PENNSTATE

Center for Statistical Ecology and Environmental Statistics

DETECTION AND DELINEATION OF CRITICAL AREAS USING ECHELONS AND SPATIAL SCAN STATISTICS WITH SYNOPTIC CELLULAR DATA

G. P. Patil[1], J. Bishop[2], W. L. Myers[3], C. Taillie[1], R. Vraney[1], and Denice Wardrop[2]

[1]Center for Statistical Ecology and Environmental Statistics,
[2]Cooperartive Wetlands Center, [3]Office of Remote Sensing for Earth Resources

The Pennsylvania State University, University Park, PA 16802

EPA Project Officer: Barbara Levinson

Prepared with partial support from U.S. EPA Star Grant for Atlantic Slope Consortium, Cooperative Agreement Number R-82868401. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agency and no official endorsement should be inferred.

Based on invited paper for the International Society for Ecosytem Health, Washington DC, June 6-11, 2002.
Special Thematic Conference on Healthy Ecosystems/Healthy People:
Linkages Between Biodiversity, Ecosystem Health and Human Health

To appear in *Environmental and Ecological Statistics, 11(1), 2004:*
Special Institutional Thematic Issue on Center for Statistical Ecology and Environmental Statistics

Technical Report Number 2002- 0501
TECHNICAL REPORTS AND REPRINTS SERIES
May 2002

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
http://www.stat.psu.edu/~gpp

DETECTION AND DELINEATION OF CRITICAL AREAS USING ECHELONS AND SPATIAL SCAN STATISTICS WITH SYNOPTIC CELLULAR DATA

G. P. Patil[1], J. Bishop[2], W. L. Myers[3], C. Taillie[1], R. Vraney[1], and Denice Wardrop[2]

[1]Center for Statistical Ecology and Environmental Statistics
Department of Statistics
[2]Cooperartive Wetlands Center
[3]Office of Remote Sensing for Earth Resources

The Pennsylvania State University
University Park, PA 16802

Abstract:

Geographical surveillance for hotspot detection and delineation has become an important area of investigation both in geospatial ecosystem health and in geospatial public health. In order to find critical areas based on synoptic cellular data, geospatial ecosystem health investigations apply recently discovered echelon tools. In order to find elevated rate areas based on synoptic cellular data, geospatial public health investigations apply recently discovered spatial scan statistic tools. The purpose of this paper is to conceptualize a joint role for these together in the spirit of a cross-disciplinary cross-fertilization to accomplish more effective and efficient geographical surveillance for hotspot detection and delineation, and early warning system.

Keywords: Atlantic slope watersheds, critical areas, early warning system, echelon analysis, elevated rate area identification, hotspot detection and delineation, landscape analysis, spatial scan statistics.

1. Introduction

Geographical surveillance for hotspot detection and delineation has become an important area of investigation both in geospatial ecosystem health and in geospatial public health. In order to find critical areas based on synoptic cellular data, geospatial ecosystem health investigations apply recently discovered echelon tools. In order to find elevated rate areas based on synoptic cellular data, geospatial public health investigations apply recently discovered spatial scan statistic tools. The purpose of this paper is to conceptualize a joint role for these together in the spirit of a cross-disciplinary cross-fertilization to accomplish more effective and efficient geographical surveillance for hotspot detection and delineation, and early warning system.

The spatial scan and echelon methods are applicable, and can be valuable, in various areas of concern, such as: (1) disease epidemiology with prevalence, incidence, or mortality data, (2) medical imaging involving brain scan or mammography for the breast, (3) reconnaissance involving mine fields of antipersonnel mines or deposits of mineral, oil, or uranium, (4) astronomy involving stars or galaxies, (5) archaeology and history involving geography of pottery or odontological feature, (6) biodiversity issues involving species-rich and species-poor areas, (7) urban and regional planning involving distribution of population and businesses, (8) regional studies involving water resources at watershed scales, (9) power lines and electromagnetic effects on objects around, and (10) networks of water

distribution systems, subway systems, and road transportation systems. The thrust of this paper is focused on biodiversity, ecosystem health and human health.

2. Echelon Analysis for Critical Areas

Quantitative spatial data are important data sources of many environmental process models to determine future implications of current resource use, policies, and interventions. End products of applying such models are often mappings of indices for level of potential environmental impact, which then become guides to allocation of economic and technical resources for amelioration.

Errors in quantitative spatial data layers will propagate through environmental models and find expression in the resulting indexes of environmental impact. However, the consequences of such errors for decision-making may well depend upon where the errors occur. There may be relatively little confusion introduced by moderate errors occurring in a vicinity that otherwise has consistently high values of a variable. In contrast, errors compound confusion in areas that are highly variable. Errors can also substantially distort the apparent state of areas that otherwise have consistently low values of a variable.

It is therefore desirable to have a systematic means of determining spatial organization in mappings of quantitative variables, both for quantitative input variables to environmental models and for indexes of potential impact generated by the models. Contemporary computer capabilities for visualization of surfaces are helpful in this regard, but their interpretation is substantially subjective. Echelons present an innovative alternative for objectively determining quantitative spatial structure for direct mapping either with or without computer-assisted visualization (Johnson et al.,1998; Kurihara et al., 2000; Myers et al., 1997,1995,1999; Patil and Taillie, 1999; Ramakomud, 1998; Rodriguez-Iturbe and Rinaldo, 1997). Thus, they can facilitate analysis of implications of errors associated with environmental models that take quantitative layers as input, or produce quantitative output layers, or both.

The tessellated cellular surfaces arise, also when one records values on each cell, even if without any modeling or measurement error. The records may consist of counts or rates of events of interest.

Echelons of Spatial Variation

The spatial variables for echelon analysis can be considered as topographies, whether real or virtual. Such terrain information is typically formatted for processing in a geographic information system (GIS) as a digital elevation model (DEM). This comprises a raster in which an `elevation' value is specified for the center of each cell. Echelons divide the (virtual) terrain into structural entities consisting of peaks, foundations of peaks, foundations of foundations, and so on in an organizational recursion. Saddles determine the divisions between entities. Each entity is assigned an echelon number for identification purposes. The peaks constitute one series of structural entities, being numbered in decreasing order of summit elevation. The foundations constitute a second series of entities that are likewise numbered in order of decreasing top level, starting with the next number after that assigned to the lowest peak. Consider, for example, the following terrain depicted in profile with division as seen in Figure 1.

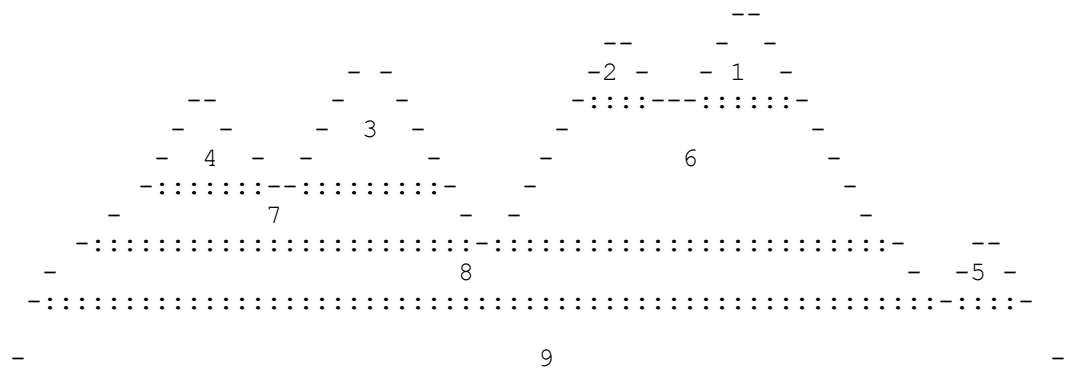


Figure 1.

The numbered entities thus determined are called echelons. Echelons are determined directly by organizational complexity in the spatial variable, and not by either absolute `elevation' or steepness.

Echelons form extended families of terrain entities having a genealogy similar to that of an extended human family, except that each echelon has only one parent. In the case of echelons, an entity that rises from another is more aptly termed an`ascendant' than a `descendant'. Likewise, a `parent' entity is termed a `founder'. The echelon relations determine a family tree as illustrated in Figure 2.

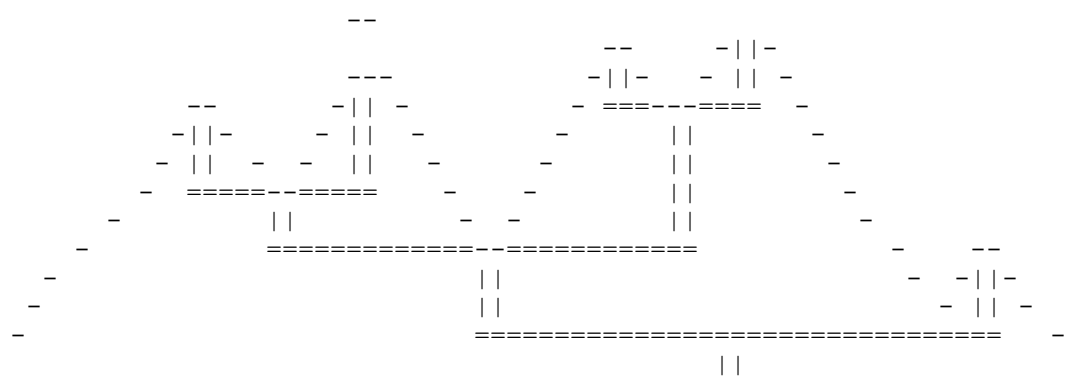


Figure 2.

This is a 'scaled tree' in the sense that the height of each vertical edge corresponds to the height of the echelon above its founder (parent). The cumulated height above the root is the height of the terrain. The number of 'ancestors' for an echelon is a local measure of regional complexity.

The echelons also comprise a structural hierarchy of organizational orders. The orders of the hierarchy are assigned and numbered in the same manner as for a network of streams and tributaries (Rodriguez-Iturbe and Rinaldo, 1997). Thus, peaks are akin to unbranched tributaries, and have order 1. A foundation for two or more order 1 entities is of order 2. Likewise, a foundation for two or more order 2 entities is of order 3. A low order entity (see 5 in Figure 1) sharing a foundation with a higher order entity does not increase the order of the foundation. This is like the case of an unbranched tributary entering a higher order stream.

Echelon Characteristics

A suite of form attributes can be determined for each echelon, including area extent of the basal slice and vertical projection above its founder. Some form attributes may depend upon an interval scale of measure for the vertical dimension, but the echelon decomposition only requires an ordinal scale of measurement. A standard table of echelon characteristics contains a record (row) with ten fields for each echelon, including echelon ID number, order, founder, maximum level, minimum level, relief, cells, progeny, ancestors, and setting within the tree. The table is associated with an echelon map file giving the 'level' value and echelon ID number for each cell. Echelons thus formalize the structural complexity of the (surface) variable without incurring any loss of information with respect to (surface) level.

