_{j \in \text{WN}_i} \frac{v_{ij}}{v_{i+}} \eta(j; \gamma), \frac{1}{\lambda_\gamma v_{i+}} \right)$$

where  $\text{WN}_i$  represents the watersheds  $j$  that are neighbors of  $i$  and  $v_{i+}$  is the sum over  $j \in \text{WN}_i$  of  $v_{ij}$ . The joint distribution of the between watershed effects is then an intrinsic Gaussian random field. The basic continuity scheme is contiguity, although alternative length schemes can clearly and fruitfully be considered, for example, length of common boundary, percentage of common boundary. The parameter  $\gamma$  includes  $\lambda_\gamma$  and others necessary to further specify the weights.

- $\phi(x; \nu)$  latent within watershed effects within the watershed of location  $x$ .

We model each  $\{\phi(x; \nu) : x \in W_i\}, i = 1, \dots, H$  as a spatial random field on the riverine system within each watershed. For simplicity, we shall initially specify that the inter-watershed dependence is captured by  $\eta(W(x); \gamma)$  and the within watershed spatial fields are independent between watersheds. This assumption can be relaxed if significant variability can be explained by doing so. The model within each watershed is a pairwise-difference model (Besag 1989). For example, consider a neighborhood system for  $x$  based on being on the same stream segment (according to RF3). That is, two location are neighbors if, and only if, they belong to the same stream segment. One would expect that, all else being equal, two location on the same stream would more likely have closer values on the measure than two locations on separate streams. We capture this effect with a modified class of non-stationary Gaussian intrinsic autoregressions. The riverine system represented by the RF3 is composed on a finite, albeit large, number of elements. Let  $s(x)$  be the stream element that  $x \in R$  is on, and there are a finite number  $M$ , say of such elements. We specify that  $\phi(x; \nu)$  is constant over the stream element  $s(x)$ . While a continuum random field on the riverine system is more appealing in principle, the hybrid irregular lattice version proposed below is designed to parsimoniously capture the stream-to-stream spatial variation. The main disadvantage of the continuum ap-

proach based on geostatistical models is the difficulty of specifying the variogram due to a lack of information at local scales. However for general processes the geostatistical approach has many advantages, as Zimmerman and Harville (1991) show with application to agricultural experiments. Progress on continuum models has been made (Kelsall and Wakefield 1997, Moller 1998, Ecker and Gelfand 1997, and Best, Ickstadt and Wolpert 1998). See Besag and Higdon (1999) for additional references and a discussion of these issues.

Returning to our model, prescribed non-negative weights,  $w(x, y) = W(s(x), s(y))$ , with  $w(x, y) = 0$  unless  $x$  and  $y$  are neighbors. As there are  $M$  stream elements the values of  $w(x, y)$  form a  $M \times M$  symmetric matrix  $W$ . Let  $\lambda_\nu$  be a scale parameter to which we give a Gamma distribution. The conditional distribution of  $\phi(x; \nu)$  given the other effects in the watershed is specified to be Gaussian:

$$\phi(x; \nu) \mid \phi(y; \nu), W(y) = W(x), y \neq x; \nu \\ \sim N \left( \sum_{s(y) \in N_1(x)} \frac{w(x, y)}{w(x, +)} \phi(y; \nu), \frac{1}{\lambda_\nu w(x, +)} \right)$$

where  $N_1(x)$  represents the stream elements that are neighbors of  $x$ , and  $w(x, +)$  is the sum over  $s \in N_1(x)$  of  $W(s(x), s)$ . The joint distribution of the within watershed effects for each stream element is then an intrinsic Gaussian random field. The simplest choice of the weights is  $w(x, y) = 1$  if  $x$  and  $y$  are on the same stream segment. However we can explore choosing weights proportional to those from a continuous geostatistically motivated semivariogram model to additionally capture the decay with distance between the locations (Raftery and Banfield 1991, and the discussion of Besag and Higdon 1999).

A number of neighborhood schemes could drive the spatial variation. For example:

1.  $N_1$  segment: locations belong to the same stream segment
2.  $N_2$  stream: locations belong to the same stream, at the same order
3.  $N_3$  siblings: locations belong to the same stream, but at different orders of the stream
4.  $N_4$  cousins: locations belong to different streams, but have the same order and source.

The above model can be generalized to this case where the weights are adjusted accordingly. The parameter  $\nu$  defines the structure of the within watershed spatial variation and includes  $\lambda_\nu$  and others necessary to further specify the weights. We expect that the precise form of these neighborhood schemes, and weights depends on the nature of the spatial variation identified during the data analysis process – this will be explored further in future work.

- $\epsilon(x; \sigma)$  residual spatial variation.

The residual spatial variation is assumed to be independent of the other factors in the model. The form of the variation depends on the models specified for the spatial dependence. If an auto-normal is used for the other terms then  $\epsilon(x; \sigma)$  will be assumed to be mean zero Gaussian with standard deviation  $\sigma$ .

### 3.1. Inferential Procedures

Based on this model, we use likelihood-based inference for  $Z(x)$  to infer the parameters  $\beta_1, \beta_2, \beta_3, \gamma, \nu$  and  $\sigma$ . Most of our likelihood-based inference is within the Bayesian paradigm, mainly as it provides an elegant way of incorporating parameter uncertainty into the final inference and the incorporation of expert knowledge when it exists (Gelman, Carlin, Stern and Rubin 1995). Inference with the Bayesian paradigm, implemented via the now standard Markov Chain Monte Carlo (MCMC) methods can address, and even solve, many very difficult inferential problems, often making it the only realistic option. However, this approach is not always necessary as Zimmerman and Cressie (1992) show.

