Title
Model-based combination of spatial information for stream networks

Journal
Environmental and Ecological Statistics, 14(3)

ISSN
1352-8505 1573-3009

Author
Handcock, Mark S

Publication Date
2007-07-10

DOI
10.1007/s10651-007-0015-2

Peer reviewed

Environ Ecol Stat (2007) 14:267–284
DOI 10.1007/s10651-007-0015-2

Model-based combination of spatial information for stream networks

Mark S. Handcock

Received: 1 March 2005 / Revised: 1 September 2005 / Published online: 10 July 2007

Abstract Evolutionary improvements in Geographic Information Systems (GIS) now routinely allow the management and mapping of spatial-temporal information. In response, the development of statistical models to combine information of different types and spatial support is of vital importance to environmental science. In this paper we develop a hierarchical spatial statistical model for environmental indicators 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. These models are used to estimate the indicators throughout the riverine system based on information from multiple sources and aggregate scales. The analysis is based on information underlying the Landscape Atlas of the mid-Atlantic region produced by the US Environmental Monitoring and Assessment Program (EMAP). We also combine information from two overlapping separate monitoring surveys, the EMAP Stream and River Survey and the Maryland Biological Streams Survey. We present a general framework for comparative distributional analysis based on the concept of a relative spatial distribution. As an application, the spatial model is used to predict spatial distributions and relative spatial distributions for a watershed.

Keywords Spatial statistics · GIS · Hierarchical models · Relative distribution · EMAP

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

M. S. Handcock (✉)
Department of Statistics, University of Washington,
Seattle, WA 98195-4322, USA
e-mail: handcock@stat.washington.edu

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, display 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 a extraordinary achievement in combining multiple data sources, and will clearly be the basis for the development of environmental statistics in the future.

In this paper we take advantage of the opportunities that EMAP Landscape Atlas of the mid-Atlantic region provides. We 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 of traditional 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 et al. (2000) and Handcock et al. (2005) 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.

In many papers on stochastic modeling the data structure is clear and simple. In such papers, the data sections can be abstracted. While model specification is usually driven by the structure of the available data, it is especially true here where a primary motivation for the model is to combine multiple sources of information. As such the description of the data sources and types in Sect. 2 is more detailed than usual. This is followed in Sect. 3 by a development of the model. In Sects. 4 and 5 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 approach in 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. 2000). 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 regions. 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 USGS defined hydrologic accounting units. Roughly speaking, these units follow watershed boundaries—areas of land that are 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 Sect. 6 we consider counties.

The basic features of the study region are given in Figure 3.4 of Jones et al. (1997). It provides an overview of the land cover and land use type. Figure 3.3 of Jones et al. (1997) represents the hydrography of the region, that is, the major rivers, streams and watersheds. The watersheds are represented by the (8-digit) hydrological units within the region.

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 et al. (2000), Mugglin and Carlin (1998), Mugglin et al. (1999), and Gotway and Young (2002). 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

In this section we briefly review the data sources used. We focus 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 R).

This paper uses as its foundation the work of the EPA/ORD Mid-Atlantic Integrated Assessment (MAIA). For an description of MAIA see http://www.epa.gov/emap/maia. A major part of this work is the Landscape Atlas of the mid-Atlantic region (Jones et al. 1997). 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 m), 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. For an description of the Atlas, see http://www.epa.gov/maia/html/maia-atlas.html.

Here we use two specific component surveys. The first is the *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/datal/surfwatr/data.

The second source is the *Maryland Biological Streams Survey (MBSS)* (Heimbuch et al. 1998). 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 been completed (http://www.epa.gov/maia/html/mbss.html). For a description of the MBSS see http://www.dnr.state.md.us/streams.

These surveys are supplemented by the *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 uses all first- through third-order (i.e. wadeable) streams. There are 230,400 kms of wadeable streams in the study region. 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.

2.3 Indicators of environmental condition

As detailed in above references, the two monitoring surveys (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 affect key characteristics of aquatic ecosystems: water quality, habitat structure, hydrologic regime, and biologic interactions (Karr and Dudley 1981). In the next section we develop a model for a single indicator, with natural extensions left for Sect. 7.

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 Sects. 4–6. 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), Molliè 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 $\{W_i : i = 1, \ldots, H\}$ 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 Sect. 6. Throughout we will use as an example the fish index of biotic integrity (IBI) (Karr et al. 1986).