Species Rich Areas for Biodiversity Monitoring and Assessment

As an important component of biodiversity monitoring, species richness is a response variable of considerable interest in conservation biology. Monitoring over a large spatial extent requires an objective method for characterizing the spatial distribution for species richness. Objectivity is needed for delineation of areas that are relatively species rich or poor, and to also provide a way to compare other spatial distributions, as with temporal monitoring for change detection.

Johnson et al. (1998) apply the echelon method to the characterization of breeding bird species richness across the state of Pennsylvania, using USEPA EMAP hexagons of 635 km². The statewide echelon map is presented in Figure 2a.

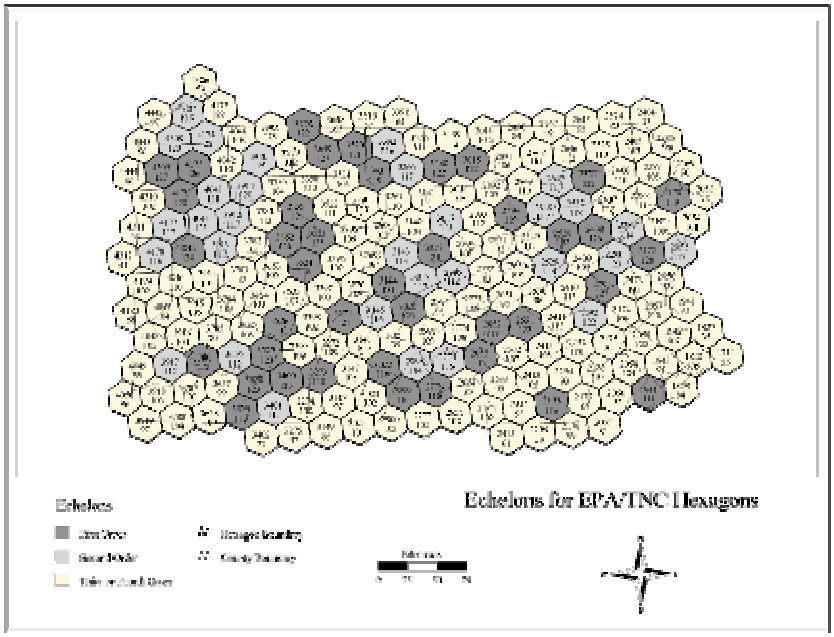


Figure 2a. Statewide echelon map based on EMAP hexagons. The 4-digit number in each hexagon is the EPA-EMAP identifier, while the number below is species richness.

Regions of high species richness relative to their surroundings in a statewide perspective are delineated by complexes of the first order and second order echelons. General correspondence can be seen with the thematic map in Figure 2b.

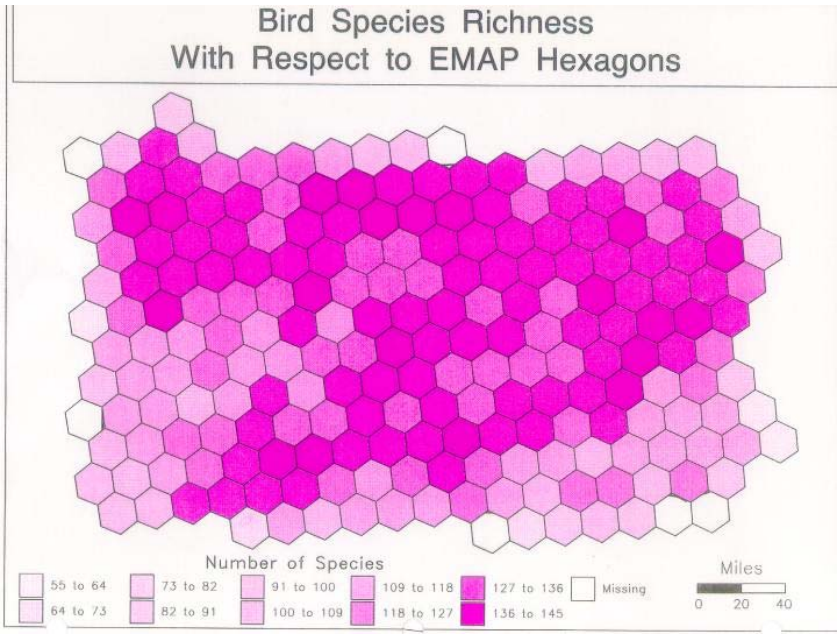


Figure 2b. Bird richness in the hexagons.

Note that the boundaries of regionally high areas in Figure 2b are not clearly defined, whereas the echelon objects of Figure 2a serve to objectively define such boundaries.

Johnson et al. (1998) further analyze selected species rich areas at the hexagon 635 km² scale at a much finer scale of the breeding bird atlas blocks of 24 km² scale to help with biodiversity assessment in terms of explanatory variables at a finer scale responsible for the species richness in species rich areas at a higher coarser scale. See Figure 2cde.

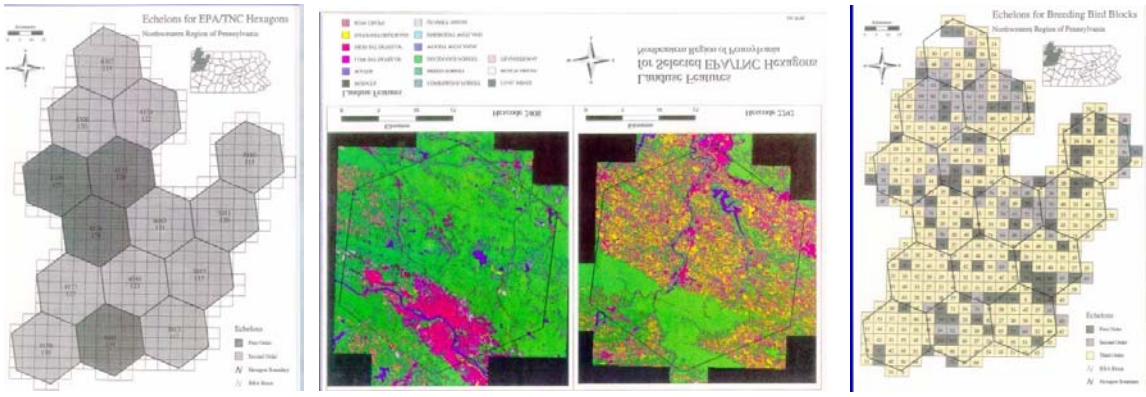


Figure 2cde

3. Echelon Trees, Principal Families, and Upper Echelons

Since most echelon trees are much too complicated for visual study as dendrograms, characterization and comparison of echelon trees is done through analytical processes such as pruning. (Myers and Patil 2002). The current stage of development for echelon information technology provides a mature descriptive capability for characterizing quantitative spatial variables. A major question concerning quantitative spatial variables with respect to many applications is whether there are substantial sectors of the surface having particularly high or particularly low values relative to the mean level. These are the 'uplands' and 'lowlands' of the virtual surface. Currently the manager or investigator is obliged to resort to subjective examination of visualizations on maps and/or computer displays in an attempt to gain such insights regarding what should be 'focal' areas.

In the domain of echelons, candidate focal areas may be conceptualized as principal families and the sectors that they occupy can be considered as being principalities. The information needed for determining principal families resides in the echelon table and tree representation. Once the principal families are identified, the sectors that they occupy can be extracted by exploiting the linkage between the echelon map and echelon table. Analytical and computational strategies need to be formulated for segregating the principal families from what typically are, hundreds of upper-level echelon families.

Probabilities based on a null model using a planar random process could allow the user to specify a criterion for areas of potential concern to be extracted computationally. In other words, an echelon family would be seen as a candidate for focus if the probability of its extent receiving observed amounts is less than the criterion under a random distribution of quantity over area.

Since echelon determination is computationally intensive, there would be further advantage in capability to extract principal families from partially determined echelons. This scenario would terminate the top-down progression of echelon determination for an area when the probability of observing encountered values under planar randomization exceeds the criterion level. The echelon table would then consist of a series of subtrees, with a subtree for each principal family.

Echelons synthesize the topological structure of synoptically mapped environmental indicators in a regional context for comparative purposes and for objective analyses of complex hierarchies of spatial variation across landscapes. The environmental indicator is regarded as a surface variable that represents a virtual topography as depicted (in one-dimensional profile) in Figure 3. Echelons are structural entities consisting of peaks, foundations of peaks, foundations of foundations, and so on in an organizational hierarchy. It is natural to display the echelon hierarchy as a tree and, in this form, echelons have proven effective for elucidating patterns of concentration and connectivity for biodiversity, landscape change, urban sprawl, etc. (Myers *et al*., 1995, 1997; Myers *et al*., 1999; Kurihara *et al*., 2000; Smits and Myers, 2000; Myers and Patil, 2002).

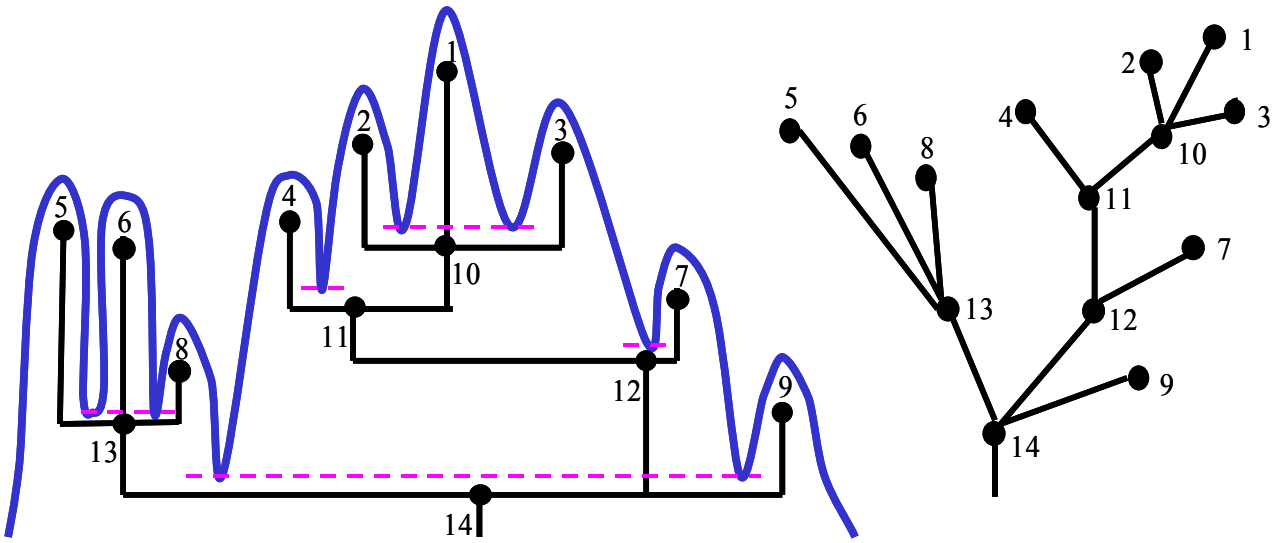


Figure 3. Echelon decomposition of a surface (left) and associated echelon tree (right).

