

PENNSTATE



---

## Center for Statistical Ecology and Environmental Statistics

---

### HOTSPOT GEOINFORMATICS, ENVIRONMENTAL RISK, AND DIGITAL GOVERNANCE

by  
**G.P. Patil<sup>1</sup>, S.W. Joshi<sup>2</sup>, and S.L. Rathbun<sup>3</sup>**

<sup>1</sup>Center for Statistical Ecology and Environmental Statistics, Department of Statistics  
The Pennsylvania State University, University Park, PA, 16802

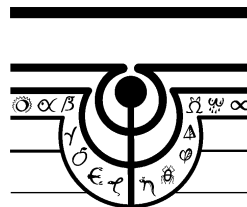
<sup>2</sup>Department of Computer Science, Slippery Rock University,  
Slippery Rock, PA 16057

<sup>3</sup>Department of Health Administration, Biostatistics, and Epidemiology,  
University of Georgia, Athens, GA 30602

This material is based upon work supported by (1) the National Science Foundation under Grant No. 0307010, and (ii) The United States Environmental Protection Agency under Grant No. CR-83059301 and No. R-828684-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

[Invited Paper for the Encyclopedia of Quantitative Risk Analysis]

Technical Report Number 2007-0226  
TECHNICAL REPORTS AND REPRINTS SERIES  
February 2007



---

Department of Statistics  
The Pennsylvania State University  
University Park, PA 16802

G. P. Patil  
Distinguished Professor and Director  
Tel: (814)865-9442 Fax: (814)865-1278  
Email: [gpp@stat.psu.edu](mailto:gpp@stat.psu.edu)

<http://www.stat.psu.edu/~gpp>  
<http://www.stat.psu.edu/hotspots>  
DGOnline News

[Environmental and Ecological Statistics-Springer](#)

# Hotspot Geoinformatics, Environmental Risk, and Digital Governance \*

by  
G.P. Patil<sup>1</sup>, S.W. Joshi<sup>2</sup>, and S.L. Rathbun<sup>3</sup>

<sup>1</sup>Center for Statistical Ecology and Environmental Statistics, Department of Statistics  
The Pennsylvania State University, University Park, PA, 16802

<sup>2</sup>Department of Computer Science, Slippery Rock University,  
Slippery Rock, PA 16057

<sup>3</sup>Department of Health Administration, Biostatistics, and Epidemiology,  
University of Georgia, Athens, GA 30602

## Abstract:

Government agencies continue to require meaningful summaries of georeferenced data to support policies and decisions involving risk maps, geographic targets, and resource allocations. So also the public with initiatives of digital governance, enabling transparency, efficiency, and efficacy in the risk assessment and management.

This article briefly introduces hotspot geoinformatics for hotspot detection and prioritization, and provides examples of societal importance involving environmental risk.

*Keywords: Hotspot geoinformatics, digital governance, spatial scan statistic, upper level set scan, poset prioritization, geospatial environmental risk mapping, biodiversity, drinking water system, syndromic surveillance, ecological indicators, environmental risk indicators*

---

\* This material is based upon work supported by (i) the National Science Foundation under Grant No. 0307010, and (ii) the United States Environmental Protection Agency under Grants No. CR-83059301 and No. RD-83244001-0. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

## **1. Introduction and Background**

Quantitative environmental and ecological risk assessment has been a relatively new and rapidly developing area during the past quarter century. Numerous important issues have been discussed in the two recent NAS/NRC reports entitled science and judgment in risk assessment, and issues in risk assessment. On the methodological side, a special issue of the international journal, *Environmental and Ecological Statistics*, is devoted to statistical issues and approaches to some current topics in environmental risk assessment (West and Piegorsch, in press).

We now live in the age of geospatial technologies. The age old geographic issues of societal importance are now thinkable and analyzable. The geography of disease is now just as doable as genetics of disease, for example. And it is possible to pursue it in an intelligent manner with the rapidly advancing geospatial information technology around.