The likelihood framework makes available exploratory graphical tools useful for inference about the underlying random field (Handcock, Meier and Nychka 1994). These tools can identify when an approach is lacking.

## 4. Models for Spatial Cumulative Distributions

In an increasing number of environmental applications, the comparison of an environmental indicator across regions requires consideration of more than the usual summary measures of level and variation. Environmental scientists are increasingly interested in techniques for comparing changes in distributional shape as well as changes in mean-levels. Traditionally, comparative research has relied heavily on measures that capture differences in average indices between regions or rough measures of dispersion over time. These summary measures leave

untapped much of the information inherent in a distribution.

A common distributional tool is the spatial cumulative distribution function (SCDF) defined over each watershed:

$$F_i(r) = \int_{x \in W_i} \mathcal{I}(Z(x) \leq r) dx \quad (4.1)$$

The SCDF is a largely under appreciated characteristic of spatial random fields. The motivation and use of SCDFs is reviewed in Lahiri, Kaiser, Cressie and Hsu (1999). By focusing as they do on the SCDF, scientific attention is moved from idealized point spatial units to larger, often more relevant, regional units. These distribution functions, and numerical summary measures derived from them are the basis of output from the model.

We now propose a model-based approach for the prediction of  $F_i(r)$ . The model (3.1) satisfies:

$$E\{Z(x)\} = f(x)' \beta \quad \text{for } x \in R,$$

where  $f(x) = \{L(x), C(x), S(x)\}'$  is a known vector function and  $\beta = (\beta_1, \beta_2, \beta_3)$  is a vector of unknown regression coefficients. Furthermore, we represent the covariance function by

$$\text{cov}\{Z(x), Z(y)\} = \alpha K_\theta(x, y) \quad \text{for } x, y \in R$$

where  $\alpha > 0$  is a scale parameter,  $\theta = (\nu, \gamma, \sigma) \in \Theta$  is a  $q \times 1$  vector of structural parameters, and  $\Theta$  is an open set in  $\mathbb{R}^p$ . For the models illustrated here, the exact form of  $K_\theta(x, y)$  can be derived directly from the neighborhood dependence structure given for the model (3.1). Under that structure,  $\{Z(x) : x \in R\}$  is Gaussian, although the covariance structure is not stationary. The development for alternative and more general models follows the same principles.

If we wish to predict characteristics of  $\{Z(x) : x \in R\}$ , then we need to express our uncertainty about the unknown dependence structure through  $\theta$  and the mean through  $\beta$ . Under a simple Bayesian formulation (see Handcock and Stein 1993), we can specify the prior as

$$\text{pr}(\alpha, \beta, \theta) \propto \text{pr}(\theta)/\alpha$$

so that the marginal posterior distribution:

$$\text{pr}(\theta | Z) \propto \text{pr}(\theta) \cdot |K_\theta|^{-1/2} |F' K_\theta^{-1} F|^{-1/2} \hat{\alpha}(\theta)^{-(N-q)/2}$$

captures our knowledge about  $\theta$ . Here

$$\hat{\alpha}(\theta) = (1/N)(Z - F\hat{\beta}(\theta))' K_\theta^{-1} (Z - F\hat{\beta}(\theta))$$

and



$$\hat{\beta}(\theta) = (F'K_\theta^{-1}F)^{-1}F'K_\theta^{-1}Z$$

are the maximum likelihood estimators (MLEs) of  $\alpha$  and  $\beta$  conditional on  $\theta$ ,  $F = \{f_j(x_i)\}_{N \times q}$ , and  $K_\theta = \{K_\theta(x_i, x_j)\}_{N \times N}$ . To predict the SCDF for a given watershed,  $F(z)$ , say, we need to express our understanding of  $Z(x)$  at each point  $x \in R$ . Operationally choose a large finite subset of locations  $v_1, \dots, v_m$ , from the watershed as a surrogate for the continuum. For example, we could choose a realization from a high-intensity spatial Poisson point process restricted to the riverine system in the watershed. or a design adapted for numerical integration (Owen 1994).

Let  $Z = \{Z(x_1), \dots, Z(x_N)\}'$  be the sample, and let  $Z_0 = \{Z(v_1), \dots, Z(v_m)\}'$  then

$$\begin{pmatrix} Z \\ - \\ Z_0 \end{pmatrix} \sim N_{N+m} \left[ \begin{pmatrix} F\beta \\ - \\ \tilde{F}\beta \end{pmatrix}, \alpha \begin{pmatrix} K_\theta & | & H_\theta \\ - & & - \\ H'_\theta & | & J_\theta \end{pmatrix} \right]$$

It is well known that:

$$Z_0 | \theta, Z \sim t_m \left( \hat{Z}_0(\theta), \kappa \hat{\alpha}(\theta) \{ J_\theta - H'_\theta K_\theta^{-1} H_\theta + B'_\theta (F'K_\theta^{-1}F)^{-1} B_\theta \} \right)$$

$$\text{pr}(Z_0 | Z) = \int_{\Theta} \text{pr}(Z_0 | \theta, Z) \text{pr}(\theta | Z) d\theta \quad (4.2)$$

where

$$\begin{aligned} B_\theta &= \tilde{F}' - F'K_\theta^{-1}H_\theta \\ \hat{Z}_0(\theta) &= H'_\theta K_\theta^{-1}Z + B'_\theta \hat{\beta}(\theta) \\ \kappa &= N/(N - q) \end{aligned}$$