Upper Echelons

The following computational strategy for upper echelons uses these informational properties:

- The echelon table records the extent and top height of each echelon.
- Any path through the tree of foundation relations also orders the echelons by top height.

We first want to suppress the echelons that are entirely below the mean level for the population of cells. To do so, we progressively suppress echelons in order of increasing top height. As the suppression proceeds, we track the cumulative number of cells suppressed and stop when the next echelon would cumulatively encompass more than half the cells. This has the effect of pruning the tree from the bottom, leaving a series of subtrees as prospective upper echelons.

We suppress all remaining cells having height less than or equal to the greatest top height for a suppressed echelon. We will have thus suppressed topologically (objectively) determined lower levels of the surface.

We assign the remaining cells identification numbers that indicate membership in subtrees. We then cumulate the surface levels over cells within in each subtree, and reassign subtree sequence numbers according to decreasing cumulatives. We retain up to 250 of the highest ranking subtrees as principal families that can be mapped as a one-byte binary image.

We prepare a color lookup table for rendering the principal families in which red intensity expresses top height, green intensity expresses area extent of family, and blue expresses complexity of the family according to number of family members.

The image of principal families is then presented to the user via GIS for interactive query and investigation.

4. Spatial Scan Statistic Analysis for Elevated Rate Areas

The spatial scan statistic was developed as a surveillance tool to detect and test the significance of local disease clusters, without making prior assumptions about their location or size, and adjusting the inference for the multiple testing inherent in the many potential sizes and locations considered. In mathematical terminology it is defined in terms of using a scanning window of any shape (Kulldorff, 1997), but it has so far been implemented using a circular scanning window.

The method imposes a circular window on the map and lets its center move over the area so that at any given position, the window includes different sets of neighboring counties. If the window contains the centroid of a census area, then that whole census area is included in the window. For practical reasons, the center of the window is positioned only at the census tract centroids, and at each position, the radius of the circular window is varied continuously from zero up to a maximum radius so that the window never includes more than 50 percent of the total population. In this way, the circular window is flexible both in location and size. In total, the method creates a very large number of distinct circular windows, each with a different set of neighboring counties within and each a possible candidate for containing a cluster. For each window, the method tests the null hypothesis against the alternative hypothesis that there is an elevated risk of the event within the window compared to outside.

For each circle, the observed and expected number of cases inside and outside the circle is noted, and the corresponding likelihood is calculated. This likelihood is then maximized over all circles, identifying the window that constitutes the most likely disease cluster. The likelihood ratio for this window is noted and constitutes the maximum likelihood ratio test statistic. Adjusting for the multiple testing of multiple cluster locations and sizes, its distribution under the null-hypothesis and its corresponding *p*-value are obtained by repeating the same analytic exercise for 9999 random replications of the data set generated under the null hypothesis, in a Monte Carlo simulation. Calculations can be performed using the SaTScan software (SaTScan) developed at the National Cancer Institute specifically to implement the spatial scan statistic.

The spatial scan statistic has the following features, which make it particularly suitable for geographical surveillance: (1) it adjusts both for the inhomogeneous population density and for any number of confounding variables; (2) by searching for clusters without specifying their size or location the method ameliorates the problem of preselection bias; (3) the likelihood ratio based test statistic takes multiple testing into account, and delivers a single p-value for the test of the null hypothesis; and (4) if the null hypothesis is rejected, we can specify the approximate location of the cluster that caused the rejection.

The spatial scan statistic deals with the following situation. A region R of Euclidian space is tessellated or subdivided into cells that will be labeled by the symbol a. Data is available in the form of a count Y_a (non-negative integer) on each cell a. In addition, a "size" value A_a is associated with each cell a. The cell sizes A_a are regarded as known and fixed, while the cell counts Y_a are independent random variables. Two distributional settings are commonly studied:

1) **Binomial**

$A_a = N_a$ is a positive integer and $Y_a \sim$ Binomial (N_a, p_a), where p_a is an unknown parameter attached to cell a with $0 < p_a < 1$.

2) **Poisson**

A_a is a positive real number and $Y_a \sim$ Poisson ($\lambda_a A_a$), where $\lambda_a > 0$ is an unknown parameter attached to cell a .

Each distributional model has a simple interpretation. For the binomial, a collection of N_a items resides in cell a and each item has a certain trait independently with probability p_a. The cell count Y_a is the number of items having the trait. For the Poisson, A_a is the size (perhaps hypervolume) of the cell *a*, and Y_a is the realized number of points in a Poisson process of intensity λ_a across the cell.

The spatial scan statistic seeks to identify "hotspots", i.e., "clusters" of cells that have an elevated response compared with the rest of the region. However, by elevated response we do not mean large values for the raw counts Y_a; instead, we mean large values for the rates,

$$G_a = Y_a / A_a .$$

In other words, the cell counts are adjusted for cell sizes before comparing cell responses. The scan statistic easily accommodates other rate adjustments, such as for age or for gender. In some situations, one also wishes to find regions of low response; this is easily accomplished by reversing the direction of the rate function.

A collection of cells from the tessellation should satisfy several geometrical properties before it could be considered as a candidate for a hotspot cluster. First, the union of the cells should comprise a connected subset of the region *R*. All such collections of cells will be referred to as zones and the set of all zones is denoted by Ω. Thus, a zone $Z \in \Omega$ is any collection of cells that are connected. Second, the zone should not be excessively large. Otherwise, the zone instead of its exterior would constitute background. This restriction is generally achieved by limiting the search for hotspots to zones that do not comprise more than, say, fifty percent of the population.

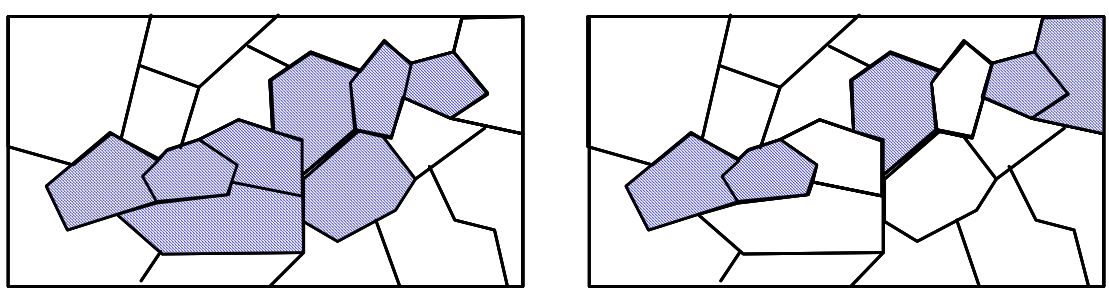


Figure 4. A tessellated region. The collection of shaded cells in the left-hand diagram is connected and, therefore, constitutes a zone in Ω. The collection on the right is not connected.

The following hypothesis testing scenario formalizes the cluster detection and delineation problem. For definiteness, we use the binomial distributional setting.

H_0 : p_a is the same for all cells in region R, i.e., there is no cluster.

H_1 : There is a non-empty zone Z (connected union of cells) and parameter values $0 < p_0, p_1 < 1$ such that

$$p_a = \begin{cases} p_1 & \text{for all cells } a \text{ in Z} \\ p_0 & \text{for all cells } a \text{ in } R - Z \end{cases} \qquad \text{and} \qquad p_1 > p_0 .$$

The zone Z specified in H_1 is an unknown parameter of the model. The full model, $\mathrm{H}_0 \bigcup \mathrm{H}_1$, involves three unknown parameters:

$$Z, p_0, p_1 \quad \text{with} \quad Z \in \Omega \ \text{ and } \ p_0 \le p_1 .$$

The null model, H_0, is the limit as $p_1 \to p_0$; however, the parameter Z is not identifiable in the limit. For given Z, the likelihood estimates of p_0 and p_1 can be written down explicitly which allows us to determine the profile likelihood for Z:

$$L(Z) = \max_{p_0, p_1} L(Z, p_0, p_1) = L(Z, \hat{p}_0, \hat{p}_1) .$$

The difficult part of hotspot estimation lies in maximizing $L(Z)$ as Z varies over the collection Ω of all possible zones.

The SaTscan performs a conditional likelihood ratio test, conditional on the sum $\mathrm{Y} = \sum \mathrm{Y}_a$ with N_a and $N = \sum N_a$ known and fixed. The simulation performs the independent identical realizations of the data set $\mathrm{Y} = \sum \mathrm{Y}_a$ given. This turns out to be the classical multivariate hypergeometric in the Binomial Model and multinomial in the Poisson Model. It is interesting to observe here the role of sum-symmetric distributions as defined in Patil and Ratnaparkhi (1988), Joshi and Patil (1970), and Janardan and Patil (1972).

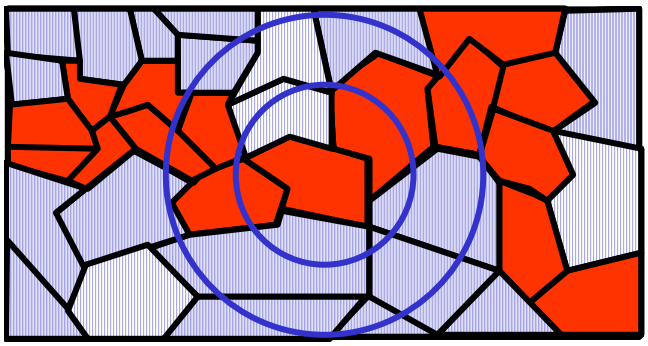
A primary advantage of spatial scan statistics is that the approach is designed specifically to detect clusters and test their significance. The cluster sizes and regions do not have to be specified in advance. The hypotheses are clearly defined, and the test statistic is based on a likelihood ratio, and not on an ad hoc procedure.

5. Comparative Features of Echelons and Satscans

Contemporary study of human disease as a component of ecosystem health entails the spatial scan statistic (Kulldorff and Nagarwalla, 1995) for detecting geographic clusters of disease and other responses that are significantly elevated with respect to the regional setting. In conjunction with the circular spatial scan statistic, echelon analysis can more clearly delineate the cluster boundaries for focus of investigation. A primary advantage of the spatial scan statistic is that the approach is designed specifically to detect clusters and test their significance. The cluster sizes and regions do not have to be specified in advance. The hypotheses are defined precisely, and the test statistic is based on a likelihood ratio and not on some ad hoc index.

The spatial extent of a disease cluster can reflect physical and man-made features of the environment such as rivers, highways, wind patterns, and topography (highlands *versus* lowlands). The circular

scan statistic, which is designed to detect compact clusters, provides poor delineation of irregularly shaped clusters resulting from such features (Figure 5). Irregularity of shape can also reduce the detection power of the circular scan statistic, or force it to report a single irregular cluster as a series of small clusters (Figure 6).