Government agencies continue to require meaningful summaries of georeferenced data to support policies and decisions involving risk maps, geographic targets, and resource allocations. So also the public with initiatives of digital governance, enabling transparency, efficiency, and efficacy in the risk assessment and management.

This article briefly introduces hotspot geoinformatics for hotspot detection and prioritization, and provides examples of societal importance involving environmental risk.

## **2. Ecological Diversity as a Motivating Example**

### **2.1 Quantification for Ecological Diversity**

Conservation biology, landscape ecology, and ecosystem-oriented natural resources management lend considerable urgency to issues and approaches concerning biodiversity assessment. Most of the traditional approaches and statistical tools are plot-based with a goal of definitive characterization. Diversity, however, is relative to a spatial scale, temporal scale, and taxocene spectrum. Patterns may be more informative than absolutes in this regard.

The issues are fundamental in that explaining the effects of environment on the distribution and abundance of species is the essence of much ecological work. The controversies arise from the intrinsic scientific importance of diversity theories, as well as from the broad economic and social ramifications of considering biodiversity in land use decisions. At the heart of the scientific and social controversies regarding diversity are problems of quantification, interpretation, and analysis.

The classical view of diversity remains important for intensive studies of particular ecological communities and forest stands (Gove et al., 1994). However, the emerging sciences of landscape ecology and conservation biology have made evident the logistical and economical impracticality of such intensive observational coverage for regions in the order of square kilometers and larger (Scott et al., 1989). These spatial scales are necessarily encompassed by contemporary ecosystem-oriented resource management and design of regional/national

networks of biodiversity reserves. Furthermore, species/area and minimum viable population issues become fundamental in these matters.

The multidimensional character of diversity can be revealed by establishing an intrinsic, and index-free, diversity ordering. In effect, diversity may appear to have decreased when viewed from one vantage point (i.e., index), and increased when viewed from a different perspective.

In view of the inadequacy of a single index, Patil and Taillie (1979, 1982) quantify diversity by means of diversity profiles. A diversity profile is a curve depicting the simultaneous values of a large collection of diversity indices. Thus, the profile portrays the views of diversity from many different vantage points simultaneously and in a single picture.

Differences in community diversity are studied by comparing profiles. If the two communities are intrinsically comparable, then one profile will lie uniformly above the other. Conversely, when the communities are not intrinsically comparable, their profiles may intersect. But even here, the profiles can reveal which portions of the community have undergone opposing diversity changes.

## **2.2. Indicators for Ecological Diversity**

We proceed to consider ways of coping with complexity and confounding that embrace multiple indicators rather than agonizing over choices and conflicts of diversity measures. We contemplate enlarging the orders of indicators to encompass some interactions in a formal manner that accommodates both parameterization and visualization. We conclude by noting the convergence of biodiversity and ecological community concepts at meter scales, but not for broader landscape, regional, and global scales of ecological organization.

The indefiniteness regarding biodiversity that can give rise to frustration is well expressed by L. R. Taylor (1978) in the following quote:

*Diversity so pervades every aspect of biology that each author may safely interpret the word as he wishes and there is consequently no central theme to the subject. We cannot be sure if this flexibility is healthy or due to lack of discipline, but it can be traced back to the beginnings of interest in biological diversity ...*

The recent programs of the U.S. National Science Foundation probing biocomplexity in many contexts serve to provide evidence that the flexibility addressed by Taylor is both healthy and indicative of need for strengthening discipline with regard to scientific constructs and means by which they are made operational.

Indicators/expressions of this nature are appropriate, and therefore of value, if they convey the desired information within the budget and delivery delays that are acceptable. One method is more efficient than another if it conveys the required information either more rapidly or at less cost. Conveying more information at the same cost and timing is not necessarily desirable if unwanted information has to be processed or filtered.

Increasingly sophisticated management, intervention, remediation, and regulation require a continuing flow of multiple indicators for various aspects of ecosystems.