These calculations are straightforward even for large  $m$  as the conditional predictive distribution is multivariate  $t$  with the appropriate covariance matrix and inversion of the covariance matrix of  $Z_0$  is not necessary. In some circumstances, it will be easier to use the formula

$$\text{pr}(Z_0 | Z) = \frac{\text{pr}(Z_0 | \theta, Z) \text{pr}(\theta | Z)}{\text{pr}(\theta | Z, Z_0)}$$

(Besag 1989), rather than do the  $p$ -dimensional integral directly. The posterior distribution of  $F(z)$  can then be approximated by that of

$$F^m(z) = \frac{1}{m} \sum_{i=1}^m \mathcal{I}(\tilde{Z}(v_i) \leq z) \quad (4.3)$$

where  $\{\tilde{Z}(v_1), \dots, \tilde{Z}(v_m)\}$  is a random draw from (4.2). The approximation can be made arbitrarily accurate by choosing  $m$  large, and more importantly

the accuracy of the approximation can be easily assessed for any  $m$ . One simple approach is to draw samples directly from  $\text{pr}(Z_0 | Z)$  and use (4.3) for a range of  $z$  values to obtain draws from posterior of  $F(z)$ . The analysis of these draws would be very useful in understanding the behavior of  $F(z)$ . In particular they can be used to define pointwise probability limits and prediction bounds for  $F(z)$ .

### Example: Spatial Distribution of Fish IBI in the Gunpowder-Patapsco watershed

As an illustration, the model described in Section 3 has been applied to a single watershed in Maryland on the shores of Chesapeake Bay. The location and hydrology of the watershed are given in the Figure 1. For simplicity, a simple contiguity neighborhood for the streams with distance decay specified by the Matérn class of covariances and prior distributions described by Handcock and Wallis (1994) is used. The resulting mean posterior SCDF for the fish IBI metric is given in Figure 2. The point prediction is smooth as the model averages over many possible spatial dependence structures.

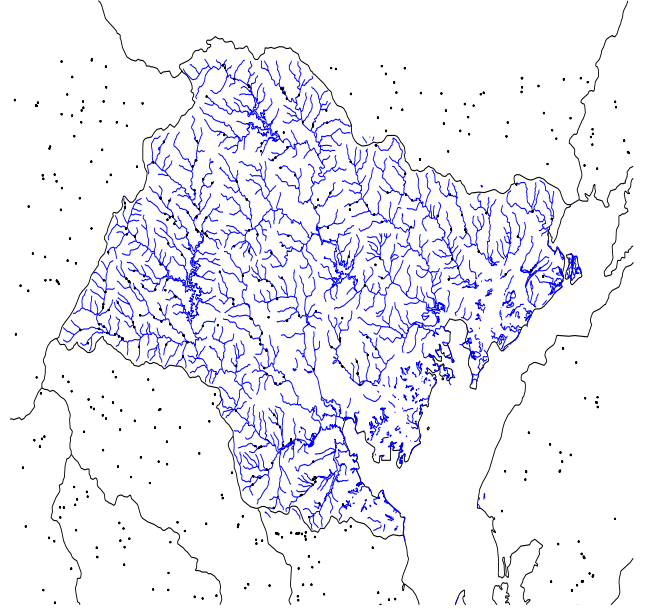


Figure 1: Hydrologic detail of the Gunpowder-Patapsco watershed on the Chesapeake Bay in Maryland. The figure provides the detail of the wadeable streams in the watershed and outlines of the surrounding watersheds. The monitoring sites from the Maryland Biological Streams Survey are marked.

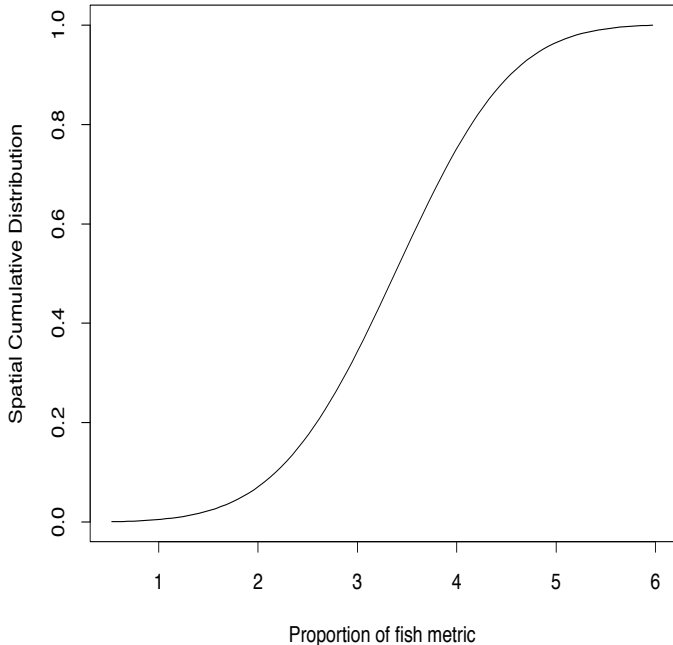


Figure 2: The mean posterior spatial cumulative distribution function for the fish IBI within the Gunpowder-Patapsco watershed.