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

$$Z(x) = L(x)\beta_1 + C(x)\beta_2 + S(x)\beta_3 + \eta(W(x); \gamma) + \phi(x; \nu) + \epsilon(x; \sigma) \qquad (1)$$

where the first three terms capture variation due to differences in covariates, the ϕ and η 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. That is, they are a characteristic of an area associated with the location x. 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 Sect. 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 a requirement of the model. While the division is artificial from a modeling

perspective it serves the theme of combining data sources via a model by clarifying the precise linkage of contextual, complete coverage and location-specific data types with the random field $Z(x)$. Within the model the components are treated similarly. 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. Whether we believe in the existence of true spatial proximity effects depends on the philosophical interpretation of the model. If we believe the model is a causal representation then the latent variables only approach is compelling. If we believe the model is descriptive then there is room for the residual spatial proximity effects.

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 $\{\eta(i; \gamma) : i = 1, \ldots, H\}$. 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 $\{\eta(i; \gamma) : i = 1, \ldots, H\}$ 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 1991). Let v_{ij} be prescribed non-negative weights, with $v_{ij} = 0$ unless watersheds i and j are neighbors and let λ_γ be a scale parameter. The conditional distribution of $\eta(i; \gamma)$ given the other effects in the watershed is specified to be Gaussian:

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

where NW_i represents the watersheds j that are neighbors of i and v_{i+} is the sum over $j \in \mathrm{NW}_i$ 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 γ includes λ_γ and others necessary to further specify the weights.

$\phi(x; \nu)$ *latent within watershed effects within the watershed of location* x. We model each $\{\phi(x; \nu) : x \in W_i\}, i = 1, \ldots, H$ 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 1989b). For example, consider a neighborhood system for x based on being on the same stream segment (according to RF3). That is, two locations are neighbors if, and only if, they belong to the same stream segment. One would expect that, all else being equal, two locations on the same stream would more likely have closer values on a 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 approach 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 2002; Moller 1998; Ecker and Gelfand 1997, Best et al. 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. 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, λ_ν is a scale parameter, 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; 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 ν defines the structure of the within watershed spatial variation and includes λ_ν 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 σ.

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 σ. 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 et al. 2003). 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.

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

In addition to Bayesian inference for the parameters, we will usually be interested in posterior distributions for the latent effects $\phi(x;\cdot)$ and $\eta(W(x);\cdot)$. These can be plotted spatially and use to create maps of summarizing knowledge about $Z(x)$ over the stream network. Here, however, we emphasize inference for areal measures, the topic of the next section.

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:

$$F_i(r) = \frac{1}{|\mathrm{W}_i|}\int_{x\in \mathrm{W}_i} \mathcal{I}(Z(x)\le r)dx \tag{2}$$

where $|\mathrm{W}_i|$ is the length of the stream elements in watershed i. The SCDF is a largely under appreciated characteristic of spatial random fields. The motivation and use of SCDFs is reviewed in Lahiri et al. (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 (1) satisfies:

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

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

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

where $\alpha > 0$ is a scale parameter, $\boldsymbol{\theta} = (\nu,\gamma,\sigma)\in\boldsymbol{\Theta}$ is a $p\times 1$ vector of structural parameters, and $\boldsymbol{\Theta}$ is an open set in $\mathbb{R}^p$. For the models illustrated here, the exact form of $K_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y})$ can be derived directly from the neighborhood dependence structure given for the model (1). This is one of the advantages of modeling the covariance structure via latent variables in a hierarchical manner. Direct modeling of such global covariance structure is subject to peril as it relies on global specifications. Our approach builds from local properties to global outcomes. This representation averages over the components of spatial variation in model (1) to produce the global covariance. This covariance function describes the overall covariance

between points in the stream network but obscures the structure of the covariation. Under this structure, $\{Z(\boldsymbol{x}) : \boldsymbol{x} \in \boldsymbol{R}\}$ 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 $\{Z(\boldsymbol{x}) : \boldsymbol{x} \in \boldsymbol{R}\}$, then we need to express our uncertainty about the unknown dependence structure through θ and the mean through β. Under a simple Bayesian formulation (see Handcock and Stein 1993), we can specify the prior as

