

## 22 Spatial Cluster Analysis

---

GEOFFREY M. JACQUEZ

“We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression. In the theory of probability, as developed in its application to games of chance, we have the classic example proving this possibility. If the gambler's apparatus are really true or unbiased, the probabilities of the different possible events, or combinations of events, can be inferred by a rigorous deductive argument, although the outcome of any particular game is recognized to be uncertain. The mere fact that inductive inferences are uncertain cannot, therefore, be accepted as precluding perfectly rigorous and unequivocal inference” (Fisher 1935).

“Humility is indeed wise for the spatial analyst!” (Bailey and Gatrell 1995)

Spatial cluster analysis plays an important role in quantifying geographic variation patterns. It is commonly used in disease surveillance, spatial epidemiology, population genetics, landscape ecology, crime analysis and many other fields, but the underlying principles are the same. This chapter provides an overview of a probabilistic approach that is the foundation of spatial cluster analysis. It first provides a working definition of a cluster, founded on the type of data to be analyzed. The role of cluster analysis in Exploratory Spatial Data Analysis (ESDA) is discussed,

and provides an entrée into five components that underlie statistical pattern recognition. The clustering typology of global, local and focused methods is then defined, followed by an overview of descriptors of cluster morphology. Approaches to quantifying cluster change and persistence are summarized, and issues of multiple testing are addressed. This chapter concludes with an overview of some software resources for undertaking cluster analyses.

## **1 What is a Cluster?**

In order to define a spatial cluster we first must consider the kinds of data that are being studied. The information to be clustered may be event-based, population-based, field-based, or feature-based. Event-based data include point locations (such as the places of residence and time of diagnosis of cases of disease in people, or the locations of a species of tree in a forest) and counts (accidents at particular road intersections). Population-based data incorporate information on the population from which the events arose, and include disease rates with case counts in the numerator and size of the at-risk population in the denominator. Field-based data are observations that are continuously distributed over space, and include concentrations and temperatures. Feature-based data include boundaries and polygons that may be derived from field-based data, such as zones of rapid change in an attribute's value.

A spatial cluster might then be defined as an excess of events (for event- and population-based data, such as a cancer cluster) or of values (for field-based data, such as a grouping of excessively high concentrations of cadmium in soils) in geographic space. For feature-based data, a cluster might be a spatial aggregation of boundaries. But in practice, the term “cluster” is too generic and does not convey information on *cluster morphology*, such as descriptors of magnitude of the excess or deficit, geographic size and shape of the cluster and the locations of spatial

outliers, and descriptors of boundary shape, as described in more detail later in this chapter. For now, it is useful to think of a “cluster” as a spatial pattern that differs in important respects from the geographic variation expected in the absence of the spatial processes that are being investigated. This is a key concept – that “clustering” is always measured relative to a null expectation.

## 2 Why Search for Clusters?

Cluster analysis plays important roles in the construction of spatial models and in ESDA. Model construction requires an understanding of the patterns of spatial variation as one often wants to incorporate relevant features of attribute variability into the model. ESDA involves the identification and description of spatial patterns (such as outliers, clusters, hotspots, cold spots, trends and boundaries), and has two primary objectives:

- **Objective 1:** Pattern recognition using visualization, spatial statistics and geostatistics to identify the locations, magnitudes and shapes of statistically significant pattern descriptors.
- **Objective 2:** Hypothesis generation to specify realistic and testable explanations for the geographic patterns found under Objective 1.

## 3 Statistical Pattern Recognition

Spatial patterns are of interest because they are the trace of space-time processes that are the focus of geographic studies. For example, in cancer research spatial patterns contain the geographic trace of processes, covariates, and factors (such as exposures to environmental carcinogens; access to cancer screening facilities; behaviors mediating cancer risks, and so on) that determine how cancer risk varies across and is expressed within human populations.