What ecosystem managers and regulators seek is a complementary set of indicators that captures aspects of interest. We thus have entered the exciting age of indicators. For ecological diversity and biodiversity.

And so also for environmental risk exactly in a similar manner.

### 3. Hotspot Geoinformatics of Detection and Prioritization

In geospatial and spatiotemporal surveillance, it is important to determine whether any variation observed may reasonably be due to chance or not. This can be done using tests for spatial randomness, adjusting for the uneven geographical population density as well as for age and other known risk factors. One such test is the spatial scan statistic, which is used for the detection and evaluation of local clusters or hot-spot areas. This method is now in common use by various governmental health agencies, including the National Institutes of Health, the Centers for Disease Control and Prevention, and the state health departments in New York, Connecticut, Texas, Washington, Maryland, California, and New Jersey. Other test statistics are more global in nature, evaluating whether there is clustering in general throughout the map, without pinpointing the specific location of high or low incidence or mortality areas.

#### 3.1 Scan Statistic Approach for Geospatial Hotspot Detection

Three central problems arise in geographical surveillance for a spatially distributed response variable indicator, generating a cell-wise constant cellular surface (see Figure 1). These are (i) identification of areas having exceptionally high (or low) response, (ii) determination of whether the elevated response can be attributed to chance variation (false alarm) or is statistically significant, and (iii) assessment of explanatory factors that may account for the elevated response. The circle-based spatial scan statistic (Kulldorff & Nagarwalla, 1995; Kulldorff, 1997) has become a popular method for detection and evaluation of disease clusters. In space-time, the scan statistic can provide early warning of disease outbreaks and can monitor their spatial spread. The response variable/indicator in this case is disease rate in the case control situation for binomial distribution and in the case count situation for Poisson distribution.

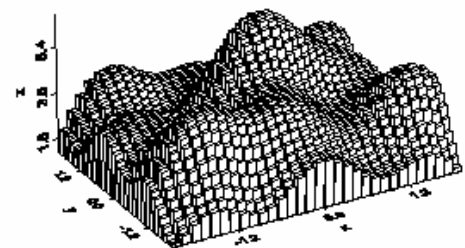


Figure 1: Cellular Surface

#### ULS Scan Statistic.

A new version of the spatial scan statistic is designed for detection of hotspots of arbitrary shapes and for data defined either on a tessellation or a network. This version looks for hotspots from among all connected components of upper level sets of the response rate and is therefore called the upper level set (ULS) scan statistic (Patil & Taillie, 2004a). The method is adaptive with respect to hotspot shape since candidate hotspots have their shapes determined by the data rather than by some *a priori* prescription like circles or ellipses.

## Continuous Response Situations.

The circle-based scan statistic methodology can be extended to include continuous response distributions, such as, three parametric families of distributions: gamma distribution, lognormal distribution, and scaled beta distribution. The first two families apply to responses that can range from zero to infinity, while the third is for bounded responses. The overall approach is to model the mean and relative variance in terms of the size variable. These moments are functions of the parameters of the response distribution, so that a likelihood function can be written down and parameters estimated by maximum likelihood.

Our strategy for handling continuous responses is thus to model the mean and variance of each cellular response distribution in terms of the size variable  $Aa$  for cell  $a$ ; modeling is guided by the principle that the mean response should be proportional to  $Aa$  and the relative variability should decrease with  $Aa$ . Just as with the Poisson and binomial models, we take the  $Y_a$  to be independent. The approach is best illustrated for the gamma family of distributions, as follows:

We parameterize the gamma distribution by  $(k, \beta)$ , where  $k$  is the index parameter and  $\beta$  is the scale parameter. Thus, if  $Y$  is a gammadistributed variate,  $E[Y] = k\beta$  and  $\text{Var}[Y] = k\beta^2$ .

