Spatial CDF Estimation and Visualization with Applications to Forest Health Monitoring

James J. Majure[1], Dianne Cook[2], Noel Cressie[2], Mark Kaiser[2], Soumendra Lahiri[2], Jürgen Symanzik[2]

[1] GIS Support and Research Facility, Iowa State University

[2] Department of Statistics, Iowa State University, Ames, IA 50011, USA

majure@iastate.edu

Abstract. This paper discusses the estimation and visualization of spatial cumulative distribution functions (CDFs) with extensions to bivariate and higher dimensional CDFs. The use of CDFs is an important part of the USEPA Environmental Monitoring and Assessment Program's (EMAP) work in assessing and monitoring the state of the nation's environmental resources. The resources in a given region can be classified broadly into nominal, marginal, or sub-nominal states. These can be obtained from the spatial CDF which, in its entirety, offers the greatest flexibility for investigation of spatial and temporal trends. The emphasis in this paper is on the computational and graphical techniques implemented in an interactive environment. The environment supports computation and visualization of CDFs over several spatial regions and features interaction between and linking of elements in the CDF plot and the map view. The work involves communication of data between a geographic information system (GIS), ArcView 2.0^{TM}, and a program for dynamic graphics, XGobi (Swayne et al., 1991).

1 Introduction

Large-scale monitoring of ecological resources is of vital importance to the nation. Changes in the environment of a long-term and global nature are almost impossible to detect from local ecological studies taken sporadically in space and time. If one observes a degradation in an ecological resource in such a local study, it is not clear whether it is more regional in scope or whether it is due to a unique condition at the study site. Furthermore, if there are no repeat visits to a site, it is not clear whether the observed ecological condition is degrading or improving over time. Further discussion of the importance of well designed, large-scale, long-term monitoring programs can be found in the articles by Messer, Linthurst, and Overton (1991) and Stevens (1994).

The U.S. Environmental Protection Agency's (EPA) Environmental Monitoring and Assessment Program (EMAP) represents a commitment by the EPA to be more proactive about the condition of the nation's environment. Up until 1987, almost all of the EPA's budget involved detecting relatively local insults to the environment and then following up with fines, litigation, and remediation recommendations. To use a medical analogy, the EPA often acted as a physician to a rather large and at times unco-operative patient, discovering and excising isolated illnesses as they arose. A public-health approach would be to give the patient periodic physicals according to a protocol made up of a list of benign diagnostic procedures. Good public health policies pay enormous dividends in the long term, in spite of what may seem like a considerable short-term expenditure. The EPA's EMAP has the potential to reap these same sorts of dividends in an area of vital importance to the nation and its environment.

The overall objectives of EMAP are to:

1. Estimate current status of and trends and changes in, selected indicators of the condition of the Nation's ecological resources on a regional basis with known statistical confidence.
2. Estimate the geographic coverage and extent of the Nation's ecological resources with known confidence.
3. Seek associations between selected indicators of natural and anthropogenic stresses and indicators of the condition of ecological resources.
4. Provide annual statistical summaries and periodic assessments of the Nation's ecological resources.

These four objectives all speak to the need for understanding the condition of the nation's ecological resources at a regional or even local level. This is explicitly stated in the first goal ("Estimate the current status...") and the second ("Estimate the geographic coverage...") with words such as "regional basis," "geographic coverage and extent". It is equally important but implicitly stated in the third goal ("Seek associations...") and the fourth ("Provide annual statistical summaries..."). What is called the ecological fallacy (Robinson, 1950) is the presence of a relationship between two variables at an aggregated level that is due simply to the aggregation rather than to any real link. It has been known for some time that successive aggregation of regionalized variables tends to increase correlations, even though the correlation at the disaggregated level is zero (Yule and Kendall, 1950, pp. 310-313; Openshaw and Taylor, 1979). Therefore, the summaries provided and the associations sought in EMAP need to be based on estimates at the most local level possible, yet still with enough precision to allow meaningful conclusions. Consequently, spatial statistics is highly relevant to the goals of EMAP.

We have formed a research team at Iowa State University to carry out spatial statistics research applied to ecological

resource monitoring programs. An important component of that research is estimating, visualizing, exploring, and comparing spatial cumulative distribution functions (CDFs) over different spatial regions. The main thrust of this paper is a discussion of the graphical and computational techniques being developed to support this work. In addition to this, we shall also discuss some of our work on associated confirmatory methods to test and compare spatial CDFs. However, before these discussions can proceed the theoretical and philosophical basis for our work must be established.

Consider a multivariate space-time process,

$$\{\boldsymbol{Z}(\boldsymbol{s}; v) : \boldsymbol{s} \in D, v \in T\},$$

where here $\boldsymbol{Z}$ is a vector of ecological indices, D is the spatial domain of interest (think of $D \subset I\!R^2$) and T is an index set representing some beginning and then all future time points. For the study that we are considering in this paper, $\boldsymbol{Z}$ is the bivariate process that represents the crown dieback index and the foliage transparency index; D is the northeast USA given by the states of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island and Connecticut, and $T = \{1991\}$. In fact, because we believe these processes are relatively stable over time, a regional investigation of the process at just one recent time point will allow us to draw conclusions about the regions that should still be relevant to the present-day. As more data become available, we intend to use our methods to look for both regional *and* temporal changes.