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

so that the marginal posterior distribution:

$$\mathrm{pr}(\theta \mid Z) \propto \mathrm{pr}(\theta) \cdot |K_{\boldsymbol{\theta}}|^{-1/2} |F' K_{\boldsymbol{\theta}}^{-1} F|^{-1/2} \widehat{\alpha}(\theta)^{-(N-q)/2}$$

captures our knowledge about θ. Here

$$\widehat{\alpha}(\theta) = (1/N)(Z - F\widehat{\beta}(\theta))' K_{\boldsymbol{\theta}}^{-1} (Z - F\widehat{\beta}(\theta))$$
$$\text{and}$$
$$\widehat{\beta}(\theta) = (F' K_{\boldsymbol{\theta}}^{-1} F)^{-1} F' K_{\boldsymbol{\theta}}^{-1} Z$$

are the maximum likelihood estimators (MLEs) of α and β conditional on θ, $F = \{f_j(x_i)\}_{N \times q}$, and $K_{\boldsymbol{\theta}} = \{K_{\boldsymbol{\theta}}(x_i, x_j)\}_{N \times N}$. The prior for θ, $\mathrm{pr}(\theta)$, can be very flexible and capture expert knowledge if available. In the application of this paper we presume little explicit knowledge and a simple structure. Explicitly, we presume λ_γ and λ_ν have Gamma distributions, and σ has an inverse χ distribution. We also presume prior independence among them.

To predict the SCDF for a given watershed, $F(z)$, say, we need to express our understanding of $Z(\boldsymbol{x})$ at each point $\boldsymbol{x} \in \boldsymbol{R}$. Operationally choose a large finite subset of locations $v_1, \ldots, v_m$, 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 $Z = \{Z(x_1), \ldots, Z(x_N)\}'$ be the sample, and let $Z_0 = \{Z(v_1), \ldots, Z(v_m)\}'$ then

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

It is well known that:

$$Z_0 \mid \theta, Z \sim t_m \left(\widehat{Z}_0(\theta), \kappa \widehat{\alpha}(\theta) \{J_\theta - H'_\theta K_{\boldsymbol{\theta}}{}^{-1} H_\theta + B'_\theta (F' K_{\boldsymbol{\theta}}^{-1} F)^{-1} B_\theta\} \right)$$
$$\mathrm{pr}(Z_0 \mid Z) = \int_\Theta \mathrm{pr}(Z_0 \mid \theta, Z) \mathrm{pr}(\theta \mid Z) d\theta \tag{3}$$

where

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

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

$$\mathrm{pr}(Z_0 \mid Z) = \frac{\mathrm{pr}(Z_0 \mid \theta, Z) \mathrm{pr}(\theta \mid Z)}{\mathrm{pr}(\theta \mid Z, Z_0)}$$

(Besag 1989a). This avoids computing the p-dimensional integral (3) directly, but does require the existing software code for computing $\text{pr}(\theta \mid Z)$ to be extended to compute $\text{pr}(\theta \mid Z, Z_0)$. This is typically straightforward. The posterior distribution of $F(z)$ can then be approximated by that of

$$F^m(z) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{I}(\tilde{Z}(v_i) \leq z) \tag{4}$$

where $\{\tilde{Z}(v_1), \ldots, \tilde{Z}(v_N)\}$ is a random draw from (3). 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 $\text{pr}(Z_0 \mid Z)$ and use (4) 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 Sect. 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 Fig. 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 Fig. 2. The point prediction is smooth as the model averages over many possible spatial dependence structures.