There are several approaches to pattern recognition, including visualization techniques that rely on the human visual cortex (e.g. “eye-balling”), kernel-based methods that accentuate differences on a surface (e.g. median smoothing techniques such as headbanging; Kafidar 1996, Gelman et al. 2000); artificial intelligence approaches (e.g. neural network and genetic algorithms; Pandya and Macey 1996), and methods of statistical pattern recognition that are the foundation of ESDA (Bailey and Gattrell 1995; Moore and Carpenter 1999; Jacquez 1998, 2000; Rushton and Elmes 2000). Most commonly used spatial cluster analysis methods are founded on statistical pattern recognition.

*Statistical pattern recognition* proceeds by calculating a statistic (e.g. spatial cluster statistic, autocorrelation statistic, boundary statistic, etc.) that quantifies a relevant aspect of spatial pattern in event-based, population-based, field-based or feature-based data. For human disease this might be a health outcome (e.g. case/control locations, incidence or mortality rate), putative cause (e.g. exposure to agricultural pesticides), risk factor (e.g. smoking prevalence) or access measure (e.g. availability of prostate screening). The numerical value of this statistic is then compared to the distribution of that statistic’s value under a null spatial model. This provides a probabilistic assessment of how unlikely an observed spatial pattern is under the null hypothesis (Gustafson 1998). Waller and Jacquez (1995) formalized this approach by identifying five components of a test for spatial pattern:

- The *test statistic* quantifies a relevant aspect of spatial pattern (e.g. Moran’s I, Geary’s c, LISA, a spatial clustering metric, etc.).
- The *alternative hypothesis* describes the spatial pattern that the test is designed to detect. This may be a specific alternative, such as clustering near a focus, or it may be the omnibus “not the null hypothesis”.

- The *null hypothesis* describes the spatial pattern expected when the alternative hypothesis is false (e.g. uniform cancer risk).
- The *null spatial model* is a mechanism for generating the reference distribution. This may be based on distribution theory, or it may use randomization (e.g. Monte Carlo) techniques. For example, many disease cluster tests employ heterogeneous Poisson and Bernoulli models for specifying null hypotheses (see Lawson and Kulldorff 1999).
- The *reference distribution* is the distribution of the test statistic when the null hypothesis is true.

Comparison of the test statistic to the reference distribution allows calculation of the probability of observing that value of the test statistic under the null hypothesis of no clustering. This five-component mechanism underpins most tests commonly used in spatial cluster analysis.

*Null or Neutral Hypothesis?* There still is some debate as to how the null hypothesis and null spatial model should be specified. Many implementations of spatial cluster tests employ the null hypothesis of Complete Spatial Randomness or CSR. In the real world, CSR is an appropriate model for pure noise processes such as the static on a television screen when a channel with no signal is selected. But most geographic systems are highly complex and a null hypothesis of CSR is rarely, if ever, appropriate. While CSR is useful in some situations, it is not a relevant null hypothesis for highly complex and organized systems such as those encountered in the physical, environmental and health sciences including fields such as geography, spatial epidemiology, and exposure assessment (Liebisch et al. 2002). CSR is not relevant because spatial randomness rarely, if ever, occurs – some spatial pattern is almost always present. Hence, rejecting CSR has little scientific value because it does not describe any plausible state of the system. The term “Neutral Model” captures the notion of a plausible system state that can be used as a reasonable

null hypothesis (e.g. “background variation”). A typology of neutral models that account for differences in underlying population sizes and for regional and local variation in mean values is one mechanism for constructing null hypotheses that are more plausible than CSR (Goovaerts and Jacquez 2004, 2005). The problem then is to identify spatial patterns above and beyond that incorporated into the neutral model. In this chapter we will continue to use the terms “null model” and “null hypothesis”, with the proviso that they denote appropriate levels of spatial pattern expected in the absence of the hypothesized alternative spatial process.

#### **4 Types of Tests**

There are dozens of cluster statistics (see Jacquez et al. 1996a, b; Lawson and Kulldorff 1999, among others, for reviews), and presentation of these statistics would fill most of this book. Instead, we now present the characteristics of *global*, *local*, and *focused* tests with a few commonly used statistics as examples.