Both  $k$  and  $\beta$  can vary from cell to cell but additivity with respect to the index parameter suggests that we take  $k$  proportional to the size variable:

$$ka = Aa/c,$$

where  $c$  is an unknown parameter but whose value is the same for all  $a$ . This gives the following mean and squared coefficient of variation:

$$E[Y_a] = \beta Aa/c \text{ and } \text{CV}^2[Y_a] = c/Aa.$$

## Multivariate Hotspotting

When we have multiple indicators representing multidimensional landscape level cellular environmental risk, multivariate hotspotting has been conceptualized. See Modarres and Patil (2008, in press).

### 3.2 Hotspot Prioritization with Multi-criteria Indicators

At times, several hotspots are discovered and in response to several stakeholders resulting in their criteria, scores, and/or indicators, the hotspots need to be prioritized and ranked without crunching the indicators into an index. This gives rise to a data matrix of rows for hotspots and columns for indicator scores. And the problem becomes a partial order theory problem and has been addressed in Patil and Taillie, 2004b. Broadly speaking, this paper is concerned with the question of ranking a finite collection of objects when a suite of indicator values is available for each member of the collection. The objects can be represented as a cloud of points in indicator space, but the different indicators (coordinate axes) typically convey different comparative messages and there is no unique way to rank the objects while taking all indicators into account. A conventional solution is to assign a composite numerical score to each object by combining the indicator information in some fashion. Consciously or otherwise, every such composite

involves judgments (often arbitrary or controversial) about tradeoffs or substitutability among indicators.

Rather than trying to combine indicators, we take the view that the relative positions in indicator space determine only a partial ordering and that a given pair of objects may not be inherently comparable. Working with Hasse diagrams of the partial order, we study the collection of all rankings that are compatible with the partial order (linear extensions). See Figure 2. In this way, an interval of possible ranks is assigned to each object. The intervals can be very wide, however. Noting that ranks near the ends of each interval are usually infrequent under linear extensions, a probability distribution is obtained over the interval of possible ranks. This distribution, called the rank-frequency distribution, turns out to be unimodal (in fact, log-concave) and represents the degree of ambiguity involved in attempting to assign a rank to the corresponding object. See Table 1.

Stochastic ordering of probability distributions imposes a partial order on the collection of rank-frequency distributions. This collection of distributions is in one-to-one correspondence with the original collection of objects and the induced ordering on these objects is called the cumulative rank-frequency (CRF) ordering; it extends the original partial order. See Figure 3. Although the CRF ordering need not be linear, it can be iterated to yield a fixed point of the CRF operator. We hypothesize that the fixed points of the CRF operator are exactly the linear orderings. The CRF operator treats each linear extension as an equal “voter” in determining the CRF ranking. It is possible to generalize to a weighted CRF operator by giving linear extensions differential weights either on mathematical grounds (e.g., number of jumps) or empirical grounds (e.g., indicator concordance). Explicit enumeration of all possible linear extensions is computationally impractical unless the number of objects is quite small. In such cases, the rank-frequencies can be estimated using discrete Markov chain Monte Carlo (MCMC) methods.

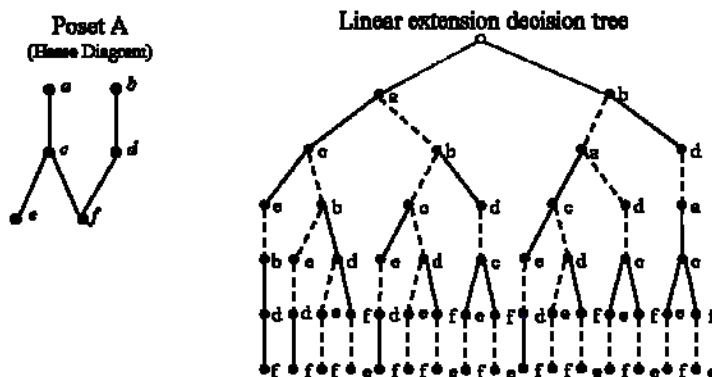


Figure 2. Hasse diagram of Poset A (*left*) and a decision tree enumerating all possible linear extensions of the poset (*right*). Every downward path through the decision tree determines a linear extension. Dashed links in the decision tree are not implied by the partial order and are called *jumps*. If one tried to trace the linear extension in the original Hasse diagram, a “jump” would be required at each dashed link. Note that there is a pure-jump linear extension (path *a, b, c, d, e, f*) in which every link is a jump.