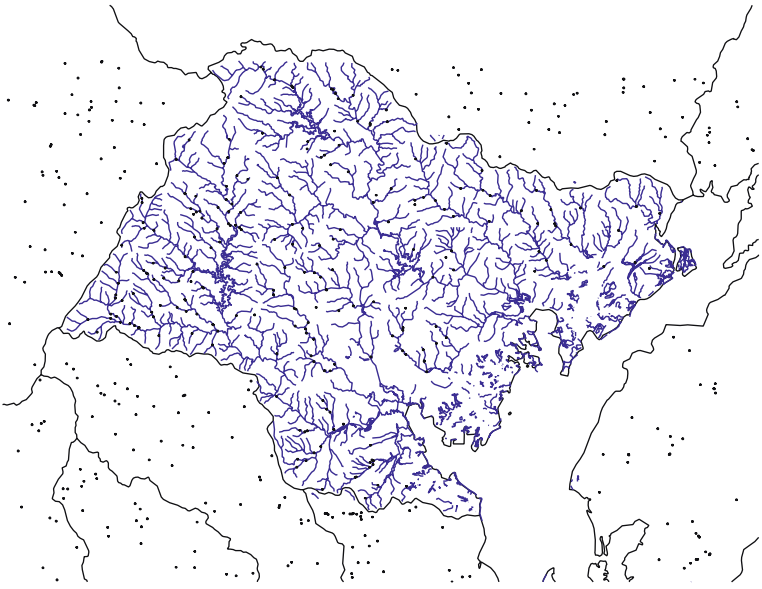

Fig. 1 Hydrologic detail of the Gunpowder-Patapsco watershed on the Chesapeake Bay in Maryland. The figure provides the detail of the wadeable streams in the watershed and outlines of the surrounding watersheds. The monitoring sites from the Maryland Biological Streams Survey are marked

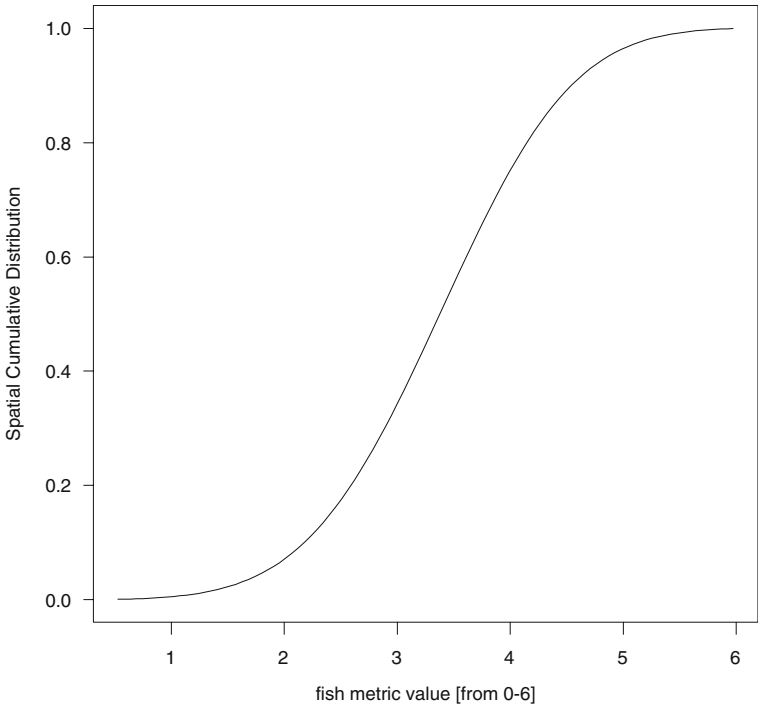

Fig. 2 The mean posterior spatial cumulative distribution function for the fish IBI within the Gunpowder-Patapsco watershed

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://csde.washington.edu/∼handcock/RelDist). This site contains software, example data-sets, manuals, and example code.

The relative distribution is 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 et al. 1996). Recently it has been used to study changes in environmental characteristics over time and between demographic groups. For example, Morris et al. (1994) study changes in yearly earnings by race and gender from 1967 to 1987. Bernhardt et al. (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):

$$R = F_0(Y) \tag{5}$$

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

$$G(r) \;=\; F(F_0^{-1}(r)) \qquad\qquad 0 \le r \le 1 \tag{6}$$

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

$$g(r) = \frac{f\left(F_0^{-1}(r)\right)}{f_0\left(F_0^{-1}(r)\right)} \qquad\qquad 0 \le r \le 1 \tag{7}$$

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 pth 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.

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). As we have the joint posterior for F_0 and F, we can use (6) and (7) to produce the Bayesian inference for both the relative SCDF and the relative spatial density.

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 Figs. 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 are over-represented in the