Henceforth, consider the multivariate spatial process

$$\{\boldsymbol{Z}(\boldsymbol{s}) : \boldsymbol{s} \in D_0\}, \tag{1.1}$$

where D_0 represents forested locations in D. There is a scaling issue of when individual trees begin to look like a forest, in which case one can represent the ecological index as a random field with continuous spatial index. Because the field data are taken over a small study site, which we shall denote as Δ, we chose this as our standard area. Its geometric configuration is given in Figure 1.

Henceforth, we shall define Δ as the *spatial support unit* (SSU). Thus, at location $\boldsymbol{s}$, we have SSU $\Delta(\boldsymbol{s})$ and

$$\boldsymbol{Z}(\boldsymbol{s}) \equiv \frac{\sum_{i=1}^{n(\boldsymbol{s})} DBH_i \cdot (CDB_i, FTR_i)}{\sum_{i=1}^{n(\boldsymbol{s})} DBH_i}; \ \boldsymbol{s} \in D_0, \tag{1.2}$$

where the indices $\boldsymbol{Z}(\boldsymbol{s})$ represent the nature of the tree crown as a response to stressors and are composed of two observables, *crown dieback* (CDB) and *foliage transparency* (FTR). The size of each tree is measured by its *diameter at breast height* (DBH).

The choice of SSU can affect inferences, a phenomenon we referred to earlier as the ecological fallacy. In the geography literature, this is referred to as the modifiable areal unit problem (e.g., Openshaw and Taylor, 1979). The choice we have made comes from a desire to match SSUs with the ones

Figure 1: Geometric configuration of forest sampling site. Reprinted from Tallent-Halsell, 1994.

that are used in the field. Further, SSU Δ also represents an area large enough to allow sufficient spatial averaging in (3) yet small enough to capture local fluctuations in small geographic areas.

A number of definitions are needed. First, we shall use the convention

$$\{\boldsymbol{Z}(\boldsymbol{s}) \leq \boldsymbol{z}\} \equiv \{Z_1(\boldsymbol{s}) \leq z_1\} \cap \{Z_2(\boldsymbol{s}) \leq z_2\}, \tag{1.3}$$

where $\boldsymbol{Z}(\boldsymbol{s}) = (Z_1(\boldsymbol{s}), Z_2(\boldsymbol{s}))$ and $\boldsymbol{z} = (z_1, z_2)$. The spatial CDF F_∞ is defined as follows:

$$F_\infty(\boldsymbol{z}; B_0) \equiv \int_{B_0} I(\boldsymbol{Z}(\boldsymbol{u}) \leq \boldsymbol{z}) d\boldsymbol{u} / |B_0|; \boldsymbol{z} \in I\!R^2, \tag{1.4}$$

where $B \subset D$ is a well defined region of interest whose forested part is

$$B_0 \equiv B \cap D_0; \tag{1.5}$$

$|B_0|$ denotes the area of B_0; and $I(A)$ denotes the indicator function equal to one if A is true and equal to zero otherwise. Should $B_0 = \emptyset$, we say that F_∞ is undefined. On the other hand, if

$$B_0 = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots\}, \tag{1.6}$$

a countable collection of locations, then

$$F_\infty(\boldsymbol{z}; B_0) \equiv \lim_{M \to \infty} \sum_{i=1}^{M} I(\boldsymbol{Z}(\boldsymbol{u}_i) \leq \boldsymbol{z}) / \sum_{i=1}^{M} 1. \tag{1.7}$$

Indeed, given the nature of sampling, we shall have to approximate the conceptual CDF (1.4) with a (finite) version like (1.7). In what is to follow, we shall tesselate the region B_0 into "tiles" made up of Thematic Mapper (TM) pixels (30m×30m). Let

$$B_0 = \bigcup_{i=1}^{N(B_0)} \{A(\boldsymbol{u}_i)\}, \tag{1.8}$$

where $A(\boldsymbol{u}_i)$ represents the TM pixel defined at center point $\boldsymbol{u}_i$. There are $N(B_0)$ such pixels that make up B_0. On occasions, when discretization of the continuum in B_0 is necessary, we use (1.8) and replace the definition (1.4) with

$$F_\infty(\boldsymbol{z}; B_0) \equiv \sum_{i=1}^{N(B_0)} I(\boldsymbol{Z}(\boldsymbol{u}_i) \leq \boldsymbol{z})/N(B_0), \tag{1.9}$$

where $\boldsymbol{Z}(\boldsymbol{u}_i)$ refers to the CDB and FTR indices defined over $\Delta(\boldsymbol{u}_i)$ located at the point $\boldsymbol{u}_i; i = 1, ..., N(B_0)$.

Notice that we have effectively replaced the process $\boldsymbol{Z}(\boldsymbol{u}); \boldsymbol{u} \in D_0$, with a discrete process

$$\{\boldsymbol{Z}(\boldsymbol{u}_i) : i = 1, \ldots, N(D_0)\}, \tag{1.10}$$

where $N(D_0)$ is the number of TM pixels that tesselate D_0 in a manner analogous to (1.8). This discretization is essential for making progress but does introduce an approximation, the effect of which deserves further study.