Cholera outbreak along a river flood-plain

•Small circles miss much of the outbreak

•Large circles include many unwanted cells

Figure 5. Circular spatial scan statistic zonation.

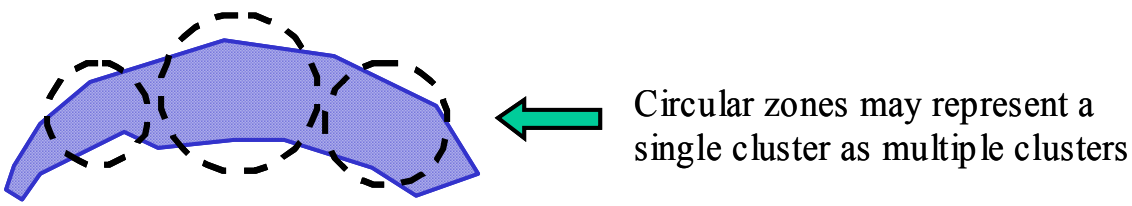


Figure 6. Circular scan statistic represents a single actual cluster as a series of small clusters.

Consider a tessellation of a geographic region with a response value assigned to each cell of the tessellation. Echelon analysis is an objective means of analyzing the spatial structure of the response function. In environmental and remote sensing applications, the tessellation is often regular (for example, pixels in an image or EMAP hexagons) but the basic logic of the method does not require this. Examples of environmental response functions include: (*i*) grey-scale level in a raster image, (*ii*) a pixel-referenced index of change between two images, (*iii*) abundance of a species across a cell, and (*iv*) species richness across a cell.

There is an analogy between echelon analysis and Morse theory. The latter considers smooth response functions $G(x)$ defined over a differentiable manifold. For each value g of the response function, the "upper level set" is defined as $\{x : G(x) \geq g\}$. Morse (1934) studied how the topological structure (specifically, the higher order connectivity) of the upper level sets changes as the level g decreases; see Milnor (1963) and Matsumoto (2002).

In echelon analysis, the response function G is defined over a finite set M of cells instead of a differentiable manifold. The differentiable structure on M is replaced by a nearest-neighbor structure. For example, a raster image has two standard nearest-neighbor structures depending on whether

diagonally adjacent cells (pixels) are considered to be neighbors. A "path" in M is then a sequence of cells in M with the property that every two consecutive cells in the sequence are nearest neighbors. A subset Z of M is "connected" if any two cells in Z can be joined by path that lies entirely within Z. Every subset X of M can be written uniquely as a disjoint union of maximal connected subsets called the connected components of X. The term "connectivity" of X refers to the number and identity of the connected components Z of X.

Echelon analysis studies how the connectivity of the upper level sets $\{x : G(x) \geq g\}$ changes as the level g decreases. For example, if the response is species richness in a cell then for large values of g the upper level set may consist of a few isolated pockets of high species richness. As g decreases, corridors of lower richness may join up these pockets. This changing pattern of connectivity can be quite varied and complex. At the simplest, there might be only one primary region of high richness that grows by gradually accumulating neighboring cells as g decreases. A slightly more complex pattern has two primary regions, each growing by gradually accumulating their neighboring cells, but which do not join up until g is very small. Echelon analysis represents each of these various patterns as a rooted tree with a height function defined on the nodes of the tree. Nodes at height g are in one-to-one correspondence with the connected components of the upper level set $\{x : G(x) \geq g\}$.

Echelon analysis may be used in conjunction with the spatial scan statistic to more clearly delineate cluster boundaries, since echelon families identify the spatial connectivity of a response surface. For example, two isolated first order echelons may be connected by a common second order echelon, as identified by "saddle point" mapping units. Echelons at any hierarchical level may be tested for statistical significance by the spatial scan statistic approach. Therefore, the combination of these two different methods will result in the determination of spatially disjoint areas of significantly elevated disease rates. Essentially, echelon analysis mechanizes and objectifies the way a person may look at a map and quickly determine a reasonable set of candidate zones, while eliminating many other zones as obviously uninteresting.

Therefore, echelon analysis used in conjunction with the spatial scan statistic may improve disease surveillance for programs that currently apply the scan statistic. For example, we have revisited the North Carolina SIDS mortality data over counties during 1974–84 analyzed by Kulldorff (1997). The following diagrams show the hotspot results of the two approaches based on circles and echelons as the choices for candidate zones. It would be interesting to conduct several comparative investigations.

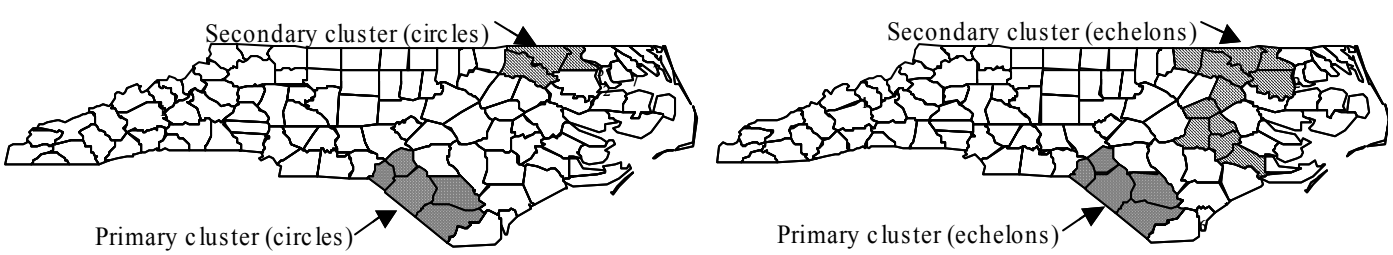


Figure 7

6. Atlantic Slope Watersheds: Disturbance and Vulnerability Pixels

Definitions of Disturbance and Vulnerability

In order to delineate disturbed, vulnerable pixels, two criteria must be established:

- Land uses or land use patterns which are considered to be indicative of human alteration of the landscape, or "disturbance", and
- Landscape parameters that indicate susceptibility to adverse environmental impact, or "vulnerability".

Specific criteria are listed as follows.

Disturbance

The most basic indication of potential alteration of the land surface from its natural condition is a change in land cover from forested to other cover types which are indicative of various human activities. Land cover, therefore, is assumed to be representative of land use. In addition, land use/land cover is one piece of information that is universally available in the study area, and is thus a convenient indicator of the extent of human disturbance. Land cover data that were used in this study were produced as part of a cooperative project between the U.S. Geological Survey (USGS) and the U.S. Environmental Protection Agency (USEPA) to produce a consistent, land cover data layer for the conterminous U.S. based on Landsat Thematic Mapper (TM) data. National Land Cover Data (NLCD) were developed from TM data acquired by the Multi-Resolution Land Characterization (MRLC) Consortium. The MRLC Consortium is a partnership of federal agencies that produce or use land cover data. Partners include the USGS (National Mapping, Biological Resources, and Water Resources Divisions), USEPA, the U.S. Forest Service, and the National Oceanic and Atmospheric Administration (Loveland & Shaw 1996). The goal of the MRLC is to provide data across a range of spatial and temporal scales for the analysis and monitoring of environmental change (Loveland & Shaw 1996). Spatial resolution, or pixel size, is 30m on a side (900 m^2) and the images were preprocessed to correct for electronic problems and geographic referencing by the MRLC. Images for the mid-Atlantic region were acquired from dates ranging from 1989 to 1994.

The general NLCD procedure is to classify regions from mosaiced TM scenes using spectral information (Vogelmann et al. 1998, and Brown et al. 1993). The initial classification was then interpreted and corrected using any number of ancillary data sets that include: aerial photographs, Digital Terrain Elevation Data (DTED) and derived slope, aspect and shaded relief; population and housing density data from census data; USGS land use and land cover (LUDA); and National Wetlands Inventory (NWI) data if available. Other ancillary data sources may include soils data, unique state or regional land cover data sets, or data from other federal programs such as the National Gap Analysis Program (GAP) of the USGS Biological Resources Division (BRD). Frequently derived data sets based on the TM data were also used to separate confusing areas, for example the normalized vegetation index (NDVI) and various TM band ratios. To help discriminate areas seasonally different two distinct TM mosaics are produced, a leaves-on version (summer) and a leaves-off (spring/fall) version. The first NLCD version for the mid-Atlantic region was completed in 1996. Based on local user input and new ancillary data sets, improvements, on a state-by-state basis, have continued and the currently available version was updated in April 2000. The NLCD presents 25 land cover classes in the Mid-Atlantic region, corresponding to Level I and II Andersen Land Cover/Land Use categories

Anderson et al., 1976. These were subsequently grouped into 9 more general classes, considered to be more appropriate for our assessment, as presented in Table 1.

Table 1. Land cover/land use classes utilized in the Atlantic Slope project, and their corresponding NLCD classes.

Atlantic Slope	NLCD – Anderson 1	NLCD – Anderson 2
Water	Water	11 Open Water
Not Applicable		12 Perennial Ice/Snow
Suburban	Developed	21 Low Intensity Residential
Urban		22 High Intensity Residential
Urban		23 Commercial/Industrial/Transportation
Rock	Barren	31 Bare Rock/Sand/Clay
Rock		32 Quarries/Strip Mines/Gravel Pits
Transitional		33 Transitional
Forest	Forest Upland	41 Deciduous Forest
Forest		42 Evergreen Forest
Forest		43 Mixed Forest
Forest	Shrubland	51 Shrubland
Row Crops	Non-natural Woody	61 Orchards/Vineyards/Other
N-A	Herbaceous Upland	71 Grasslands/Herbaceous
Pasture	Herbaceous Planted/Cultivated	81 Pasture/Hay
Row Crops		82 Row Crops
N-A		83 Small Grains
N-A		84 Fallow
Suburban		85 Urban/Recreational Grasses
Forest	Wetlands	91 Woody Wetlands
Emergent Wetlands		92 Emergent Herbaceous Wetlands

It is generally assumed that pre-colonial land cover in the majority of the Mid-Atlantic region consisted of forest, naturally-occurring waterbodies, wetlands, and rock. Although specific forest type has been altered over the entire region due to widespread clear-cutting in the latter part of the 19th century, mature forest, along with the other classes listed above, is assumed to represent the least humanly altered type of landscape. In the Mid-Atlantic, dominant human activities which alter land use are agriculture and urbanization. While other human activities certainly exist, e.g., mining, agriculture and urbanization represent the largest visible footprint of human activity in this region. Any study of the impact of human activity on aquatic resources should thus investigate the marginal distribution of land

use in the study watersheds based on these three land use classes: agriculture, urban, or forest. These classes could clearly be further subdivided, for example the impacts of row crops are different than those of pasture, which are both lumped into agriculture under this classification system. But for the purpose of devising a simple but still informative watershed classification system, these three categories seemed broad, but still distinct and ecologically relevant.