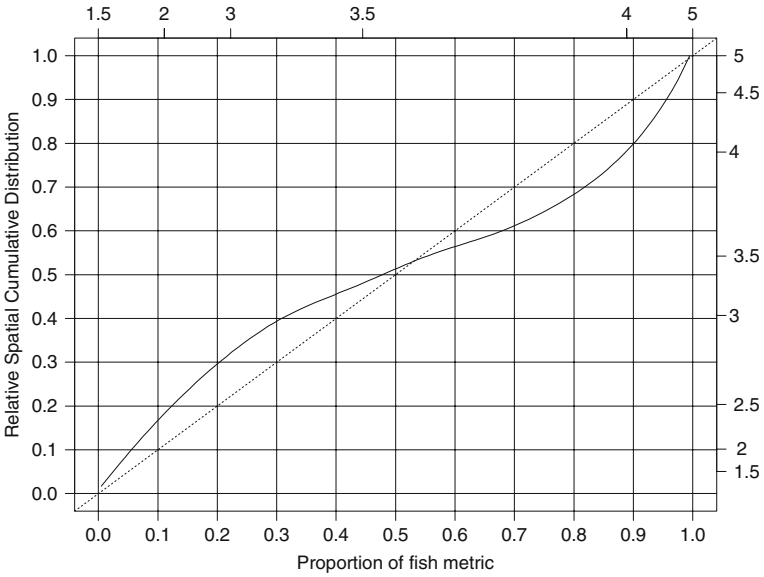

Fig. 3 The relative spatial CDF of fish IBI in the Gunpowder-Patapsco watershed to the Severn watershed. The upper and right axes is labeled in the fish IBI units

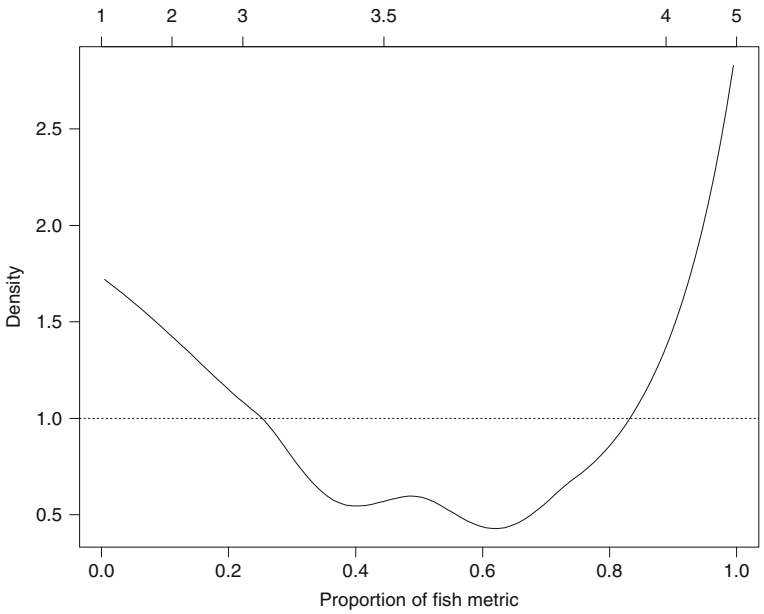

Fig. 4 The relative spatial density of fish IBI in the Gunpowder-Patapsco watershed to the Severn watershed. The upper axis is labeled in the fish IBI units

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 Fig. 2.

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 proportionally 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 Fig. 2); rather it complements it by focusing on those characteristics of the individual distributions essential for scale-free comparison. Figures 2 and 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 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 variation. 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 et al. 1998) or even discrete. We may also wish to consider derived measures of exceedences useful for risk assessment e.g.,

$$E(x) = \mathcal{I}\,(Z(x) \leq L) \qquad \text{for given } L \tag{8}$$

where L is a pre-specified limit on the measure. The generalized linear spatial models approach of Diggle et al. (1998) greatly broaden the form of spatial variation that can be represented by the framework described here. Theirs 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 hierarchal 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 Mugglins and Carlin (1998).

Acknowledgements We would like to thank James McDermott (Penn State University), Tony Olsen and Barbara Rosenbaum (EPA-Corvallis) for help with computing and data compilation issues. This research was supported by the Environmental Protection Agency (EPA) under the Science to Achieve Results (STAR) program Grant # R-82867401-0.