At this juncture, it is worthwhile emphasizing the difference between spatial CDFs and theoretical CDFs. Consider observations $\{\boldsymbol{Z}(\boldsymbol{s}_1), \ldots, \boldsymbol{Z}(\boldsymbol{s}_m)\}$ on the process $\boldsymbol{Z}(\cdot)$ at any set of locations $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m\}$ in D_0. Then their theoretical CDF is given by,

$$G_{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_m}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m) \equiv Pr(\boldsymbol{Z}(\boldsymbol{s}_1) \leq \boldsymbol{z}_1, \ldots, \boldsymbol{Z}(\boldsymbol{s}_m) \leq \boldsymbol{z}_m). \tag{1.11}$$

In particular, in Section 4, we shall be interested in $G_{\boldsymbol{s}}(\boldsymbol{z})$ and $G_{\boldsymbol{s}_1,\boldsymbol{s}_2}(\boldsymbol{z}_1, \boldsymbol{z}_2)$. Notice that the theoretical CDF is a *parameter* of the process $\boldsymbol{Z}(\cdot)$, whereas the spatial CDF is a measurable function of $\boldsymbol{Z}(\cdot)$ and hence it is a random quantity. Our goal in this research is to predict or, loosely speaking, estimate the spatial CDF F_∞ of a region B_0. In EMAP, the emphasis is on the ecological populations rather than the theoretical population of an assumed statistical model.

Available to researchers are data from the field. Define

$$\boldsymbol{Z} \equiv (\boldsymbol{Z}(\boldsymbol{s}_1)', \ldots, \boldsymbol{Z}(\boldsymbol{s}_n)')', \tag{1.12}$$

obtained from sampling locations $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$. Notice that it is highly unlikely that $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{N(D_0)}\}$ and $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ share any locations, which means it is important to maintain the *conceptual* continuous spatial model.

Having established the theoretical basis for our work, Section 2 of this paper will briefly consider estimation of (1.4) or (1.9) based on data $\boldsymbol{Z}$. Section 3 will describe the dynamic graphical environment developed to visualize and explore spatial CDFs. Visualization of both univariate and bivariate spatial CDFs will be illustrated. Section 4 will introduce the associated confirmatory methods being developed to test and compare univariate spatial CDFs. Section 5 will conclude with a summary and discussion of future research directions.

2 Spatial CDF Estimation

As developed here, estimation of a spatial CDF refers to estimation of either the continuous version or the discrete version of $F_\infty(\boldsymbol{z}; B_0)$, given in equations (1.4) and (1.9), respectively. In what is to follow, we use the discrete version (1.9), which allows a unified development of the estimation problem. We assume that, connected with the spatial sampling locations $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, there are a set of known weights $\{u(\boldsymbol{s}_1), \ldots, u(\boldsymbol{s}_n)\}$. These might, for example, correspond to inclusion probabilities from a sampling design or to importance weights in a resource management plan. A basic estimator of (1.9) is

$$\hat{F}_n(\boldsymbol{z}; B_0) = \frac{\sum_{i=1}^n u(\boldsymbol{s}_i) I(\boldsymbol{Z}(\boldsymbol{s}_i) \leq \boldsymbol{z})}{\sum_{i=1}^n u(\boldsymbol{s}_i)}. \tag{2.1}$$

Notice that the CDF (1.9) is of the same form as the estimator (2.1) with $u(\boldsymbol{s}_i) \equiv 1$, n replaced by $N(B_0)$, and $\boldsymbol{Z}(\boldsymbol{s}_i)$ replaced by $\boldsymbol{Z}(\boldsymbol{u}_i)$. In the rest of this section, we consider the simpler univariate case, $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D_0\}$. The first two moments of $\hat{F}_n(z; B_0)$ given by (2.1) are

$$E\left(\hat{F}_n(z; B_0)\right) = \frac{\sum_{i=1}^n u(\boldsymbol{s}_i)\, Pr(\, Z(\boldsymbol{s}_i) \leq z\,)}{\sum_{i=1}^n u(\boldsymbol{s}_i)}, \tag{2.2}$$

which is a weighted average of the theoretical CDFs $Pr(\, Z(\boldsymbol{s}_i) \leq z\,) \equiv G_{\boldsymbol{s}_i}(z)$; and

$$E\left(\hat{F}_n^2(z; B_0)\right) = \frac{\sum_{i=1}^n \sum_{j=1}^n u(\boldsymbol{s}_i)\, u(\boldsymbol{s}_j)\, Pr\,(\, Z(\boldsymbol{s}_i) \leq z\,,\, Z(\boldsymbol{s}_j) \leq z\,)}{\left(\sum_{i=1}^n u(\boldsymbol{s}_i)\right)^2}.$$

Therefore,

$$\text{var}\left(\hat{F}_n(z; B_0)\right) = \sum_{i=1}^n \sum_{j=1}^n u(\boldsymbol{s}_i)\, u(\boldsymbol{s}_j)\, \{\, Pr\,(\, Z(\boldsymbol{s}_i) \leq z\,,\, Z(\boldsymbol{s}_j) \leq z\,) \; - \; Pr\,(\, Z(\boldsymbol{s}_i) \leq z)\; Pr\,(\, Z(\boldsymbol{s}_j) \leq z)\} \cdot \left(\sum_{i=1}^n u(\boldsymbol{s}_i)\right)^{-2} \tag{2.3}$$