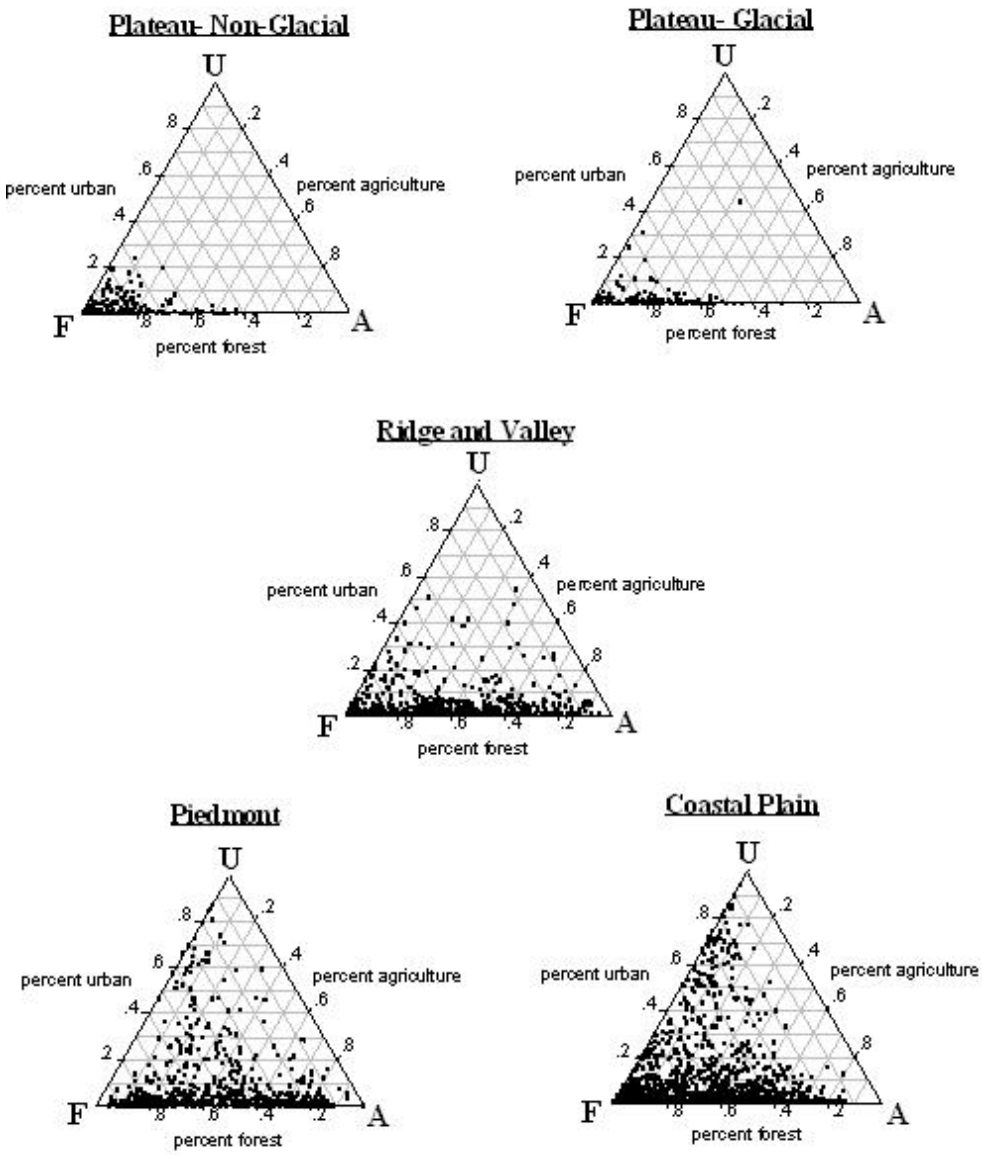


Figure 8. Ternary plots of watersheds of Atlantic Slope by physiographic province.

As a way to visually represent the composition of land use in watersheds by these three categories, we plotted watersheds on a ternary plot. Like the triangular graphs used to plot soil types based on the composition of sand, silt, and clay, these diagrams have three axes, one each for the percentage of agriculture, urban, or forest land cover in the watershed. We plotted all of the study area watersheds on ternary plots by the five study area Physiographic Provinces: Coastal Plain, Piedmont, Glaciated and Non-glaciated Appalachian Plateau, and the Ridge and Valley. Distribution of land cover within the 14

digit HUC watersheds are depicted, for each physiographic province, in ternary plot format in Figure 8. In the Mid-Atlantic, it is obvious from these plots that land use departs from forested along two major trajectories: conversion of forested to agricultural use, or conversion of forested to urban land use.

However, for the purposes of this analysis, we are not concerned with the actual direction of the trajectories, only that a departure from reference (i.e., forested) has occurred. Therefore, we need only to split the land cover into two categories: natural or disturbed. As was discussed previously, natural land covers are considered to be forest, water, and wetlands land cover classes. All other uses are considered to be altered by human activity. For these reasons, all land covers other than forested, water, and wetland are considered to be "disturbed".

Vulnerable

A primary goal of the Atlantic Slope project is the development of an intellectual model relating human activity to aquatic resource condition. This implies a relationship between land use in the watershed (as a source of environmental stressors) and the receiving water body (the sink). Relevant characteristics of the source are the type of stressors generated, and their magnitude. Therefore, if we utilize the definition of "disturbance" developed in the previous section, we also need a descriptor of its magnitude as an indicator of connection between land use and aquatic resources. A primary determinant of the magnitude of a stressor (e.g., nutrient enrichment, sedimentation) derived from a specific land use is the slope of the land surface. In general, higher slope areas represent a greater risk of adverse environmental impact from human disturbance. For example, higher slope leads to decreased infiltration, increased runoff, rapid runoff, and minimal contact time between runoff and soil for remediation and/or impact. We recognize that in some instances, steep slopes can represent a lower risk for some environmental stressors, such as susceptibility to atmospheric deposition. However, since the project is primarily concerned with water-mediated processes, we generally assume that higher slope areas represent a greater risk for adverse environmental impact associated with human disturbance.

The U.S. Natural Resources Conservation Service (NRCS) maps soil types in three major slope classes: 0-3%, 3-8%, and >8%. Applications such as the Universal Soil Loss Equation (USLE) utilize these major classes in predicting soil loss through erosion, and a number of other applications utilize similar classes to predict hydrologically derived processes such as nutrient transport. In general, without regard to soil type, higher slope areas represent a greater risk of adverse environmental impact from human disturbance.

There is a clear connection between land uses and slope. For example, high slope areas (>8%) are not generally used for row crop agriculture or urban areas, since lower slope areas are more feasible and cost effective. Agricultural activities on higher slope areas require specialized practices to minimize soil loss, loss of soil additives such as herbicides and pesticides, and are generally not recommended by conservation agencies such as NRCS. However, human activities such as agriculture and urban use do occur on a range of slopes from 0-8%. Given the existence of two slope classes within this range (0-3% and 3-8%), we consider human activities on slopes of 3-8% to be vulnerable.

A special note is required for the Coastal Plain physiographic province. Slopes greater than 3% do not occur in the Coastal Plain. In addition, it is likely that the definition of disturbance would need to be specifically defined for this province, since the nature of human disturbance is unique. Most of the

regions large urban centers occur on the Coastal Plain, along large rivers. The nature of these activities are radically different from those upstream in the other physiographic provinces, and are not adequately described by the definitions utilized in this exercise. Therefore, the Coastal Plain is not included in this assessment.

7. Atlantic Slope Watersheds: Nodes, Nodal Variance, and Disturbance Nodes

Given the nature of the Atlantic Slope project, and its aquatic resource mandate, it was critical to investigate other landscape metrics that could assist in the elucidation of the relationship between land use and aquatic resource condition. While many studies have found watershed-wide land use to be helpful in the prediction of various impacts, a more specific characterization of the riparian area was desired, given its proximity to the stream and its potential impact on stream condition. Therefore, a more specific characterization of human disturbance in the riparian area was required.

It must be noted that a primary objective of our research is the relationship between human activities and aquatic resource condition. Therefore, the manner in which we characterize the landscape must be compatible with the manner in which we measure condition. Ecological condition of streams and wetlands is determined at an individual point in the stream, and is assumed to vary more or less continuously along the length of the stream to the next sampled point. Thus, ecological condition is an attribute associated with a point. This determines the form of the associated landscape metric, which must also be associated with a point to be compatible with the ecological characterization. The next question is obviously the shape of the bounded area, which should satisfy the following constraints (Schuft et al., 1999). The bounded area should:

- Adequately capture the stream network;
- Be conceptually defensible in relation to the functions being considered;
- Capture most of the riparian vegetation and an appropriate amount of surrounding land use to allow for associative analysis; and
- Be cost-efficient.

Thus, the concept of a circle emerges, where relevant landscape properties are measured and attributed to the center point. A number of studies have utilized a circular bounded area for determination of landscape metrics associated with ecological and physical processes, with good success. Thus, we utilize the concept here.

The next question is the location of the circles themselves. Certainly, we strive to increase the power of the relationship between landscape metrics and ecological condition of streams, and therefore consider bounded areas physically neighboring the aquatic resource of concern. Once we narrow our view to the riparian corridor (that area buffering the stream on either side, and of undetermined width), two locations present themselves for consideration. The first are locations which have been ecologically characterized during previous investigations. Unfortunately, these locations are often randomly selected along the entire length of a stream, and may not be as informative as one wishes. These points are often thinly distributed, with only minimal representation of a high number of stream miles. For example, a point may be located in a section of a stream, downstream of a number of tributaries draining watersheds with varying land uses. Any specificity or direct correlation between specific land uses and aquatic condition is lost. The second is at stream convergences, or those points which represent connections of the stream network. We hereby term these locations nodes, since they

satisfy the common use of the term as intersections areas of corridors, and as sources or sinks of flowing objects (Forman and Godron, 1986). In this context, nodes function as relayers of water movement, rather than as ultimate destinations. Nodes offer a number of advantages as locations at which to describe land use. As indicated previously, they have a near stream location, suggesting that more precise landscape metrics can be calculated. However, their main advantage is their amenability to network analysis. Nodes are a direct reflection of the complexity of the drainage network, and will increase in density along with the drainage complexity.

Nodes were placed at convergences of stream tributaries of third order (Strahler method) or higher, as depicted in Figure 9. First and second order stream nodes were not utilized for the following reasons:

- At least in some physiographic provinces, such as the Ridge and Valley, headwaters are less impacted or mostly forested. Use of the 1st order nodes might under-represent the actual impact in the watershed.
- First order streams were not represented consistently in the streams coverage throughout the study area, and are really only available in Pennsylvania. Therefore, in Pennsylvania nodes were selected that were not headwaters, where in other states all nodes were used.
- Feasibility of analysis was reasonable utilizing nodes on third order streams and higher. Inclusion of the 1st and 2nd order nodes doubled the number of nodes in Pennsylvania (where these streams were represented in the stream coverage).