References

Anselin L, Florax R (1995) New directions is spatial econometrics. Springer-Verlag, Berlin
Bernhardt AD, Morris M, Handcock MS (1995) Women's gains or men's losses? A closer look at the shrinking gender gap in earnings. Am J Sociol 101:302–328
Besag J (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). J Roy Stat Soc B 36:192–236
Besag J (1975) Statistical analysis of non-lattice data. The Statistician 24:179–195
Besag J (1989a) A candidate's formula: a curious result in Bayesian prediction. Biometrika 76:183–183
Besag J (1989b) Towards Bayesian image analysis. J Appl Stat 16:395–407
Besag J, Kooperberg C (1995) On conditional and intrinsic autoregression. Biometrika 82:733–746
Besag J, York J, Mollič A (1991) Bayesian Image Restoration, with two applications is spatial statistics (with discussion). Ann Inst Stat Mathemat 43:1–59
Besag J, Higdon D (1999) Bayesian analysis of agricultural field trials (with discussion). J Roy Stat Soc B 61:691–746
Bernardinelli L, Montomoli C (1992) Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. Stat Med 11:983–1007
Best NG, Ickstadt K, Wolpert RL (1998) Spatial Poisson regression for health and exposure data measured at disparate resolutions. Discussion paper, 98-36, Institute of Decision Sciences, Duke University
Butler RJ, McDonald JB (1987) Interdistributional Income Inequality. J Busi Econ Stat 5:13–18
Cox LH, Piegorsch WW (1996) Combining environmental information. I: Environmental monitoring, measurement and assessment. Environmetrics 7:299–308
Cox LH (1998) Workshop: statistical methods for combining environmental information. In: Nychka D, Piegorsch WW, Cox LH (eds) Case studies in environmental statistics, 143–158. Lecture Notes in Statistics, 132. Springer-Verlag
Cressie NAC (1993) Statistics for spatial data. Wiley, New York
Cressie NAC (1995) Bayesian smoothing of rates in small geographic areas. J Region Sci 35:659–673
Cwik J, Mielniczuk J (1989) Estimating density ratios with application to discriminant analysis. Commun Stat 18:3057–3069
Cwik J, Mielniczuk J (1993) Data-dependent bandwidth choice for a grade density kernel estimate. Stat Probabil Lett 16:397–405
Dewald T, Olsen M (1994) The EPA reach file: A national spatial data resource. U.S. Environmental Protection Agency, Office of Water
Ecker MD, Gelfand AE (1997) Bayesian variogram modeling for an isotropic process. J Agric Biol Environ Stat 4:347–369
Fausch KD, Lyons J, Karr JR, Angermeier PL (1990) Fish communities as indicators of environmental degradation. In: Adams SM (ed) Biological Indicators of Stress in Fish, 123–144. American Fisheries Society Symposium 8. Bethesda, Maryland
Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian Data Analysis. Chapman and Hall, London
Handcock MS (1994) Discussion of "Epidemics: Models and Data" by D. Mollison, V. Isham and B. Grenfell (1994). J Roy Stat Soc A 157:115–149
Handcock MS (1998) Discussion of "Model–based Geostatistics" by P. J. Diggle, J. A. Tawn and R. A. Moyeed (1998). J Roy Stat Soc C 47(3):299–350
Handcock MS (1999) Discussion of "Prediction of Spatial Cumulative Distribution Functions Using Subsampling" by S. Lahiri, M. Kaiser, N. Cressie and N. Hsu (1999). J Am Stat Assoc 94(445):100–102
Handcock MS, Morris M (1998) Relative distribution methods. Sociol Methodol 28:53–97
Handcock MS, Morris M (1999) Relative distribution methods in the social sciences. Springer, New York
Handcock MS, Stein ML (1993) A Bayesian analysis of kriging. Technometrics 35(4):403–410
Handcock MS, Wallis J (1994) An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). J Am Stat Assoc 89:368–378. rejoinder, 388–390
Handcock MS, Meier K, Nychka D (1994) Comment on "Kriging and Splines: An Empirical Comparison of their Predictive Performance" by G. M. Laslett. J Am Stat Assoc 89:401–403