Table 1. Rank-frequency table for the poset of Figure 2. Each row gives the rank-frequency distribution for the corresponding element of the poset.

	Rank						
Element	1	2	3	4	5	6	Totals
<i>a</i>	9	5	2	0	0	0	16
<i>b</i>	7	5	3	1	0	0	16
<i>c</i>	0	4	6	6	0	0	16
<i>d</i>	0	2	4	6	4	0	16
<i>e</i>	0	0	1	3	6	6	16
<i>f</i>	0	0	0	0	6	10	16
Totals	16	16	16	16	16	16	

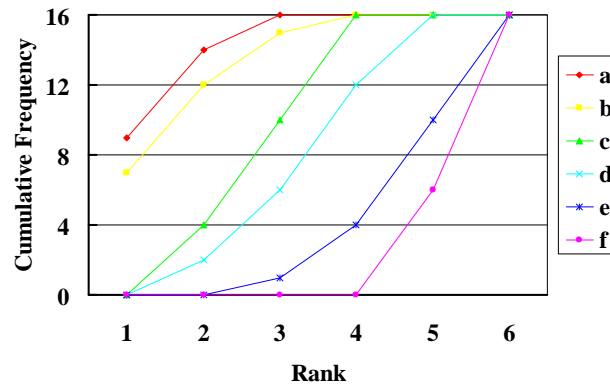


Figure 3. Cumulative rank-frequency distributions for Poset A.

## 4. Some Examples

### 4.1 Drinking Water Quality and Water Utility Vulnerability

New York City has installed 892 drinking water sampling stations across the five boroughs. Each 4.5-foot high station is located outdoors and draws water from a nearby water main. The purpose is to monitor general water quality, detect potential health threats, and thwart bioterror activity. Sampling frequency was increased after the 9/11 attacks and, currently, about 47,000 water samples are analyzed annually. Parameters analyzed include: bacteria, chlorine, pH, inorganic and organic pollutants, color, turbidity, odor, and many others. The network version of the ULS scan statistic can provide a real-time surveillance system for detecting and evaluating water quality hotspots within the distribution system on a parameter-by-



parameter basis.

An overall assessment of water quality at each sampling station taking all parameters into account is achieved by employing recent progress on multi-criterion ranking using poset (partially ordered set) prioritization (Patil & Taillie, 2004b).

#### 4.2 New York City Subway System Syndromic Surveillance

For certain problems, there is an underlying network structure on which we will want to perform the cluster detection and evaluation. For example, the New York City Health Department is monitoring the New York subway system and water distribution networks for bioterrorism attacks. In such a scenario, a circular scan statistic is not useful as two individuals close to each other in Euclidian distance may be very far from each other along the network. However, the ULS methods will be employed for the detection and evaluation of clusters on a predefined network. The essentially linear structure of these networks, compared with tessellation-derived networks, is expected to have a major impact on the form of the null distributions and their parametric approximations.

The New York City Department of Health (DOH) and Metropolitan Transportation Authority (MTA) began monitoring subway worker absenteeism in October 2001 as one of several surveillance systems for the early detection of disease outbreaks. Each day the MTA transmits an electronic line list of workers absent the previous day, including work location and reason for absence. DOH epidemiologists currently monitor temporal trends in absences in key syndrome categories (e.g., fever-flu or gastrointestinal illness). Analytic techniques are needed for detecting hotspots within the subway network.