*Global* cluster statistics are sensitive to spatial clustering, or departures from the null hypothesis, that occur anywhere in the study area. Many early tests for spatial pattern, such as Moran’s *I* (Moran 1950) were global in nature, and provided one statistic, such as a global autocorrelation coefficient, that summarized spatial pattern over the entire study area. While global statistics can identify whether spatial structure (e.g. clustering, autocorrelation, uniformity) exists, they do not identify where the clusters are, nor do they quantify how spatial dependency varies from one place to another.

*Local* statistics such as Local Indicators of Spatial Autocorrelation (Anselin 1995, Ord and Getis 1995) quantify spatial autocorrelation and clustering within the small areas that together comprise the study area. Many local statistics have global counterparts that often are calculated

as functions of local statistics. For example, Moran's global autocorrelation statistic is the scaled sum of the LISA statistics that are calculated as:

$$L_i = z_i \sum w_{ij} z_j \quad (1)$$

Here  $L_i$  is the LISA statistic for area  $i$ ,  $z_i$  is the observation at location  $i$ , scaled to have a mean of 0 and unit standard deviation (a  $z$ -score), and the term in the summation is the average within those areas immediately adjacent to the  $i$ th area. Local statistics thus can tell you the nature of spatial dependency (e.g. not significantly different from the null expectation, cluster of high values, cluster of low values, and high or low spatial outlier) in a given locality, while also providing a global test.

*Focused* statistics quantify clustering around a specific location called a *focus*. These tests are particularly useful for exploring possible clusters of disease near potential sources of environmental pollutants. For example, Lawson (1989) and Waller et al. (1992) proposed tests that score each area for the difference between observed and expected disease counts, weighted by degree of exposure to the focus:

$$U = \sum_{i=1}^N g_i (o_i - e_i) \quad (2)$$

Here there are  $N$  areas,  $g_i$  is a function defining the exposure to the focus,  $o_i$  is the observed number of cases in area  $i$ , and  $e_i$  is the expected number of cases in that area. A commonly used exposure function is the inverse distance to the focus ( $1/d$ ). The null hypothesis is no clustering relative to the focus, and the expected disease count thus is calculated as the Poisson expectation using the population at risk in each area and the assumption that disease risk is uniform over the study area.

Waller et al. (1992) used this test to explore whether cases of leukemia clustered near 12 hazardous waste sites in upstate New York that were injecting trichloro ethylene (TCE) into the groundwater. The Score test found some of the foci to be associated with high leukemia risk, and was significant after adjusting for the 12 repeated tests.

## 5 Tests for Cluster Association

Once clusters are identified they define feature sets that can be compared to the configuration of other feature sets. This is an exercise in pattern matching, rather than statistical pattern recognition. So, for example, one can ask whether the edges of disease clusters are near the edges of pollutant plumes. An ecologist might ask whether the spatial distribution of species abundance is associated with habitat patches, and so on. This kind of pattern matching task has been called the map comparison problem (Jacquez 1995), and has been addressed using methods of geographic boundary analysis (Jacquez et al. 2000). Two approaches will be discussed. The first quantifies boundary overlap, the second quantifies area overlap.

### 5.1 Boundary Overlap

Jacquez (1995) proposed four tests for the overlap of geographic boundaries. For ease of reference, we will term one set of boundaries boundary **G** and the other Boundary **H**. For example, **H** might correspond to the edges of clusters in a health related variable and **G** might be the cluster edges for a pollutant plume. The statistics are based on the nearest neighbor distances between *boundary elements* (BE's), which are those geographic coordinates that define the cluster boundary. The first statistic,  $O_s$ , is the count of the number of BE's that are included in both sets of boundaries, and is a measure of exact boundary overlap. The second statistic,  $O_G$ , is the mean minimum distance from the BE's in **G** to the nearest BE in **H**.  $O_H$  is the mean



minimum distance from the BE's in **H** to the nearest BE in **G**.  $O_{GH}$  is the mean distance from a BE in either boundary set to the nearest BE in the other boundary set. These statistics are useful for evaluating whether the edges of geographic features, such as zones of rapid change and clusters, are significantly near one another. These statistics thus evaluate spatial association and are a tool that can be used in conjunction with non-spatial tests for association such as correlation and regression.