The next criterion to be established was the size of the bounded area to be characterized at the node. Based on previous work, we established circles of three radii for assessment: 1 km, 2 km, and 4 km. Analysis of landscape metrics of the three circle sizes showed that the highest resolution of differing land uses occurs with the 1-km radius circle; landscape metrics based on the 2 and 4 km radius circles begin to converge on watershed-wide values. Depiction of the 1-km circles centered on the nodes is presented in Figure 9. Calculation of landscape metrics at these nodes, rather than watershed-wide values, allows statistical descriptions of the watershed to be assessed, since a population of nodal values is available. Their nested nature also allows observation of the change in landscape parameters as one moves away from the stream. Finally, nodes are a direct reflection of the complexity of a stream, with more nodes occurring in a more complex drainage system.

It is worth noting that previous studies have utilized the characterization of stream buffer areas rather then nodes. This study chose nodal analysis over a buffer-based one because of some unique features of the node-based approach. Buffers do not capture the complexity of the drainage network in the same manner as the nodal approach (e.g., using nodes gives a higher weight to more complex stretches of the stream). It is arguable that in areas of higher drainage complexity, more complex aquatic habitat may occur, and landscape parameters in these areas may call for a higher weighting in the final analysis. The nodal approach achieves this weighting, by virtue of the ensuing overlapping of nodal circles, as depicted in Figure 9.

Finally, some attribute of the nodes must be selected as an indicator of potential risk to the aquatic resource. Ongoing work has suggested that a threshold of 45% forested land cover in an area is critical for maintenance of ecological integrity, for physical, biogeochemical, and biologic functions. For this reason, nodes with less than 45% forested land cover were identified as areas of elevated risk.

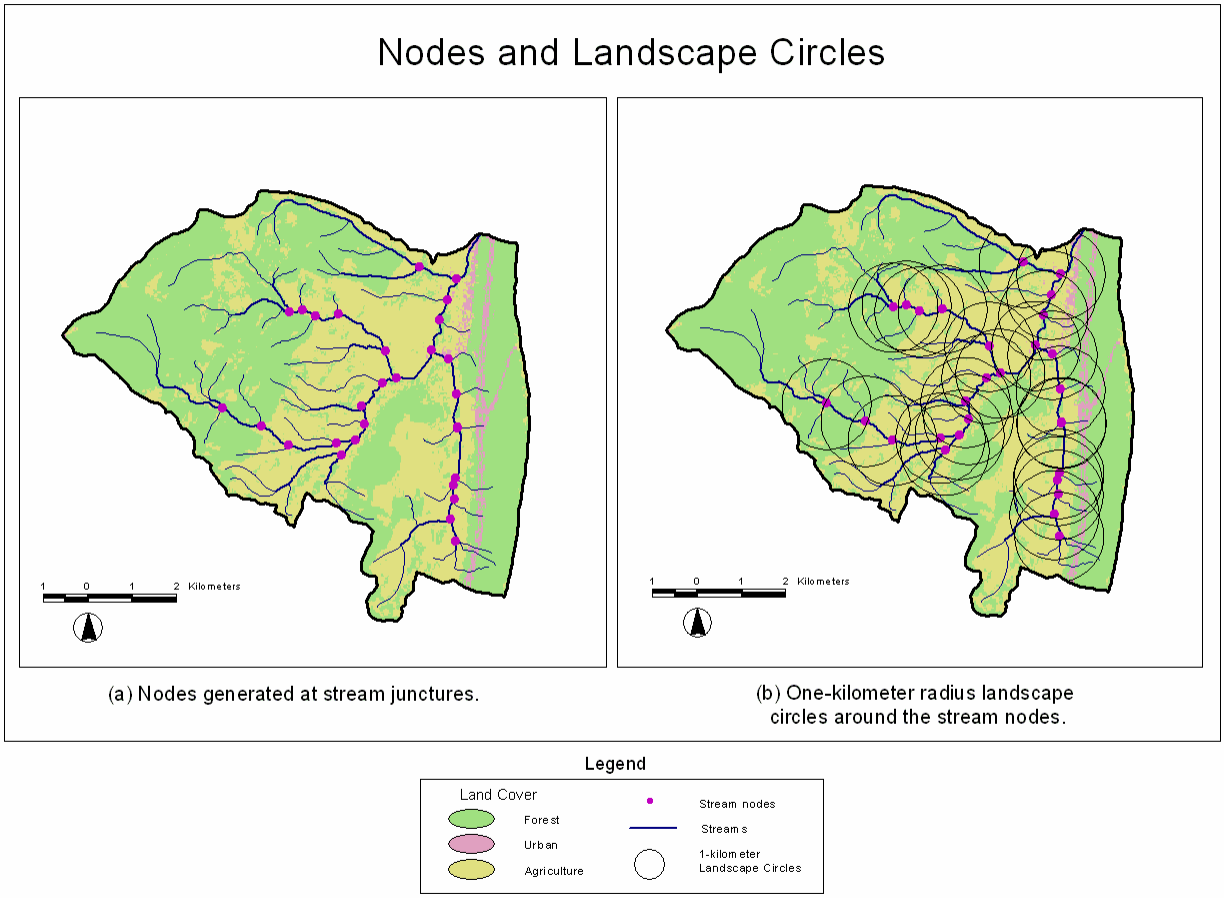


Figure 9. Nodes and landscape circles.

8. Critical Areas in Ridge and Valley Physiographic Province Based on Pixels and Nodes Using Spatial Scan Statistic

It is obvious from Figure 10 that further investigation into large hot spots, such as those present in the Ridge and Valley analysis, would be informative. For this reason, a second-level hot spot identification was performed. The results of this second-level hot spot identification, based on both the nodal and total disturbed pixels, are depicted in Figures 11. Two areas require further examination for elucidation of the land use patterns driving the critical area determination:

- The second-level hot spot derived from the use of the nodal information
- The second-level hot spot derived from the use of disturbed/vulnerable pixel information

As is evidenced for the Ridge and Valley first-level hot spot, the two analyses provide second-level hot spots in different spatial portions of the area. However, it appears that they are closely related, i.e., the hot spot identified by the nodal analysis would be a second cluster identified in the disturbed/vulnerable pixel analysis, and vice versa. This is suggested by the ordinal maps produced by each analysis, and presented as Figures 12 and 13. In general, there is a co-occurrence of high relative risk watersheds

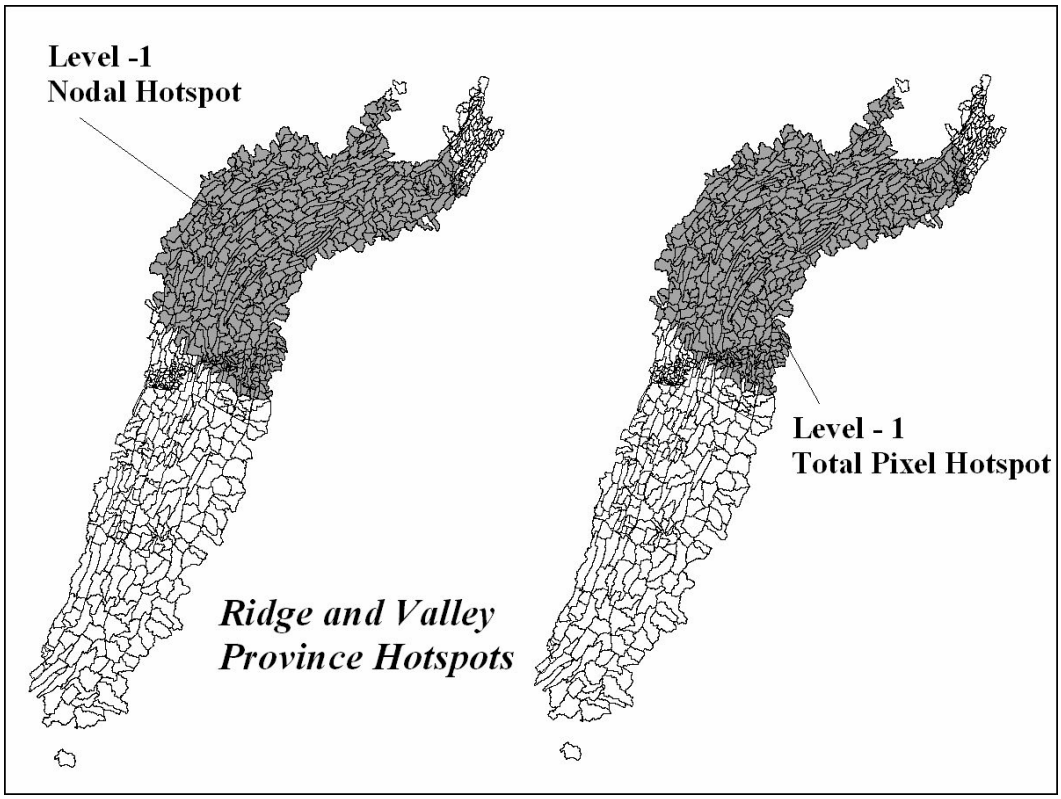


Figure 10. Ridge and Valley hotspots.

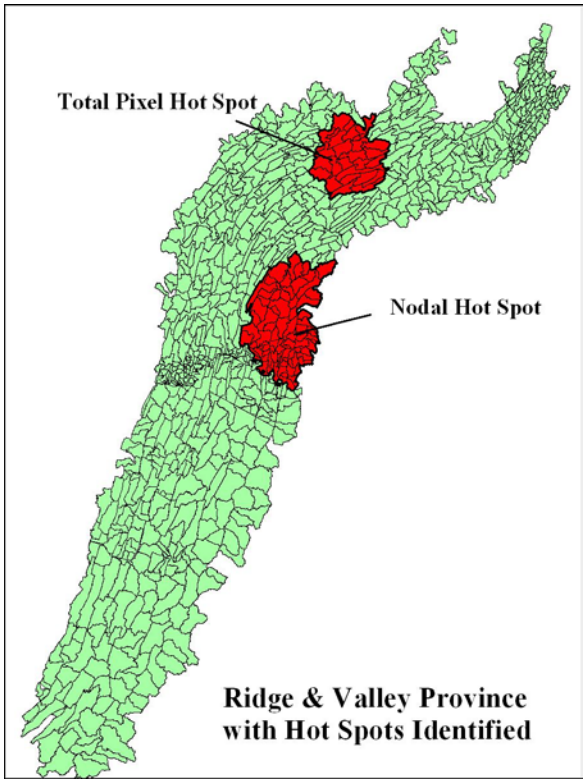


Figure 11. Ridge and Valley hotspots within hotspots.