## 5. Models for Relative Spatial distributions

Many questions of environmental justice take the form, usually implicitly, of the comparison of distributions across different groups. For example, consider comparing the pollution levels of an area with predominantly lower socioeconomic status to one with predominately higher status. This is fundamentally a distributional question - how does the SCDF of the pollution level of the lower socioeconomic area compare to that of the higher. Relative distribution methods are designed to address these questions. We review the concepts below. A book length treatment is given in Handcock and Morris (1999). See also the website (<http://www.stat.psu.edu/~handcock/RelDist>) This site contains software, example data-sets, manuals, and example code for published analyses to guide the practitioner.

The relative distribution summarizes the information required for scale-invariant comparisons between two distributions. It appears, explicitly and implicitly, in many independent research areas (Parzen 1977, 1992, Cwik and Mielniczuk 1993, Holmgren 1995, Li, Tiwari and Wells 1996). Re-

cently it has been used to study changes in environmental characteristics over time and between demographic groups. For example, Morris, Bernhardt and Handcock (1994) study changes in yearly earnings by race and gender from 1967 to 1987. Bernhardt, Morris and Handcock (1995) used it, and its extensions, to take a closer look at the shrinking gender gap in earnings. Handcock and Morris (1998) use the relative distribution to study the changes in the distribution of yearly hours worked between 1975 and 1993. In each of these areas of study the pattern of the changes has made it necessary to study differences beyond the usual differences in the summary measures of location and variation (Butler and McDonald, 1987; Karoly, 1993). Additional applications are given in Handcock and Morris (1999).

### 5.1. The Relative SCDF and the Relative Spatial Density

Let  $F_0$  be the SCDF of an environmental indicator on a reference area and  $F$  be the corresponding SCDF for a comparison area. Typically the reference area is the measurement for a separate area or the same area during an earlier time period. However the reference distribution can be from a minimally disturbed area where it represents a nominally “pristine” state. Indeed, it may even be a hypothetical distribution based on an environmental standard or regulation. The objective is to study the differences between the distributions of the environmental indicator in the reference and comparison areas. Let  $Y_0 \sim F_0$  and  $Y \sim F$ . We suppose that  $F_0$  and  $F$  are absolutely continuous with common support. The *grade transformation* of  $Y$  to  $Y_0$  is defined as the random variable (Cwik and Mielniczuk 1989):

$$R = F_0(Y) \quad (5.1)$$

$R$  is obtained from  $Y$  by transforming it by the function  $F_0$  and so it is continuous with outcome space  $[0, 1]$ . As  $R$  measures the relative rank of  $Y$  compared to  $Y_0$ , we refer to the distribution of  $R$  as the *relative spatial distribution*. We can express the CDF of  $R$  as

$$G(r) = F(F_0^{-1}(r)) \quad 0 \leq r \leq 1 \quad (5.2)$$

where  $r$  represents the proportion of values, and  $F_0^{-1}(r) = \inf_y \{y \mid F_0(y) \geq r\}$  is the quantile function of  $F_0$ . The probability density function (PDF) of  $R$  is

$$g(r) = \frac{f(F_0^{-1}(r))}{f_0(F_0^{-1}(r))} \quad 0 \leq r \leq 1 \quad (5.3)$$

If the two distributions are identical then the CDF of the relative distribution is a 45° line and the PDF of the relative distribution is the uniform PDF.

The relative distribution is an intuitively appealing approach to the comparison problem because both the density and the CDF have clear, simple interpretations. The relative spatial density  $g(r)$  can be interpreted as the ratio of the comparison population to the reference population at a given level ( $F_0^{-1}(r)$ ). The relative spatial CDF  $G(r)$  can be interpreted as the proportion of the comparison area whose attribute lies below the  $p$ th quantile of the reference area. More technically: a proportion  $G(r)$  of the  $Y$  are below the values of a proportion  $p$  of  $Y_0$ .

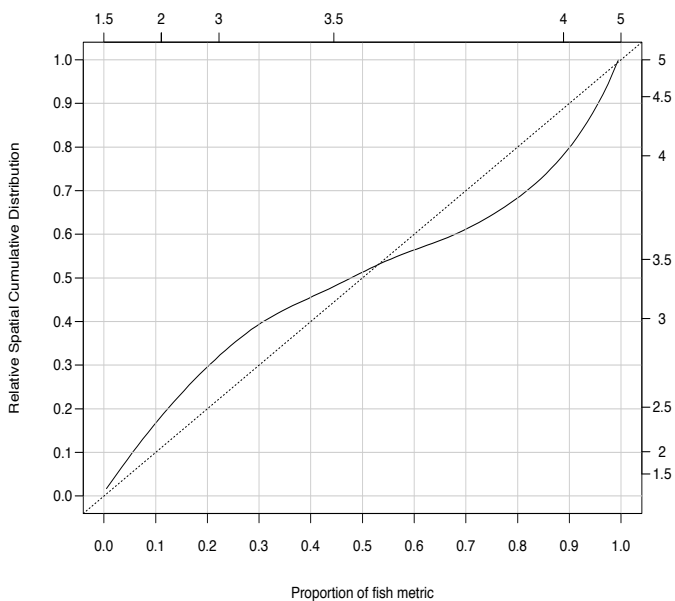


Figure 3: The relative spatial CDF of fish IBI. in the Gunpowder-Patapsco watershed to the Severn watershed. The upper and right axes is labeled in the fish IBI units.