## 5.2 Boundary Overlap Examples

Jacquez (1995) explored boundary overlap in respiratory illness and environmental ozone in southern Ontario. Exposure to high ozone can cause acute respiratory distress leading to pulmonary edema or even emphysema. Jacquez asked whether zones of rapid change in environmental ozone induced concomitant zones of rapid change in respiratory health. Ozone boundaries appeared by visualization to coincide with boundaries in hospital respiratory admissions; however, the overlap statistics were not significant. Most likely other factors were involved that may have obscured the relationship between ozone and respiratory health, and these results demonstrate the need to statistically evaluate apparent associations in order to avoid the “Gee Whiz” effect (Jacquez 1998).

Fortin et al. (1996) used boundary overlap to assess the relationships between edaphic factors (soil types and moisture) and vegetation boundaries. They found that vegetation boundaries based on species stem density and species presence/absence overlapped boundaries in edaphic factors, but vegetation boundaries based on species diversity and richness did not. This pattern suggests a hierarchy of effects, with edaphic factors predicting species presence but not plant community structure.

To determine how much the variable examined influences boundary delineation, Fortin (1997) evaluated overlap among vegetation boundaries calculated from different data sets. She found that density, percent coverage, and presence/absence for trees, shrubs, and trees and shrubs together significantly overlapped. While most variables concurred, the tree-only and the shrub-only data did not. This study demonstrated the use of boundary overlap analysis to distinguish variables that are spatially associated from those that are not.

Hall and Maruca (2001) compared vegetation boundaries to those in songbird abundance in a 45 ha swamp in Michigan. They found that bird abundance boundaries were significantly associated with vegetation boundaries, but not vice versa. Upon investigating the composition of the eight vegetation clusters, they found that the variable driving the vegetation clusters, and hence their boundaries, was the density of coniferous trees, a potentially important factor influencing the selection of songbird nesting and foraging areas. The authors suggested boundary analysis may aid in the development of monitoring and recovery plans for threatened bird species that use mosaic landscapes.

### **5.3 Cluster Overlap**

Maruca and Jacquez (2002) developed tests for identifying the amount of overlap between two spatial patterns. These tests differ from boundary overlap tests in that they focus on overlap of the areas enclosed within cluster boundaries. Recognizing the ubiquity of edge effects in natural systems and that spatial heterogeneity typically occurs on several spatial scales, they developed a test for association between two spatial patterns (e.g. sets of cluster calculated on two different variables) that is not biased by edge effects and is based on null spatial models that can incorporate spatial heterogeneity found in real-world systems. These methods can be used to

determine whether landscape classification maps have geographic partitions that overlap to a significant extent, and to determine whether spatial clusters defined on different variables (e.g. health outcomes and putative exposures) are significantly close to one another, existence of which is consistent with (but does not prove) a causal relationship.

Assume two sets of clusters, **I** and **J**, each comprised of  $N_I$  and  $N_J$  clusters and obtained as a spatial cluster analysis of different variables in the same geographic space. For cluster  $i$  in set **I** and cluster  $j$  in set **J**, *relative cluster overlap* is calculated as:

$$a_{ij} = \frac{a_{(i \cap j)}}{a_{(i \cup j)}} \quad (3)$$

Here  $a_{(i \cap j)}$  is the area of intersection and  $a_{(i \cup j)}$  is the area of union for clusters  $i$  and  $j$ . This statistic is zero for non-overlapping clusters, and increasing values represent better overlap, with a maximum value of 1 for perfectly overlapping clusters (where  $a_{(i \cap j)} = a_{(i \cup j)}$ ). For each cluster  $i$  in **I** we can then find the cluster in **J** that  $i$  overlaps best with, by finding the maximum value of  $a_{ij}$  over all clusters in **J** (called the *maximum relative cluster overlap*):

$$A_i = \max(a_{i \bullet}) \quad (4)$$

A cluster overlap statistic from the perspective of set **I** is the *average maximum relative cluster overlap* (Equation 5). This also can be calculated for set **J** (Equation 6), and a simultaneous area overlap statistic is shown in Equation 7:

$$A_I = \frac{\sum_{i=1}^{N_I} \max(A_{i \bullet})}{N_I} \quad (5)$$

$$A_J = \frac{\sum_{j=1}^{N_J} \max(A_{\bullet j})}{N_J} \quad (6)$$

$$A_{IJ} = \frac{\sum_{i=1}^{N_I} \max(A_{i\bullet}) + \sum_{j=1}^{N_J} \max(A_{\bullet j})}{N_I + N_J} \quad (7)$$