If we assume that $Z(\cdot)$ in (1.1) has an invariant distribution, so that $G_{\boldsymbol{s}}(z) = G_{\boldsymbol{0}}(z)$ does not depend on $\boldsymbol{s}$, then $\hat{F}_n(z; B_0)$ is an unbiased estimator of $G_{\boldsymbol{0}}(z)$ for any $B_0 \subseteq D_0$, and (2.3)

may be written as

$$\text{var}\left(\hat{F}_n(z;\, B_0)\right) = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} u(\boldsymbol{s}_i)\, u(\boldsymbol{s}_j) \left\{ G_{\boldsymbol{s}_i,\boldsymbol{s}_j}(z,z) - G_{\mathbf{0}}^2(z) \right\} \right\} / \left(\sum_{i=1}^{n} u(\boldsymbol{s}_i) \right)^2 , \tag{2.4}$$

where $G_{\boldsymbol{s},\boldsymbol{s}'}$ denotes the joint theoretical CDF of $Z(\boldsymbol{s})$ and $Z(\boldsymbol{s}')$. Notice that, when $\boldsymbol{s} = \boldsymbol{s}'$, $G_{\boldsymbol{s},\boldsymbol{s}'}(z,z) = G_{\boldsymbol{s}}(z) = G_{\mathbf{0}}(z)$.

Of central importance in the use of spatial CDFs is the comparison of two different estimators of the form (2.1). Such estimators may correspond to different weights for the same locations in a given region, different overlapping sets of locations in a region, or different subregions. In addition, because the quantity to be estimated, (1.9), is of the same form as the general estimator (2.1), comparison of an estimator with the corresponding true spatial CDF takes the same form as the comparison of two estimators. Consider two estimators $\hat{F}_{n,1}$ and $\hat{F}_{m,2}$ defined by equation (2.1) over subregions of D_0 with associated locations

$$\omega_1 = \left\{ i : Z(\boldsymbol{s}_i) \text{ contributes to } \hat{F}_{n,1} \right\}$$

$$\omega_2 = \left\{ j : Z(\boldsymbol{s}_j) \text{ contributes to } \hat{F}_{m,2} \right\}$$

and weights

$$\{u(\boldsymbol{s}_i) \,:\, i \in \omega_1\} \quad , \quad \{v(\boldsymbol{s}_j) \,:\, j \in \omega_2\}\,.$$

Here, $|\omega_1| = n$ and $|\omega_2| = m$ and we write the difference between the two estimators as

$$\Delta_{n,m}(z) = \hat{F}_{n,1}(z) - \hat{F}_{m,2}(z). \tag{2.5}$$

Consequently,

$$\text{var}\left(\Delta_{n,m}(z)\right) = \text{var}\left(\hat{F}_{n,1}(z)\right) + \text{var}\left(\hat{F}_{m,2}(z)\right) - 2\text{cov}\left(\hat{F}_{n,1}(z),\, \hat{F}_{m,2}(z)\right) . \tag{2.6}$$

Again, assuming an invariant distribution for $Z(\cdot)$, $E(\Delta_{n,m}(z)) = 0$ and

$$\text{var}\left(\Delta_{n,m}(z)\right) = E\left(\Delta_{n,m}^2(z)\right) = \frac{1}{\left(\sum_{i\in\omega_1} u(\boldsymbol{s}_i)\right)^2} \sum_{i\in\omega_1}\sum_{j\in\omega_1} u(\boldsymbol{s}_i)u(\boldsymbol{s}_j) G_{\boldsymbol{s}_i,\boldsymbol{s}_j}(z,z) + \frac{1}{\left(\sum_{i\in\omega_2} v(\boldsymbol{s}_i)\right)^2} \sum_{i\in\omega_2}\sum_{j\in\omega_2} v(\boldsymbol{s}_i)v(\boldsymbol{s}_j) G_{\boldsymbol{s}_i,\boldsymbol{s}_j}(z,z) - \frac{2}{\left(\sum_{i\in\omega_1} u(\boldsymbol{s}_i)\right)\left(\sum_{i\in\omega_2} v(\boldsymbol{s}_i)\right)} \sum_{i\in\omega_1}\sum_{j\in\omega_2} u(\boldsymbol{s}_i)v(\boldsymbol{s}_j) G_{\boldsymbol{s}_i,\boldsymbol{s}_j}(z,z)\,. \tag{2.7}$$

If $\hat{F}_{m,2}$ is taken to be the spatial CDF F_∞, then (2.7) represents a *mean squared prediction error* (MSPE). If both $\hat{F}_{n,1}$ and $\hat{F}_{m,2}$ are estimators of the spatial CDF, then (2.7) is the variance of the difference between these two estimators at the value z. In either case it is not sufficient for inferential purposes to know variances of CDF estimators only. One must also know covariances. Equivalently $E\left(\Delta_{n,m}^2(z)\right)$ may be obtained directly through (2.7). We have already noted that, upon taking $\hat{F}_{m,2}$ to be the true spatial CDF F_∞, the right hand side of (2.7) is the MSPE. This expression may be compared with the variance (2.4). This latter expression is a measure of expected squared difference of the estimator $\hat{F}_n$ from its own expected value, namely the theoretical CDF $G_{\mathbf{0}}$. The MSPE given in equation (2.7) is a measure of the expected squared difference of the estimator from the true spatial CDF F_∞.