#### 4.3 Choice of Indicator Class for Watershed Biological Impairment

There has been considerable work on determining a suitable method to accomplish a satisfactory ordering of a group of objects, when there are multiple evaluation criteria. Data from 21 watersheds of the Atlantic Slope Consortium (ASC) has been examined with the goal of determining an accurate ranking by overall watershed condition. In particular, there are three levels of indicators that range from Level I to Level III, increasing in the quality and accuracy of the data as well as the cost and effort needed to obtain the data. Due to the high cost, level III data is available only for six watersheds; however the interest is in ranking all 21 watersheds (Patil et al., 2007). We use elements of Poset (Partial order set) theory as a foundation for our analysis. The Poset linear extension method can be used to find rankings without using an index, relying only on pairwise comparisons of the objects.

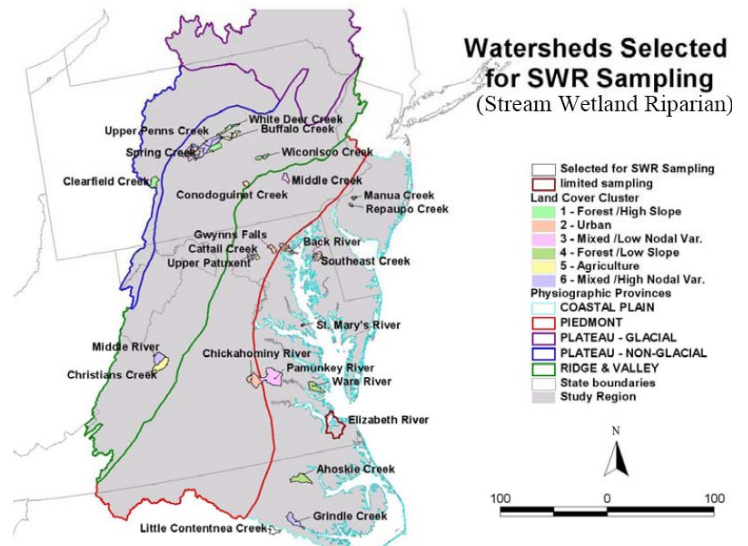


Figure 4. The 21 watersheds from the Mid-Atlantic area

The data relates to a set of 21 watersheds from the Atlantic Slope Consortium (ASC), which has the goal for determining an accurate ranking of the health of the watersheds (Brooks et al., 2006). There are 15 indicators each of which measures a facet of the watershed and are grouped into Level I to Level III, increasing in quality and accuracy as well as amount of cost and effort. For all 21 watersheds, we have data for seven Level II indicators, and for five Level I indicators. The data matrix for Level II indicators has 21 rows and 7 columns, and the data matrix for the Level I indicators has 21 rows and 5 columns. For Level I and II, the data was scored between zero and one. The three levels are as follows:

Level III – Intensive Field Assessment

Level II – Rapid Field Assessment

Level I – Landscape Assessment from GIS

The indicators are categorized by the level of accessibility and availability. The Level III Intensive Field Assessment. It is the most expensive among the three levels. We consider the indicators of Level III, the best quality of data we have regarding the watershed. Due to the money and effort in the procedure of obtaining this data, it is available for only six watersheds.

The three level III indicators are collected from the Index of Biological Integrity (IBI) developed by EPA. They are benthic IBI, fish IBI, and  $\text{NO}_3$ . The two IBI indicators are biological indicators, while the  $\text{NO}_3$  is considered a chemical indicator.

The Level II Rapid Field Assessment is obtained from onsite sampling. Certain level of expertise is involved in the field assessment. Generally, level II data is relatively cheap compared to the level III data. The Level I Landscape Assessment is the satellite data, and is the easiest to access and the least expensive. The poset prioritization and ranking analysis of the data recommends use of Level II or even Level I indicators to estimate the true ranking of the 21 watersheds by their biological condition.