Handcock MS, Rendall MS, Huovilainen SM (2000) Combining survey and population data on births and family. Demography 37(2):187–192
Handcock MS, Rendall MS, Cheadle JE (2005) Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. Sociol Methodol 35:303–346
Heimbuch D, Seibel J, Wilson H, 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
Herlihy AT, Larsen DP, Paulsen SG, Urquhart NS, Rosenbaum BJ (2000) Designing a spatially balanced, randomized site selection process for regional stream surveys: The EMAP mid-atlantic pilot study. Environ Monitor Assess 63:95–113
Holmgren EB (1995) The P–P plot as a method for comparing treatment effects. J Am Stat Assoc 90:360–365
Jones KB, Ritters KH, Wickham JD, Tankersley RD, O'Neill RV, Chaloud DJ, Smith ER, Neale AC (1997) An ecological assessment of the United States mid-atlantic region: a landscape atlas. Environmental Protection Agency: EPA/600/R-97/130 (http://www.epa.gov/emap/html/pubs/docs/groupdocs/landecol/atlas/ma_atlas.html).
Karoly LA (1993) The trend in inequality among families, individuals, and workers in the United States: A twenty-five year perspective. In: Danziger S, Gottschalk P (ed) Uneven tides: rising inequality in america. Russell Sage, New York, NY, pp 19–97
Karr JR, Dudley DR (1981) Ecological perspectives on water quality goals. Environ Manage 5:55–68
Karr JR, Fausch, KD, Angermeier PL, Yant PR, Schlosser IJ (1986) Assessing biological integrity in running waters—a method and its rationale: Illinois Natural History Survey Special Publication Number 5, 28 p.
Kelsall JE, Wakefield JC (2002) Modelling spatial variation in disease risk: a geostatistical approach. J Am Stat Assoc 97:692–701.
Kepner WG, Jones KB, Chaloud DJ, Wickham JD (1995) Mid–Atlantic landscape Indicators Project Plan. EPA/620/R-95/003. Washington, D.C.: U.S. Environmental Protection Agency
Lahiri S, Kaiser M, Cressie NAC, Hsu N (1999) Prediction of spatial cumulative distribution functions using subsampling. J Am Stat Assoc 94(445):100–102
Lazorchak JM, Klemm DL, Peck DV (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
Larsen DP, Christie SJ (eds) (1993) EMAP-Surface Waters 1991 Pilot Report. EPA/620/R-93/003. Corvallis, Oregon: U.S. Environmental Protection Agency
Li G, Tiwari RC, Wells MT (1996) Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. J Am Stat Assoc 91:689–698
McLaughlin DK, Handcock MS (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
Messer JJ, Linthurst RA, Overton WS (1991) An EPA program for monitoring ecological status and trends. Environ Monitor Assess 17:67–78
Molliè A, Richardson S (1991) Empirical Bayes estimates of cancer mortality rates using spatial models. Stat Med 10:95–112
Moller J (1998) Log Gaussian cox processes. Scand J Stat 25:451–482
Morris M, Bernhardt AD, Handcock MS (1994) Economic inequality: new methods for new trends. Am Sociol Rev 59:205–219
Mugglin AS, Carlin BP (1998) Hierarchical modeling in geographic information systems: population interpolation over incompatible zones. J Agric Biol Environ Stat 3:111–130
Mugglin AS, Carlin BP, Gelfand AE (2000) Fully Model Based Approaches for Spatially Misaligned Data. J Am Stat Assoc 95:877–887
Mugglin AS, Carlin BP, Zhu L, Conlon E (1999) Bayesian areal interpolation, estimation, and smoothing: an inferential approach for geographic information systems. Environ Plann Stat 31:1337–1352
Owen A (1994) Lattice sampling revisited: Monte Carlo variance of means over randomized orthogonal arrays. Ann Stat 22:930–945
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
Parzen E (1992) Comparison change analysis. In: Saleh A (ed) Nonparametric Statistics and related topics. Elsevier, Holland, pp 3–15
Piegorsch WW, Cox LH (1996) Combining environmental information. II: Environmental epidemiology and toxicology. Environmetrics 7:309–324

Raftery AE, Banfield JD (1991) Stopping the Gibbs sampler, the use of morphology, and other Issues in spatial statistics. Ann Inst Stat Mathemat 43:1–59

Vannote RL, Minshall GW, Cummins KW, Sedell JR, Cushing CE (1980) The River Continuum concept. Can J Fish Aquat Sci 37:130–137

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

Gotway CA, Young LJ (2002) Combining incompatible spatial data. J Am Stat Assoc 97(458):632–648

Zimmerman DL, Harville DA (1991) A Random field approach to the analysis of field-plot experiments and other spatial experiments. Biometrics 47:223–239

Biographical sketches

Mark S. Handcock is a Professor of Statistics, Department of Statistics, University of Washington, Seattle. He works in the fields of spatial statistics and inference for stochastic processes. He also works on the development of statistical models for the analysis of social network data, spatial processes and demography. Recent applications include models for combining information from demographic surveys and population-level information (http://www.stat.washington.edu/handcock). He received his B.Sc. from the University of Western Australia and his Ph.D. from the University of Chicago. He is a member of the *Center for Statistics and the Social Sciences*, the *Center for Studies in Demography and Ecology* and the *National Research Center for Statistics and the Environment.*