Estimation of both (2.4) and (2.7) depends only on the ability to estimate joint theoretical CDFs $G_{\boldsymbol{s},\boldsymbol{s}'}$ for pairs of locations. Indeed, one must estimate these joint probabilities for *any* pair of locations in D_0, not only locations that have contributed to a particular estimator.

3 Exploratory Analysis of Regional CDF Estimates

The topic of this section is the development of a dynamic graphical environment to visualize and explore spatial CDFs. We view graphical methods as an integral part of the analysis process. Graphical methods are used to identify structure within the data, as well as the type of structure (e.g., linear, non-linear, spatial) that is present. These methods are used to check the appropriateness of any assumptions that underly methods to be used. They are also used to see if fitted models are parsimonious with the data to which they are fit.

Our objectives in the development of these methods are 1) that they are very dynamic, allowing much analyst interaction; and 2) that they maintain a strong connection between the spatial locations of the sampling sites and the collected data. These objectives have been accomplished by developing a bi-directional link between ArcView 2.0[TM], a *geographic information system* (GIS), and XGobi, a dynamic graphical data analysis system (Swayne, et al., 1991). For a more detailed discussion on the technical aspects of this link see Symanzik, Majure, and Cook (1995).

The main motivation for attempting to link a GIS with XGobi was the desire to maintain a strong connection between the spatial locations of the sampling sites and the collected data. The use of a GIS allows the data to be analyzed in the visual context that concomitant geographic variables provide. In a GIS, the sampling locations can be viewed in a setting of streams, roads, cities, soil types, and topography, just to name a few.

The ArcView 2.0 GIS was chosen for several reasons, the two most important of which were the ability to conduct

[TM] *ArcView 2.0* is a trademark of Environmental Systems Research Institute, Inc.

inter-process communications and the customizable, interactive interface. The use of the *remote procedure calls* (RPCs) for interprocess communications between ArcView 2.0 and XGobi made the link quick enough for interactive use. RPC servers were established in both systems so that user actions taken in either system could be relayed to the other. ArcView 2.0's user interface also makes displaying various geographic features, and panning and zooming within the displayed region easy.

XGobi is a dynamic graphical system that allows the manipulation of scatter plots of highly multivariate data. Some of the capabilities of XGobi include three-dimensional rotation, grand tour rotation (Asimov, 1985, Buja and Asimov 1986),projection pursuit and linked color and glyph brushing. XGobi also allows points to be connected with lines, a capability used to implement the current graphics environment.

The capabilities of the developed graphics environment include: 1) the ability to delineate regions in the GIS and to see the corresponding empirical CDF estimate; 2) the ability to brush points in areas for which CDF estimates have already been defined and see where they lie in the CDF; and 3) the ability to brush quantiles of the estimated CDFs and see where the corresponding points lie in the spatial region. These capabilities are demonstrated by conducting an analysis on a univariate ecological index as described in the following paragraphs.

In this analysis we examine the CDF computed on the *crown defoliation index* (CDI) which is a weighted average of the two variables, *crown dieback* (CDB) and *foliage transparency* (FTR) introduced in Section 1.

$$Z(\boldsymbol{s}) \equiv \frac{\sum_{i=1}^{n(\boldsymbol{S})} DBH_i \cdot (CDB_i + FTR_i)/2}{\sum_{i=1}^{n(\boldsymbol{S})} DBH_i}; \; \boldsymbol{s} \in D_0,$$

where DBH_i is the *diameter at breast height* of tree $i; i = 1, \ldots, n(\boldsymbol{s})$.

Crown dieback refers to the percentage of dead branches in the upper sunlight exposed parts of the tree crown. The assumption is that these branches have died from stressors in the environment other than lack of light. It is measured as a percentage in increments of 5 from 0 to 100. Foliage transparency refers to the amount of light penetrating foliated branches. It ignores "holes" in the tree due to bare branches and is measured on the same scale as crown dieback.

Figure 2 shows the CDFs computed and plotted for two regions. The CDF estimator is that given in equation (2.1). The black CDF is computed on values from the state of Maine and the grey CDF is computed over the remaining five states in the northeast United States, indicated by the polygonal brushing done in the map view. The two CDFs differ markedly in the middle (around the value 500 on the horizontal axis), where the state of Maine has higher values. This indicates that the proportion of trees with lower crown defoliation index values in the state of Maine is higher, suggesting that the forests in the five southern New England states are reacting more negatively to stressors than those of Maine. This would need to be examined in more detail and a quantification of variability in the CDF estimates is required before conclusive statements can be made.

Once major regions are isolated it is possible to brush transiently in further subregions to examine the position of these values in the CDF plot. In Figure 3, the states of Connecticut and Rhode Island have been brushed in this way, and it can be seen that the index values fall in the central to lower portion of the CDF. This indicates that the trees in this subregion have lower crown dieback and foliage transparency, and so appear to be healthier than trees in the rest of the region.

Now we turn to the idea of examining the three categories of resource health: nominal (good), marginal, and sub-nominal (poor). The category sub-nominal corresponds to high index values. These high values are brushed in the CDF plot (Figure 4) and the sampling sites corresponding to these index values are highlighted in the map view. The map view shows that there is a smaller proportion of sampling sites with these high index values in Maine, echoing the difference just noted in the CDFs.