Inference for the relative distribution when the available information takes the form of independent sample from both reference and comparison distributions is reviewed in Handcock and Morris (1999). We are developing Bayesian inference for both the relative SCDF and the relative spatial density. As we have, in principle, the joint posterior for  $F_0$  and  $F$ , we can use (5.2) and (5.3) to produce the posteriors for both quantities. However a number of implementation issues remain to be addressed.

### Example: Comparing the Gunpowder-Patapsco and Severn watersheds

Adjacent, and upstream, from the Gunpowder-Patapsco watershed is the Severn watershed. We have repeated the modeling process for Severn and compared the two watersheds in Figures 3 and 4.

Figure 4. is the relative spatial density of the Gunpowder-Patapsco watershed to the Severn watershed. The value of one represents the relative density if the two distributions were identical. We can see, however, that there is a substantial difference between the shapes of the two distributions. The fish IBIs for Gunpowder-Patapsco over-represented in the lower and upper quantiles of the Severn distribution. They are correspondingly under-represented in the middle 60% of the distribution. The frequency of Gunpowder-Patapsco streams does not match that of Severn streams until about the 25% quantile and again at the 85% quantile of the Severn distribution. These observations are not readily apparent from the direct comparison of the SCDF for Severn with that for Gunpowder-Patapsco in Figure 2.

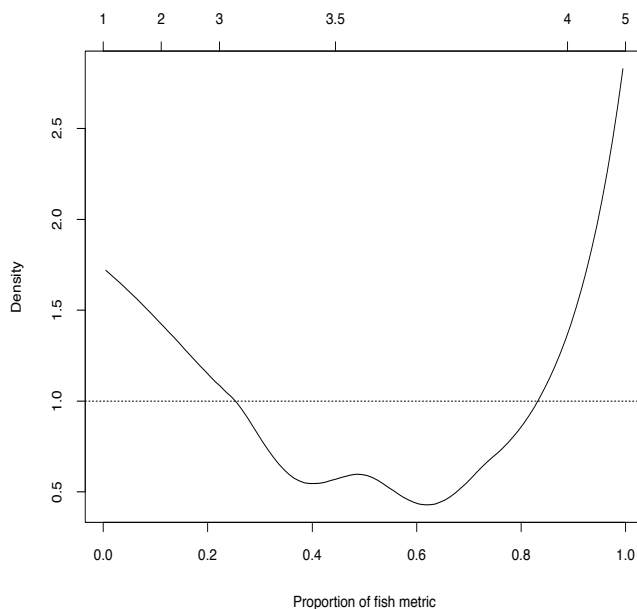


Figure 4: The relative spatial density of fish IBI in the Gunpowder-Patapsco watershed to the Severn watershed. The upper axis is labeled in the fish IBI units.

The relative density enhances comparison of the distributions in two ways. Firstly, it expresses the relative frequency in terms of a ratio, which is easier to understand both visually and numerically. Secondly, it rescales the horizontal axis so that length is equivalent to the proportion of streams in Severn with that level of fish IBI. This facilitates direct comparisons between the SCDFs because the two axes

are now in comparable units. For example, there are over 1.5 times as many Gunpowder-Patapsco than Severn stream-miles in the lower decile of the Severn fish IBI distribution.

These figures demonstrate how the relative spatial distribution can aid the comparison of distributions. This is not to suggest that they can replace the SCDF (as in Figure 2); rather it complements it by focusing on those characteristics of the individual distributions essential for scale-free comparison. Figures 2 3 provide absolute and relative description of environmental condition respectively. Figure 4 provides a relative description on a scale that may be more interpretable for most statisticians.

## 6. Application to Environmental Social Science: Modeling the Relationship Between Demographic Characteristics and Environmental Health

Our model for the environmental indicators, and the application of this model to relative spatial distributions, can also be used to address questions in the area of environmental social science. The applications described in this proposal predict spatial distributions (SCDFs) for geographical units defined by watersheds. However the same procedure can be used with appropriate modifications to predict the SCDFs for politically defined units such as counties. There is a wealth of information on human population characteristics readily available at the county level. The modeling framework proposed here will make it possible to investigate the relationship between these sociodemographic factors and environmental health in direct and innovative ways.

As a simple but practical example of this, we can relate the population density of a county to characteristics of the SCDF for indicators of environmental health. For example, this allows us to assess the impact of density on such things as the mean or ninetieth percentile of a fish biodiversity index. Any population characteristic for which data are collected at the county level – from racial composition to business indicators – can be integrated into this framework.

In future work we propose specifically to address questions of environmental justice using this approach. One of the most basic issues of environmental justice concerns the link between community economic status and environmental degradation. It is often argued that poorer communities are less effective in preventing the siting of environmentally undesirable activities in their areas. While this ar-

gument has fairly strong face validity, the empirical evidence is typically based on case studies. To systematically evaluate this claim, it is necessary to take a different approach, and the modeling framework proposed here makes that possible.