#### **4.4 More Examples and Applications**

Geospatial and spatiotemporal environmental risk mapping commands vast literature. Some publications concentrate on geospatial cellular modeling of environmental risk, the probability of adverse event, and thus providing a cellular surface. Some provide discrete or continuous profiles, allowing extraction of indicators, whereas still others, attempt to identify and provide surrogate indicators for the environmental risk. Applications abound. See, for example, publications and/or websites for:

- (a) Regional Vulnerability Assessment Indicators of the USEPA REVA program,
- (b) Indicators of Ecological Value, Ecological Sensitivity, and Anthropogenic Pressure of the Map of Italian Nature Program,
- (c) Tsunami Inundation Risk Mapping and Management Program of NOAA,
- (d) Environmental Risks and Surrogate Indicators around the World for Forest Fires, Pests, Insects and Diseases, Infectious Disease Vectors, Floods, Landslides, Droughts, and others.
- (e) The Next Generation of Ecological Indicators of Wetland Condition, *EcoHealth*, the Journal.
- (f) *Ecological Indicators*, the Journal.
- (g) For several environmental and ecological landscape level indicators and related analyses, see Johnson and Patil (2006) and Myers and Patil (2006).

## 5. Future Directions

Surveillance geoinformatics of hotspot detection and prioritization for environmental risk is a critical need of the 21<sup>st</sup> century. Next generation decision support system within this context is crucial. It will be productive to build on the present effort in the directions of prototype and user-friendly methods, tools, and software, and also in the directions of thematic groups, working groups, and case studies important at various scales and levels. The authors have such a continuation effort in progress within the context of digital governance with NSF support and would welcome interested readers to join this collaborative initiative.

## References

Brooks, R., McKinney-Easterling, M., Brinson, M., Rheinhardt, R., Havens, K., O'Brien, D., Bishop, J., Rubbo, J., Armstrong, B., & Hite, J. (2006). A Stream-Wetland-Riparian (SWR) Index for assessing condition of aquatic ecosystems in small watersheds along the Atlantic Slope of the eastern U.S. (Manuscript).

Gove, J., Patil, G.P., & Taillie, C. (1994). A mathematical programming model for maintaining structural diversity in uneven-aged forest stands with implications to other formulations. *Ecological Modelling*, 79, 11-19.

Johnson, G.D. & Patil, G.P. (2006). *Environmental and Ecological Statistics Series: Volume 1: Landscape Pattern Analysis for Assessing Ecosystem Condition*. New York, NY: Springer.

Kulldorff, M. & Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14, 799–810.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26, 1481–1496.

Modarres, R. & Patil, G.P. (2008, in press). Hotspot detection with bivariate data. *Journal of Statistical Planning and Inference*.

Myers, W. & Patil, G.P. (2006). *Environmental and Ecological Statistics Series: Volume 2: Pattern-based Compression of Multi-band Image Data for Landscape Analysis*. New York, NY: Springer.

Patil, G.P., Bhat, K.S., McKenney-Easterling, M., & Kase, M. (2007, in press). A Toolbox for the Choice of Indicator Classes for Ranking of Watersheds. *Environmental and Ecological Statistics*.

Patil, G.P. & Taillie, C. (1979). A study of diversity profiles and orderings for a bird community in the vicinity of Colstrip, Montana. In *Contemporary Quantitative Ecology and Related Ecometrics*, G.P. Patil and M. Rosenzweig (eds), pp. 23-48. Fairland, Maryland: International Cooperative Publishing House.

Patil, G.P. & Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77, 548-567. (Invited discussion paper).

Patil, G.P. & Taillie, C. (2004a). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11, 183-197.

Patil, G.P. & Taillie, C. (2004b). Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization. *Environmental and Ecological Statistics*, 11, 199-228.

Scott, J., Csuti, B., Estes, J., & Anderson, H. (1989). Status assessment of biodiversity protection. *Conservation Biology*, 3, 85-87.

Taylor, L. R. (1978). Bates, Williams, Hutchinson – a variety of diversities. In *Diversity of Insect Faunas*. I. A. Mound and N. Waloff (eds), pp. 1-18. Oxford: Blackwell Scientific Publications.

West, W. & Piegorsch, W. (2009, in press). Modern benchmark analysis for environmental risk assessment. *Environmental and Ecological Statistics*.