Finally one of the strengths of integrating these high interaction graphical tools with a GIS is that further concomitant variables can be overlaid in the map view. Figure 4 shows the population density on the map. There does not seem to be any association between population density and high crown defoliation index values, although it is clear that there are fewer sampling sites in the more heavily populated areas.

Extensions to higher-dimensional CDFs are relatively straightforward. Because XGobi is designed to handle high-dimensional data, it is a relatively simple matter to calculate high-dimensional CDFs and pass the variables, for example, three variables for a bivariate CDF (two sets of index values and a function value). The CDF can be viewed using lower dimensional projections, for example, rotation, a grand tour, or a correlation tour. The CDFs are not smooth and function values exist only where measurements have been taken. But, for the quick and exploratory approach we have adopted, the method suffices for now and can be reasonably effective when the number of sample points is large. In general, plotting of high-dimensional functions is not intuitive. Some work on visualization of three- and four-dimensional density functions can be found in Scott (1992).

Figure 5 shows three projections of two bivariate CDFs of the measurements CDB and FTR, one CDF is computed on measurements in the state of Maine (black), and the second is computed on measurements in the remaining states (grey). The left graph in this figure displays the CDF value vertically and a projection containing equal proportions of CDB and FTR horizontally. A difference between the two CDFs can be seen in the middle values and this reflects the difference observed above in the CDF for the crown defoliation index. This is not surprising since the index is based on an average of the two measurements and the projection

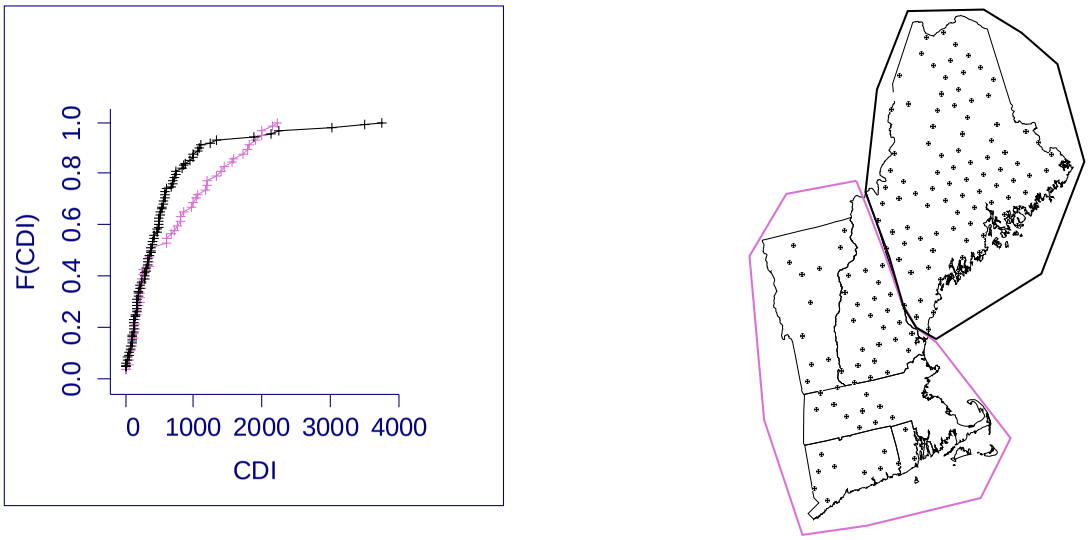

Figure 2: The map view showing two spatial regions and the CDF view showing corresponding empirical CDFs of the CDI. One can see that the CDF for Maine is steeper in the center which indicates that more trees have lower CDI; that is, the trees in Maine appear to be healthier.

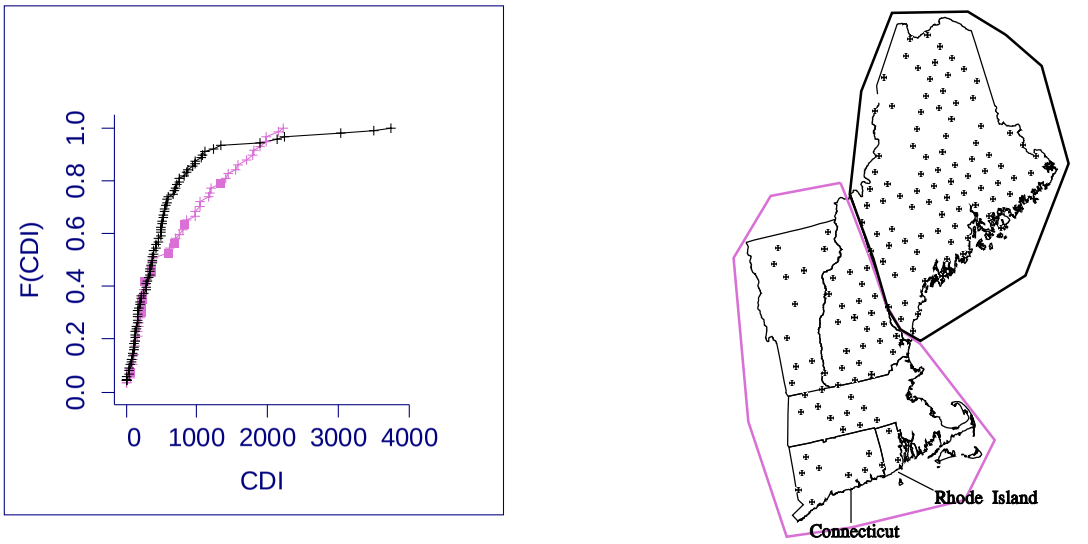

Figure 3: Highlighted values, indicated by filled boxes, are sample locations in Connecticut and Rhode Island.