At the county level, a natural index of community economic status is the level of household income. Mean income is the obvious candidate, but the mean can mask large disparities in community status within counties. To the extent that the bottom tail of the distribution drives the vulnerability of the community, some measure of the inequality in household income would be a better index. We know that household income inequality in the United States has increased dramatically during the past three decades (Karoly, 1993), and a county-level study of the spatial and temporal patterns in inequality is now underway (McLaughlin and Handcock, 1999). The approach taken in this county-level study is perfectly suited to the task here. It is based on census data and produces indicators of the level of economic inequality for each county that are easily integrated into spatial-temporal models. We propose to use the modeling approach described above to examine the relationship between summary measures of environmental indicators based on the SCDF for each county, and summary measures of county economic status. The association between economic inequality and environmental health can thus be systematically and directly assessed. For example, we may find that the mean county income has little impact on environmental health, while inequality, or the level of the lowest income decile, has a large impact. For smaller areas, where a more detailed analysis is possible, an alternative approach can be taken that uses all of the available distributional information, rather than the summary measures. This would allow us to compare the relative distribution of income between counties to the relative spatial distribution of environmental health, using the methods in Handcock and Morris (1999).

## 7. Further Issues and Extensions

It is tempting to develop a spatial-temporal model for indicators in the region. However the number of monitoring site revisits is small so that the amount of information on the temporal patterns is small. We note that the modeling framework can be extended in a straightforward fashion to include simple temporal effects.

A model-based approach such as the one described here coupled with a broad range of dependence structures can capture a wide range of spatial varia-

tion. However, in most cases the underlying random field can not be assumed to be Gaussian. The conditional distribution of  $Z(x)$  may be Gamma (Best, Ickstadt and Wolpert 1998) or even discrete. We may also wish to consider derived measures of exceedences useful for risk assessment e.g.,

$$E(x) = \mathcal{I}(Z(x) \leq L) \quad \text{for given } L \quad (7.1)$$

where  $L$  is a pre-specified limit on the measure. The generalized linear spatial models approach of Diggle, Tawn, and Moyeed (1998) greatly broaden the form of spatial variation that can be represented by the framework described here. There is a fully Bayesian approach that can also be implemented via MCMC methods. The central idea is that the observed indicator  $E(x)$ , say, satisfies a generalized linear model conditional on  $Z(x)$ . In essence we are adding another layer to the top of the hierarchical model. Extensions such as these improve on the simple model described here at the expense of some computational complexity.

The watershed effects can also be modeled hierarchically to investigate the effects of political divisions. The most direct example is the state (or states) that the watershed resides in. While the watershed does not necessarily respect state boundaries, state and local government regulations may directly influence the environmental condition and human activities. Hence the relative comparison of state-level effects is a very important way of assessing the role of institutions at the state level. This approach can be applied to other political division such as labor-market regions and counties. It is also natural to consider extensions to more sophisticated models for misaligned data, such as those of Muggins and Carlin (1998).

### Acknowledgments

We would like to thank James McDermott (Penn State University) and Barbara Rosenbaum (EPA-Corvallis) for help with computing and data compilation issues.

This project has been developed in collaboration with EPA scientists and statisticians including those from the MAIA group. The investigator has also been working with the Maryland Department of Natural Resources on the Maryland Biological Streams Survey, and environmental scientists on the Mid-Atlantic Regional Assessment (MARA) of Climate Change Impacts conducted at the Pennsylvania State University. The extensive and ongoing work by these groups on the mid-Atlantic region is a great help to this project. The projects, especially those initiated by the MAIA group, are developing

data sets and comprehensive reviews of combined data sources that greatly enhance the scientific quality of the research in this area.

### References

- [1] Anselin, L. and Florax, R. (1995). *New Directions in Spatial Econometrics* Berlin: Springer-Verlag
- [2] Bernhardt, A. D, Morris, M. and Handcock, M. S. (1995) Women's gains or men's losses? A closer look at the shrinking gender gap in earnings. *American Journal of Sociology*, **101**, 302-328.
- [3] Bernhardt, A., Morris, M., Handcock, M. and Scott, M. (1997). *Work and Opportunity in the Post-Industrial Labor Market*. Final report to the Russell Sage and Rockefeller Foundations. Institute on Education and the Economy, Columbia University New York.
- [4] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, B*, **36**, 192-236.
- [5] Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, **24**, 179-95.
- [6] Besag, J. (1989). A Candidate's Formula: A Curious Result in Bayesian Prediction, *Biometrika*, **76**, 183-183.
- [7] Besag, J. (1989). Towards Bayesian Image Analysis. *Journal of Applied Statistics*, **16**, 395-407.
- [8] Besag, J. and Kooperberg, C. (1995). On Conditional and Intrinsic Autoregression. *Biometrika*, **82**, 733-746.
- [9] Besag, J, York, J. and Mollié, A. (1991). Bayesian Image Restoration, with two applications is spatial statistics (with discussion). *Annals of the Institute of Statistics and Mathematics*, **43**, 1-59.
- [10] Besag, J, and Higdon, D. (1999). Bayesian Analysis of Agricultural Field Trials (with discussion). *Journal of the Royal Statistical Society, B*, **61**, 691-746.
- [11] Bernardinelli, L. and Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, **11**, 983-1007.