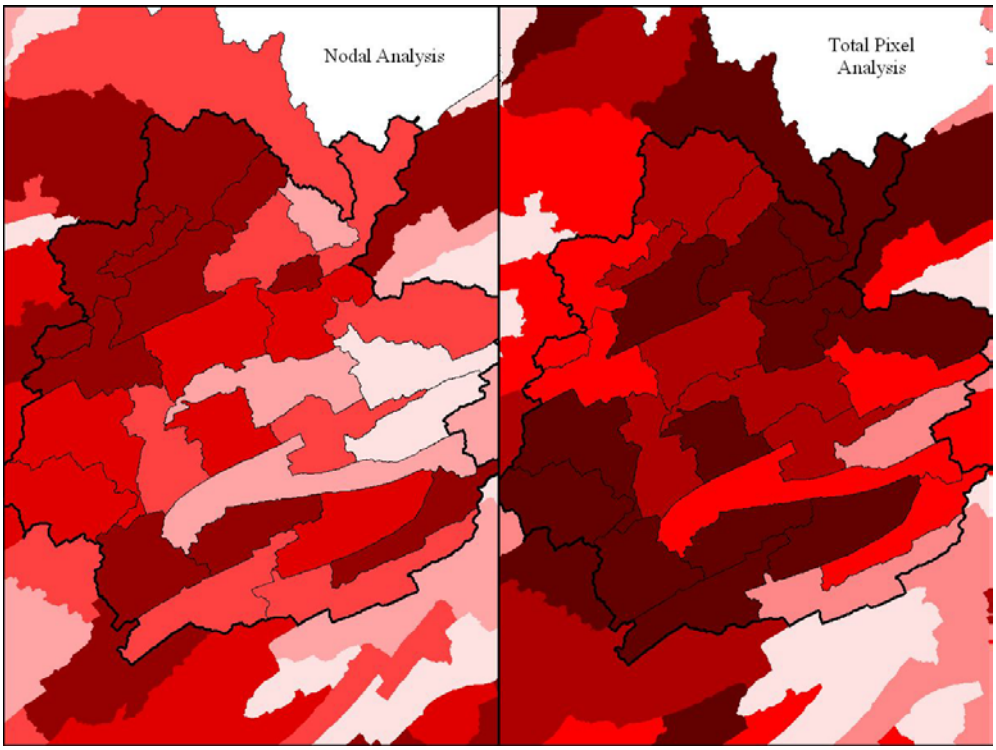


Figure 12. Shaded ordinal view of total pixel hotspot.

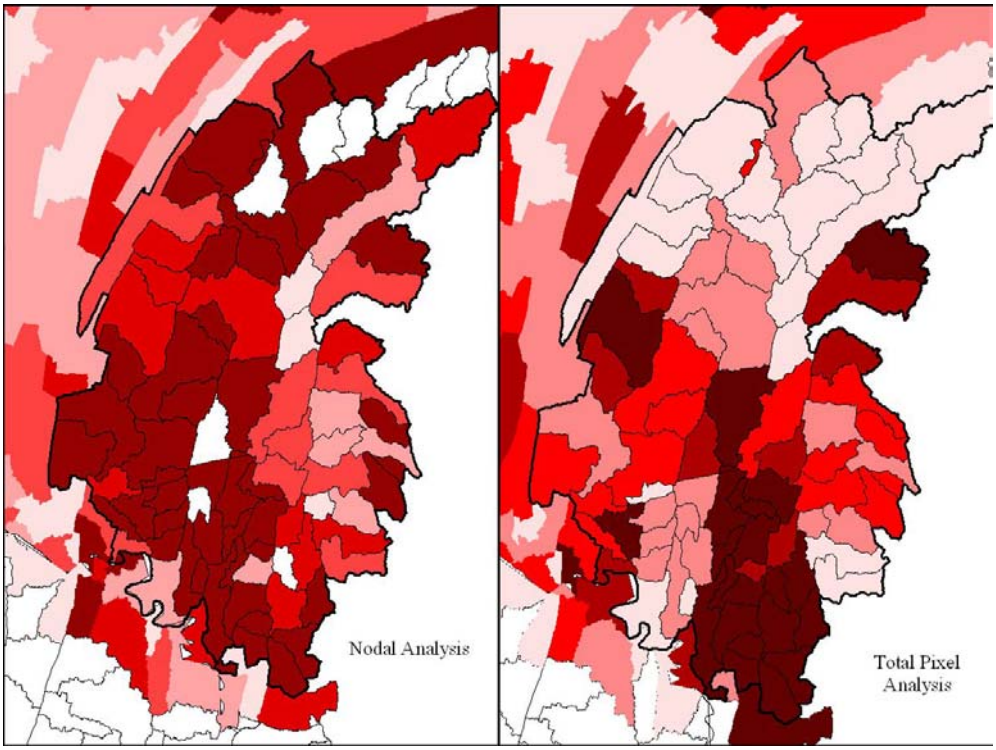


Figure 13. Shaded ordinal view of nodal hotspot.

from each of the two analyses. It is of interest, however, to investigate when the two analyses converge and depart.

In the second level hot spot identified by the nodal analysis, Figures 14 and 15 present both land cover and the occurrence of low and high slope coverage in the individual watersheds. It is obvious that the critical area is dominated by watersheds with an exceedingly high area of the watershed in the low slope category, and a general relationship exists between areas of low slope and human activity. The large low slope area is commonly identified as the Great Valley, and the extreme width of the valley makes it highly amenable to human activity such as agriculture. In these watersheds, most of the stream convergences of third order or higher will indeed be down on the valley floor, and the nodes will be placed there. The co-occurrence of broad valley floor representing a high percentage of the watershed land area, the preferential location of nodes on this valley floor, and the amenability of the topography for agricultural land use result in a high level of agreement between the two forms of analysis (node vs. pixel-based). This can be seen when watersheds of high relative risk identified by both analyses are located on the ordinal maps and then compared to the corresponding portions of the map depicting land use and slope information.

In contrast, the two analyses depart when valley floors are more narrow, and a higher percentage of the watershed is covered by forested, high slope areas. In these instances, the nodes will still be concentrated in the narrow valley floor, where human activity is also concentrated, and the watershed will be assigned a high relative risk. However, a relatively low percentage of the watershed area is available for human disturbance (due to the presence of forested, high slope ridge areas), and the overall low occurrence of disturbed/vulnerable pixels will assign a low relative risk. A related case of departure will occur when some portions of the valley floor are on slopes only slightly above the 3% threshold. In these instances, differences occur between the two analyses, but only slightly (i.e., one quintile category difference).

Slightly different patterns of convergence and departure are apparent when investigating the second level hot spot determined by the disturbed/vulnerable pixel analysis, although the general patterns remain the same. It is probable that the more even distribution of low, moderate, and high slope areas in these watersheds complicates the elucidation of strong patterns. In general, the ordinal maps from the two analyses show general agreement, although some disparity in quintile class is apparent. Further investigation into specific land use patterns is necessary in this area.

It should be noted, when interpreting the nodal analysis results, that overlapping of nodes in areas of high drainage density is not only possible, but probable. This can occur most frequently in the Ridge and Valley physiographic province, where stream convergences of third order or greater will occur in dense patterns on the valley floor. It follows that the valley floor area, with probable human activity, will be over sampled. It is not clear how this affects the identification of hot spots, although at the scale of the 14-digit HUC watersheds utilized in this analysis the effect appears to be minimal.

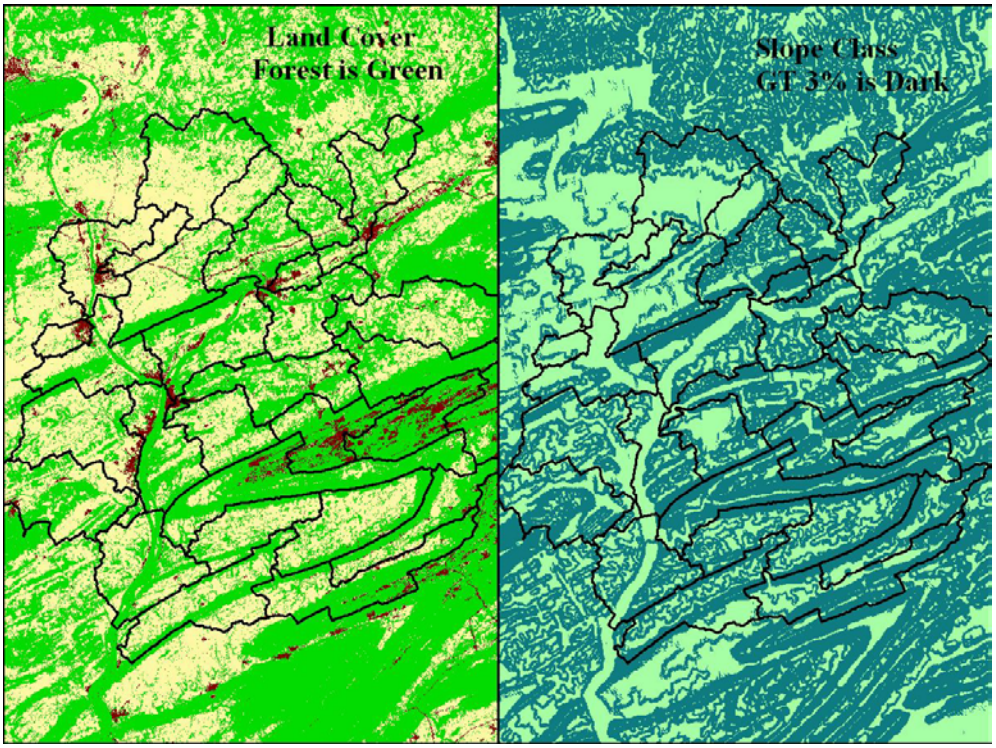


Figure 14. Land cover and slope class for the nodal hotspot.

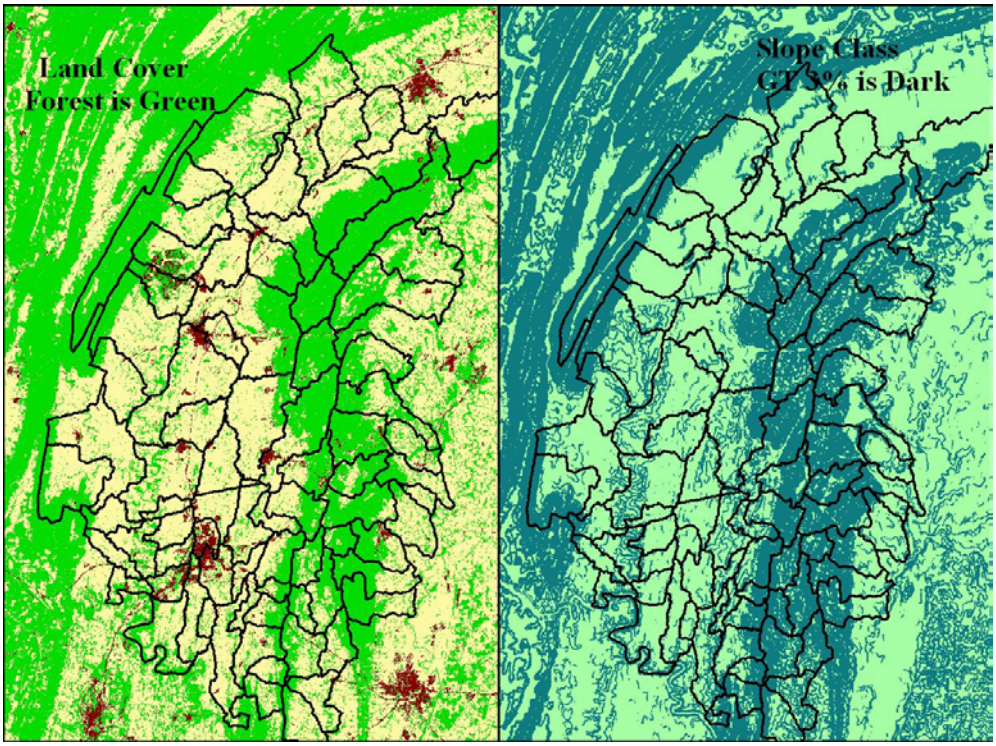


Figure 15. Land cover and slope class for the total pixel hotspot.