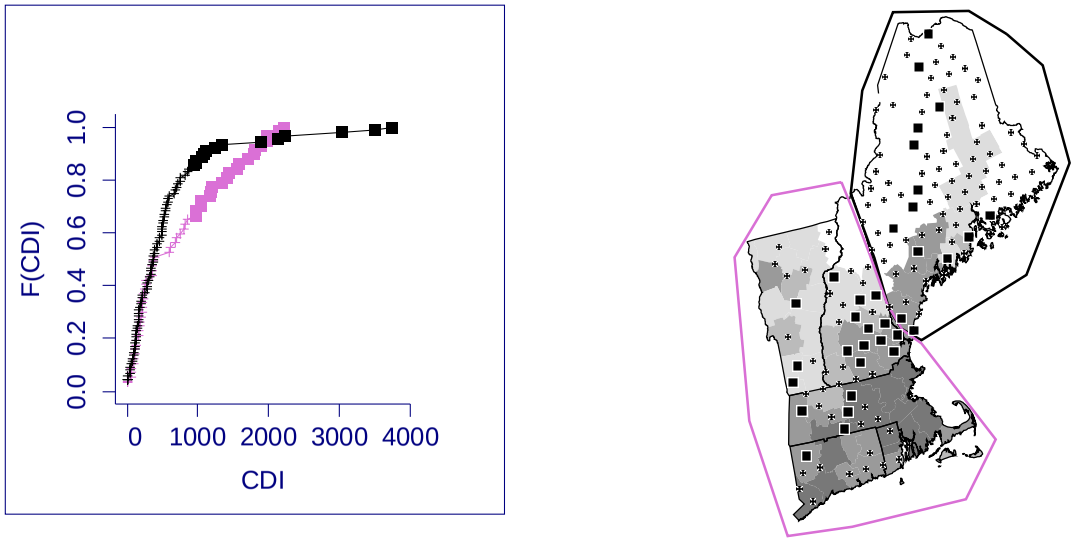

Figure 4: High CDI values brushed in the CDF view and shown as filled boxes in the map view. Population density is overlaid in the map view.

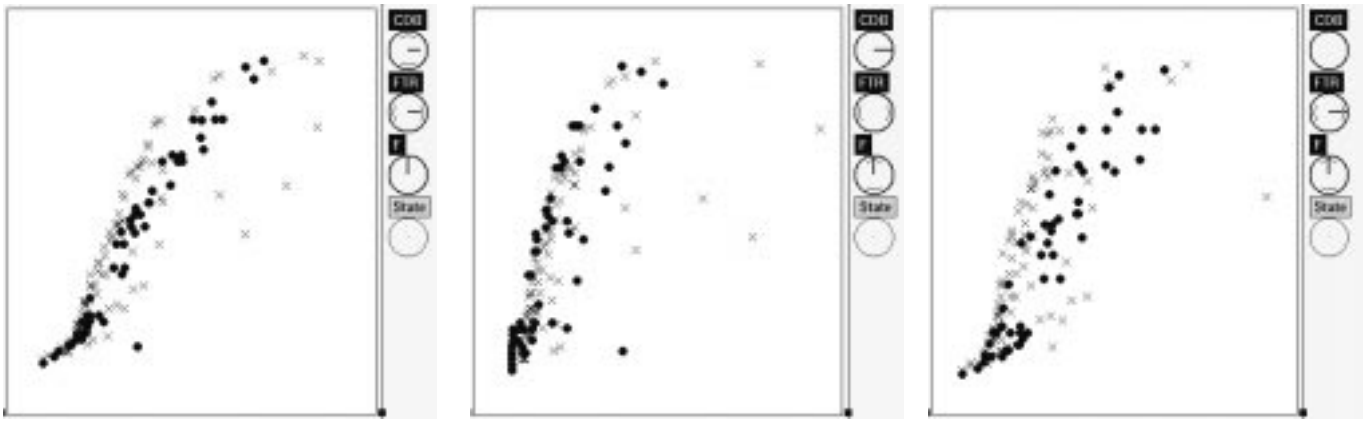

Figure 5: Projections of the estimated bivariate CDF for CDB and FTR. *(left)* equal contribution of both variables horizontally, shows difference between the two regions as seen in the univariate CDF of the crown defoliation index, *(middle)* projection in CDB direction shows very little difference between the two regions, and *(right)* projection in FTR direction shows most difference between the two regions.

also involves equal contribution from each index. The middle graph displays the projection onto CDB, horizontally. There is no difference between the two regions here. The right graph displays the projection onto FTR, horizontally, and here one can see a large difference between the two regions. The interpretation is that for this data set the only difference between the two regions occurs in the FTR measurements. Using the two-dimensional CDF instead of a one-dimensional CDF of an index of the measurements allowed a more precise interpretation. It is common in assessing environmental resources to use aggregated indices rather than the high-dimensional measurements themselves but it is not always clear in ad-

vance how to weight the indices in the aggregation. This example shows a reason why it is important to keep the full dimensionality.