- [12] Best, N. G., Ickstadt, K. and Wolpert, R. L. (1998) Spatial Poisson regression for health and exposure data measured at disparate resolutions. Discussion paper, 98-36, Institute of Decision Sciences, Duke University.
- [13] Cox, L. H. and Piegorsch, W. W. (1996). Combining Environmental Information. I: Environmental Monitoring, Measurement and Assessment. *Environmetrics*, **7**, 299-308.
- [14] Cox, L. H. (1998). Workshop: Statistical Methods for Combining Environmental Information. In D. Nychka, W. W. Piegorsch and L. H. Cox (eds.), *Case Studies in Environmental Statistics*, 143-158. Lecture Notes in Statistics, 132. Springer-Verlag.
- [15] (1993). Cressie, N. A. C. *Statistics for Spatial Data*. New York: Wiley.
- [16] Cressie, N. A. C. (1995). Bayesian smoothing of rates in small geographic areas. *Journal of Regional Science*, **35**, 659-673.
- [17] Cwik, J. and Mielniczuk, J. (1989) Estimating density ratios with application to discriminant analysis. *Communications in Statistics*, **18**, 3057-3069.
- [18] Cwik, J. and Mielniczuk, J. (1993) Data-dependent bandwidth choice for a grade density kernel estimate. *Statistics and Probability Letters*, **16**, 397-405.
- [19] Dewald, T. and Olsen, M. (1994). The EPA Reach File: A National Spatial Data Resource. U.S. Environmental Protection Agency, Office of Water.
- [20] Ecker, M. D. and Gelfand, A. E. (1997). Bayesian Variogram Modeling for an Isotropic Process. *Journal of Agricultural, Biological, and Environmental Statistics*, **4**, 347-369.
- [21] Fausch, K. D., Lyons, J., Karr, J. R., and Angermeier, P. L. (1990). Fish communities as indicators of environmental degradation. In S. M. Adams (ed.), *Biological Indicators of Stress in Fish*, 123-144. American Fisheries Society Symposium 8. Bethesda, Maryland.
- [22] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian Data Analysis*, London: Chapman and Hall.
- [23] Handcock, M. S. (1994). Discussion of "Epidemics: Models and Data" by D. Mollison, V. Isham and B. Grenfell (1994). *Journal of the Royal Statistical Society*, **A**, *157*, 115-149.
- [24] Handcock, M. S. (1998). Discussion of "Model-based Geostatistics" by P. J. Diggle, J. A. Tawn and R. A. Moyeed (1998). *Journal of the Royal Statistical Society*, **C**, *47*, 3, p. 299-350.
- [25] Handcock, M. S. (1999). Discussion of "Prediction of Spatial Cumulative Distribution Functions Using Subsampling" by S. Lahiri, M. Kaiser, N. Cressie and N. Hsu (1999). *Journal of the American Statistical Association*, **94**, 445, p. 100-102.
- [26] Handcock, M. S. and Morris, M. (1998) Relative distribution methods. *Sociological Methodology*, **28**, 53-97.
- [27] Handcock, M. S. and Morris, M. (1999). *Relative Distribution Methods in the Social Sciences*. New York: Springer.
- [28] Handcock, M. S. and Stein, M. L. (1993). A Bayesian Analysis of Kriging, *Technometrics*, **35**, 4, 403-410.
- [29] Handcock, M. S. and Wallis, J. (1994). An Approach to Statistical Spatial-Temporal Modeling of Meteorological Fields (with discussion), *Journal of the American Statistical Association*, **89**, 368-378. rejoinder, 388-390.
- [30] Handcock, M. S., Meier, K. and Nychka, D. (1994). Comment on "Kriging and Splines: An Empirical Comparison of their Predictive Performance" by G. M. Laslett. *Journal of the American Statistical Association*, **89**, 401-403.
- [31] Handcock, M. S., Rendall, M. S. and Huovilainen, S. M. (2000). Combining Survey and Population Data on Births and Family. To appear in *Demography*, **37**.
- [32] Heimbuch, D., Seibel, J., Wilson, H, and Kazyak, P. (1998). A Multi-Year Lattice Sampling Design for Maryland-Wide Fish Abundance Estimation. Presented at the "Conference on Environmental Monitoring Surveys over Time," April 20-22, 1998, University of Washington.
- [33] Herlihy, A. T., Larsen, D. P., Paulsen, S. G. , Urquhart, N. S., and Rosenbaum, B. J. (1999). Designing a Spatially Balanced, Randomized Site Selection Process for Regional Stream Surveys: The EMAP Mid-Atlantic Pilot Study. Under review by *Environmental Monitoring and Assessment*.