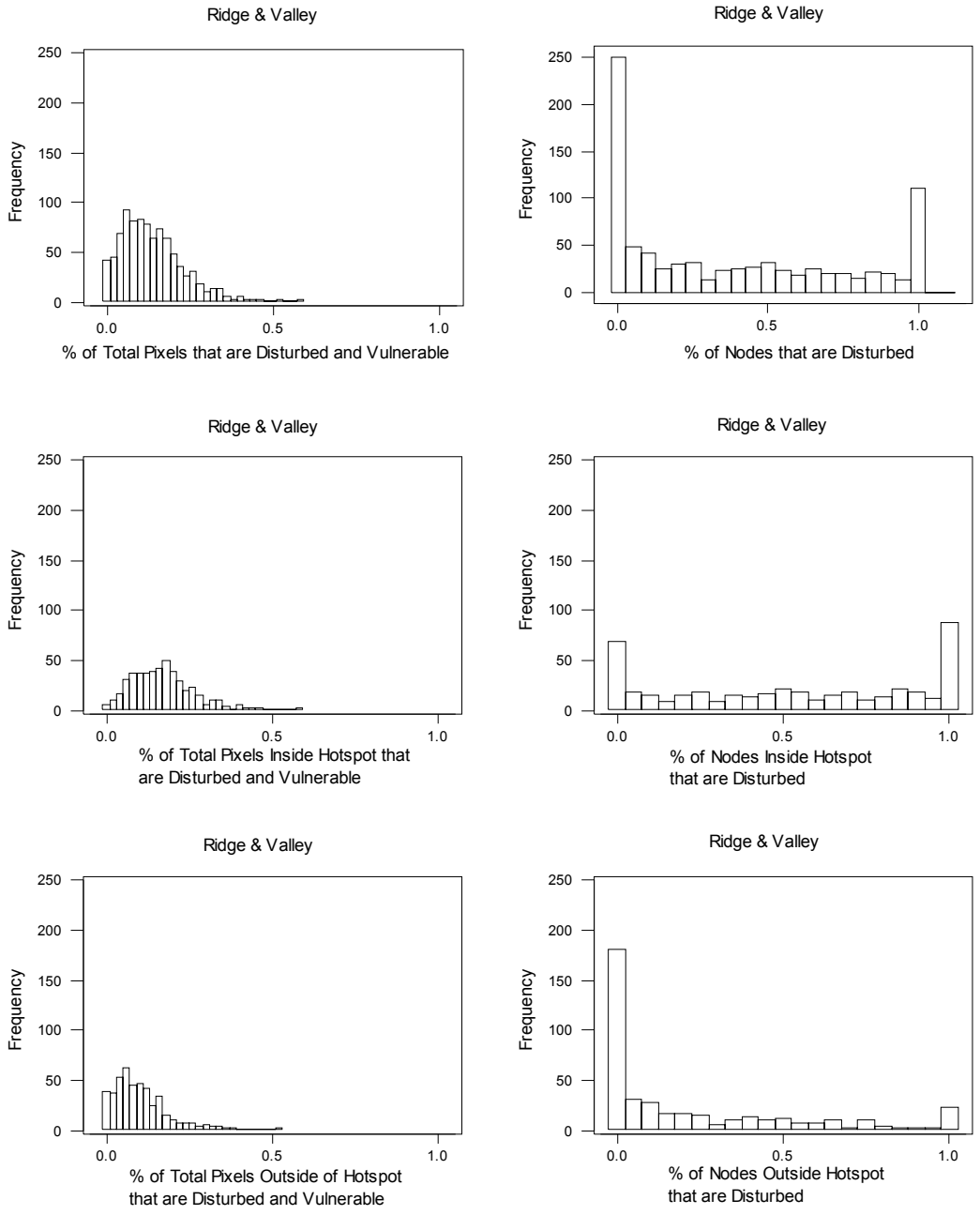


Figure 16. Frequency histograms showing how the proportions of affected pixels/nodes vary across the watersheds. The diagrams on the left refer to the total pixels analysis, where affected pixels are those that are both disturbed and vulnerable. The diagrams on the right refer to the nodal analysis, where affected nodes are those that are disturbed. The frequency distributions are for all watersheds (top), watersheds in the hotspot (middle) and watersheds outside the hotspot (bottom).

9. Literature Cited

Anderson, A. P., Hardy, E. E., Roach, J. T., and Witmer, R. E. A land use and land cover classification system for use with remote sensor data. U.S. Geological Survey Professional Paper 964, 1976.

Anyamba, A., Linthicum, K. J., Mahoney, R., Tucker, C. J., and Kelley, P. W. Mapping potential risk of Rift Valley fever outbreaks in African savannas using vegetation index time series data. *Photogrammetric Engineering & Remote Sensing*, **68(2)**, 137–145, 2002.

Brown, J. F., Loveland, T. R., Merchant, J. W., Reed, B. C., and Ohlen, D. O. Using multisource data in global land characterization – concepts, requirements, and methods. *Photogrammetric Engineering & Remote Sensing*, **59**, 977–987, 1993.

Fevre, E. M., Coleman, P. G., Oditt, M., Magona, J. W., Welburn, S. C., and Woolhouse, M. E. J. The origins of a new trypanosoma brucei rhodesiense sleeping sickness outbreak in Eastern Uganda. The Lancet, **358**, 625–628, 2001.

Forman, R.T.T., and Godron, M. *Landscape Ecology*. Wiley, New York, 1986.

Janardan, K. G., and Patil, G. P. A unified approach for a class of multivariate hypergeometric models. *Sankhya Series A*, **34**, 363–376, 1972.

Johnson, G. D., Myers, W. L., Patil, G. P., and Walrath, D. Multiscale analysis of the spatial distribution of breeding bird species richness using the echelon approach. In *Assessment of Biodiversity for Improved Forest Planning*, P. Bachmann, M. Kohl, and R. Paivinen, eds. Kluwer Academic Publishers, pp. 135–150, 1998.

Joshi, S. W., and Patil, G. P. A class of statistical models for multiple counts. In *Random Counts in Scientific Work, Volume 2: Biomedical and Social Sciences*. G. P. Patil ed. The Pennsylvania State University Press, University Park, pp. 189–203, 1970.

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

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

Kurihara, K., Myers, W. L., and Patil, G. P. Echelon analysis of the relationship between population and land cover pattern based on remote sensing data. *Community Ecology*, **1(1)**, 103–122, 2000.

Lillesand, T. M., and Kiefer, R. W. *Remote Sensing and Image Interpretation. 3rd edition*. John Wiley & Sons, New York, NY, 1994.

Loveland, T. R., and Shaw, D. M. Multiresolution land characterization: building collaborative partnerships. In *Gap Analysis: A Landscape Approach to Biodiversity Planning*, J. M. Scott, T. H.

Tear, and F. W. Davis, eds. American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, pp. 79–85, 1996.

Matsumoto Y (2002). *An Introduction to Morse Theory*, Translations of Mathematical Monographs, Vol. 208. American Mathematical Society, Providence, Rhode Island.

Milnor J (1963). *Morse Theory*. Princeton University Press, Princeton, New Jersey.

Morse M (1934). *The Calculus of Variations in the Large*. Colloquium Publications, Vol. 18. American Mathematical Society, New York.

Myers, W. L., Patil, G. P. Echelon Analysis. In *Encyclopedia of Environmetrics, Volume 2*. A. El. Shaarawi and W. W. Piegorsch, eds. Wiley, pp. 583–586, 2002.

Myers, W. L., Patil, G. P., and Joly, K. Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics*, **4(2)**, 131–152, 1997.

Myers, W. L., Patil, G. P., and Taillie, C. Comparative paradigms for biodiversity assessment. Invited paper at the IUFRO Symposium in Chiang-mai, Thailand. In *Measuring and Monitoring Biodiversity in Tropical and Temperate Forest*s, T. J. Boyle and B. Boontawee, eds. CIFOR, Bogor, Indonesia, pp. 67–85, 1995.

Myers, W. L., Patil, G. P., and Taillie, C. Conceptualizing pattern analysis of spectral change relative to ecosystem health. *Ecosystem Health*, 5**(4)**, 285–293, 1999.

Patil, G. P., and Ratnaparkhi, M. V. Sum-symmetric power series distributions. In *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz, and N. L. Johnson, eds. John Wiley, New York, pp. 80–83, 1988.

Patil, G. P., and Taillie, C. Topological concepts and definitions for echelons and echelon trees. Technical Report No. 99-0602, Center for Statistical Ecology and Environmental Statistics, Penn State University, pp. 5, 1999. (Draft).

Ramakomud, A. Change detection using hyperclustered data: the spatial averaging approach. MSc. Thesis, Pennsylvania State University, University Park, 1998.

Rodriguez-Iturbe, I., and Rinaldo, A. *Fractal River Basins: Chance and Self-Organization*. Cambridge University Press, Cambridge, UK, 1997.

Rogers, D. J., Myers, M. F., Tucker, C. J., Smith, P. F., White, D. J., Backenson, B., Edison, M., Kramer, L. D., et al. Predicting the distribution of West Nile fever in North America using satellite sensor data. *Photogrammetric Engineering & Remote Sensing*, **68**, 112–114, 2002.

Schuft, M. J., Moser, T. J., Wigington, Jr., P. J., Stevens, Jr., D. L., McAllister, L. S., Chapman, S. S. and Ernst, T. L. Development of landscape metrics for characterizing riparian-stream networks. *Photogrammetric Engineering & Remote Sensing***, 65,** 1157–1167, 1999.

Smits, P. C., and Myers, W. L. Echelon approach to characterize and understand spatial structures of change in multi-temporal remote-sensing imagery. *IEEE Trans. Geoscience and Remote Sensing*, **38(5)**, 2299–2309, 2000.

Vogelmann, J. E., Sohl, T., and Howard, S. M. Regional characterization of land cover using multiple sources of data. *Photogrammetric Engineering & Remote Sensing*, **64**, 45–47, 1998.