4 Inference for Spatial CDFs

In this section, we consider inference for estimated spatial CDFs. Quantities of central concern to the inferential process are measures of squared discrepancy between an estimator and the corresponding true spatial CDF, its expected value, or another estimator. We define the general criterion of *weighted mean integrated squared error* (WMISE) as

$$Q(n, \infty; w) \equiv E\left[\int_{-\infty}^{\infty} \left\{\hat{F}_n(z; B_0) - F_\infty(z; B_0)\right\}^2 w(z)\, dz\right], \tag{4.1}$$

where $F_\infty(z; B_0)$ is either the continuous (1.4) or discrete (1.9) version of the true spatial CDF, and $w : \mathbb{R} \to \mathbb{R}$ is a non-negative, integrable weight function. Similarly, define the *weighted mean integrated squared distance* (WMISD) between two CDF estimators as

$$Q(n, m; w) \equiv E\left[\int_{-\infty}^{\infty} \left\{\hat{F}_{n,1}(z; B_0) - \hat{F}_{m,2}(z; B_0)\right\}^2 w(z)\, dz\right]. \tag{4.2}$$

As noted in Section 2, when the spatial CDF $F_\infty(z; B_0)$ is taken as the discrete version in equation (1.9), then WMISE in (4.1) has the same form as WMISD, in which case only equation (4.2) need be considered.

An important special case is if the weight function $w(\cdot)$ is an indicator function for $z \in (a, b)$, in which case $Q(n, \cdot; w)$ is a measure of discrepancy for those z-values restricted to the interval (a, b). Thus, through appropriate choice of a and b, WMISE and WMISD can be used to assess the behavior of CDF estimators over subnominal, marginal, or nominal levels of the ecological index of concern. While the quantities WMISE and WMISD are the primary focus for inference on spatial CDFs, it may also be desired to estimate the variance of an estimator, given by equation (2.3).

We are currently pursuing two approaches to estimation of the quantities $Q(n, \cdot; w)$ and $\mathrm{var}\left(\hat{F}_n(z\,; B_0)\right)$. The first is based on straightforward estimation of joint theoretical CDFs for pairs of locations through indicator variograms. The second approach is more flexible and is based on the use of subsampling procedures. Both methods will allow inference on spatial CDFs to be carried out.

5 Future Directions

The work being reported on in this paper is on-going. The progress made thus far will provide a firm basis for further work. Specific areas for future research include:

1. *Inference on spatial CDFs.* The work on inference on spatial CDFs is continuing through indicator variograms and subsampling methods (as described above).
2. *Visualization of variation.* The graphical tools that we have developed thus far do not include any capabilities to visualize the variability of spatial CDF estimates. In order for the graphical analysis of spatial CDFs to be effective, visualization of variability is paramount. We intend to further develop our tools in this area.
3. *Inclusion of additional graphical tools for spatial analysis.* The visualization technology described in Section 3 provides a platform for additional types of graphical tools for spatial analysis. In addition to CDFs, it is possible to display spatially lagged scatter plots and variogram cloud plots while maintaining the link between the graphic and the spatial locations.
4. *Use of remotely sensed images for improving spatial CDF estimation.* Satellite images of the spatial domain provide a potentially invaluable source of concomitant information that can be exploited in order to increase the precision of spatial CDF estimates.

When completed, this research will provide a set of tools, both theoretical and applied that can be used for the effective analysis of broad-based, ecological resource monitoring problems.

Acknowledgements

Research related to this article was supported by an EPA EMAP grant under cooperative agreement # CR822919. The article has not been subjected to the review of the EPA and thus does not necessarily reflect the view of the agency and no official endorsement should be inferred. Symanzik's research was also partially supported by a German "DAAD–Doktorandenstipendium aus Mitteln des zweiten Hochschulsonderprogramms".

References

Asimov, D. (1985). The grand tour: A tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing*, 6(1):128–143.

Buja, A. and Asimov, D. (1986). Grand tour methods: an outline. *Computing Science and Statistics*, 17:63–67.

Messer, J. J., Linthurst, R. A., and Overton, W. S. (1991). An EPA program for monitoring ecological status and trends. *Environmental Monitoring Assessment*, 17:67–78.

Openshaw, S. and Taylor, P. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In Wrigley, N., editor, *Statistical Applications in the Spatial Sciences*, pages 127–144, London, UK. Pion.

Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–357.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York, NY, Wiley.

Stevens, D. (1994). Implementation of a national monitoring program. *Journal of Environmental Management*, 42:1–29.

Swayne, D. F., Cook, D., and Buja, A. (1991). XGobi: Interactive dynamic graphics in the X window system with a link to S. In *ASA Proceedings of the Section on Statistical Graphics*, pages 1–8, Alexandria, VA. American Statistical Association.

Symanzik, J., Majure, J. J., and Cook, D. (1995). Dynamic graphics in a GIS: A bidirectional link between ArcView 2.0 and XGobi. In Rosenberger, J. and Meyer, M., editors, *Proceedings of the 27th Symposium on the Interface between Comput. Sci. and Statist.*, Pittsburgh, PA. Interface Foundation of North America.

Tallent-Halsell, N. G. e. (1994). Forest health monitoring 1994 field methods guide. Technical report, U.S. Environmental Protection Agency, Washington, D.C.

Yule, G. and Kendall, M. (1950). *An Introduction to the Theory of Statistics*. London, UK., Griffin.