- [34] Holmgren, E. B. (1995) The P-P plot as a method for comparing treatment effects. *Journal of the American Statistical Association*, **90**, 360-365.
- [35] Jones, K.B., K.H. Ritters, J.D. Wickham, R.D. Tankersley, R.V. O'Neill, D.J. Chaloud, E.R. Smith, and A.C. Neale. (1997). An Ecological Assessment of the United States Mid-Atlantic Region: A Landscape Atlas. Environmental Protection Agency: EPA/600/R-97/130.
- [36] Karoly, L. A. (1993) The trend in inequality among families, individuals, and workers in the United States: A twenty-five year perspective, in *Uneven Tides: Rising Inequality In America*. S. Danziger and P. Gottschalk (ed.) Russell Sage., New York, NY, pp. 19-97.
- [37] Karr, J. R., and Dudley, D. R. (1981). Ecological Perspectives on Water Quality Goals. *Environmental Management*, **5**, 55-68.
- [38] Kelsall, J. E. and Wakefield, J. C. (1997). Spatial Modeling of Disease Risk. Manuscript
- [39] Kepner, W. G., Jones, K. B., Chaloud, D. J., and Wickham, J. D. (1995). Mid-Atlantic landscape Indicators Project Plan. EPA/620/R-95/003. Washington, D.C.: U.S. Environmental Protection Agency.
- [40] Lahiri, S., Kaiser, M., Cressie, N. A. C., and Hsu, N. (1999). Prediction of Spatial Cumulative Distribution Functions Using Subsampling. *Journal of the American Statistical Association*, **94**, 445, p. 100-102.
- [41] Lazorchak, J. M., Klemm, D. L. and Peck, D. V. (eds). (1998). Environmental Monitoring and Assessment Program Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Wadeable Streams. EPA/620/R-94/004F. Washington, D.C.: U.S. Environmental Protection Agency.
- [42] Larsen, D. P., and S. J. Christie, Eds (1993). EMAP-Surface Waters 1991 Pilot Report. EPA/620/R-93/003. Corvallis, Oregon: U.S. Environmental Protection Agency.
- [43] Li, G., Tiwari, R. C. and Wells, M. T. (1996) Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *Journal of the American Statistical Association*, **91**, 689-698.
- [44] McLaughlin, D. K and Handcock, M. S. (1999). Spatial Statistical Models for the Distribution of Income Inequality in the United States, 1980 to 1990. Presented at the Population Association of America Annual Meetings, March 26, 1999.
- [45] Messer, J.J., Linthurst, R.A. and Overton, W.S. (1991). An EPA Program for Monitoring Ecological Status and Trends. *Environmental Monitoring and Assessment*, **17**, 67-78.
- [46] Mollié, A. and Richardson, S. (1992). Empirical Bayes estimates of cancer mortality rates using spatial models. *Statistics in Medicine*, **10**, 95-112.
- [47] Moller, J. (1998). Log Gaussian Cox Processes. *Scandinavian Journal Statistics*, **25**, 451-482.
- [48] Morris, M., Bernhardt, A. D. and Handcock, M. S. (1994) Economic inequality: New methods for new trends. *American Sociological Review*, **59**, 205-219.
- [49] Mugglin, A. S. and Carlin, B. P. (1998). Hierarchical modeling in geographic information systems: population interpolation over incompatible zones. *Journal of Agricultural, Biological, and Environmental Statistics*, **3**, 111-130.
- [50] Mugglin, A. S., Carlin, B. P. and Gelfand, A. E. (2000). Fully Model Based Approaches for Spatially Misaligned Data. To appear *Journal of the American Statistical Association*.
- [51] Mugglin, A. S., Carlin, B. P., Zhu, L. and Conlon, E. (1999). Bayesian areal interpolation, estimation, and smoothing: an inferential approach for geographic information systems. *Environment and Planning Statistics*, **31**, 1337-1352.
- [52] Owen, A. (1994). Lattice Sampling Revisited: Monte Carlo Variance of Means Over Randomized Orthogonal Arrays, *The Annals of Statistics* **22**, 930-945.
- [53] Parzen, E. (1977) *Nonparametric statistical data science: A unified approach based on density estimation and testing for 'white noise'*. Technical Report 47, Statistical Sciences Division, State University of New York at Buffalo, Buffalo, NY.
- [54] Parzen, E. (1979) Nonparametric statistical data modeling. *Journal of the American Statistical Association*, **74**, 105-131.

- [55] Parzen, E. (1992) Comparison change analysis, in *Nonparametric Statistics And Related Topics*. A. Saleh (ed.) Elsevier, Holland, pp. 3-15.
- [56] Piegorsch, W. W. and Cox, L. H. (1996). Combining Environmental Information. II: Environmental Epidemiology and Toxicology. *Environmetrics*, **7**, 309-324.
- [57] Propp, J. G. and Wilson, D. B. (1996). Exact Sampling with Coupled Markov Chains and Applications to Statistical mechanics. *Random Structures and Algorithms*, **9**, 223-252.
- [58] Raftery, A. E. and Banfield, J. D. (1991). Stopping the Gibbs Sampler, The Use of Morphology, and other Issues in Spatial Statistics. *Annals of the Institute of Statistics and Mathematics*, **43**, 1-59.
- [59] Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.
- [60] Vannote, R. L., Minshall, G. W., Cummins, K. W., Sedell, J. R., and Cushing, C. E. (1980). The River Continuum Concept. *Canadian Journal of Fisheries and Aquatic Sciences*, **37**, 130-137.
- [61] U.S. Dept. of Agriculture, Natural Resource Conservation Service (1994). State Soil Geographic (STATSGO) Data Base: Data Users Guide. The National Cartography and Geospatial Center, Fort Worth, TX.
- [62] U.S. Environmental Protection Agency. Office of Environmental Justice. (1993). Serving A Diverse Society: EPA's Role in Environmental Justice. EPA/200/F-93/001. Washington, D.C.: U.S. Environmental Protection Agency.
- [63] Wilk, M. B. and Gnanadesikan, R. (1968) Probability plotting methods for the analysis of data. *Biometrika*, **55**, 1-17.
- [64] Wilson, D. B. (2000). Layered Multishift Coupling for use in Perfect Sampling Algorithms (with a primer on CFTP). To appear *Fields Institute Communications*.
- [65] Zimmerman, D. L. and Cressie, N. A. C. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistics and Mathematics*, **44**, 27-43.
- [66] Zimmerman, D. L. and Harville, D. A. (1991). A Random Field Approach to the Analysis of Field-plot Experiments and Other Spatial Experiments. *Biometrics*, **47**, 223-239.