The statistic  $A_I$  is a measure of how well the clusters in **I** overlap the clusters in **J**, and  $A_J$  is a measure of how well clusters in **J** overlap with those in **I**.  $A_{IJ}$  is a general (or bi-directional) measure of overlap between the two sets of clusters. These statistics are best suited to instances where the two sets of clusters contain roughly the same number of clusters, and with clusters of about the same size. However, in the real world we expect to find cluster sets with very different numbers of clusters. Further, a given cluster set may contain clusters of drastically different sizes, as occurs when spatial variation is heterogeneous. The constraint on cluster number and size distribution is relaxed by calculating the *average maximum relative overlap* ( $A'_I$ ) as a weighted average, where the weight is the area of the focus cluster ( $a_i$  for cluster  $i$  in set **I**;  $a_j$  for cluster  $j$  in set **J**). In this scenario, the statistics would be calculated as follows:

$$A'_I = \frac{\sum_{i=1}^{N_I} [a_i \max(A_{i\bullet})]}{\sum_{i=1}^{N_I} a_i} \quad (8)$$

$$A'_J = \frac{\sum_{j=1}^{N_J} [a_j \max(A_{\bullet j})]}{\sum_{j=1}^{N_J} a_j} \quad (9)$$

$$A'_{IJ} = \frac{\sum_{i=1}^{N_I} [a_i \max(A_{i\bullet})] + \sum_{j=1}^{N_J} [a_j \max(A_{\bullet j})]}{\sum_{i=1}^{N_I} a_i + \sum_{j=1}^{N_J} a_j} \quad (10)$$

These tests have local versions, just as the global cluster tests discussed earlier have corresponding local tests. The local version identifies those locations on the map where statistically significant cluster overlap and overlap avoidance are found. Implementation of this involves decomposing the summation for either the raw (equations 5-7) or weighted versions (from equations 8-10) into local contributions for each cluster. Thus, for the global unweighted overlap statistic in equation (5) the local version is:

$$A_{ll} = \frac{\max(A_{l\bullet})}{N_l} \quad (11)$$

Here the “index” cluster is cluster  $l$ , and  $A_{ll}$  is the contribution of cluster  $l$  to the global overlap statistic  $A_l$ . Local counterparts can also be constructed for the other global statistics in equations (6-10). The value of each local overlap statistic can indicate overlap, no association, or overlap avoidance. A probability for overlap of cluster of the size of cluster  $l$  is calculated from the distribution of  $A_{ll}$  under a Monte Carlo simulation that conditions on both the number and size distribution of the clusters, but assumes no association between the geographies of the clusters in sets **I** and **J**.

## 6 Method Selection Advisors

So far the reader has been introduced to global, local and focused cluster statistics, as well as spatial association tests for boundary and area overlap. Researchers often ask “which spatial cluster test should I use”, and online cluster analysis advisors, such as those found at [http://www.terraser.com/bsr/boundaryseer\\_advisor.html](http://www.terraser.com/bsr/boundaryseer_advisor.html) and [http://www.terraser.com/csr/clusterseer\\_advisor.html](http://www.terraser.com/csr/clusterseer_advisor.html) can aid in this regard.

But looking for the “one” suitable cluster test is appropriate only when one has prior knowledge of cluster shape. For example, if one knows cancer clusters are circular then it would be appropriate to use a spatial scan statistic with a circular scanning window. However, this reasoning is also circular since one usually undertakes a cluster analysis in order to locate and describe the clusters – prior knowledge of cluster shape therefore is lacking.

## 7 Cluster Morphology

There is a growing awareness that clusters can take on a variety of different shapes, but cluster tests are usually sensitive to only one profile (Sun 2002, Smith 2003, Jacquez 2004, Tango and Takahashi 2005). Different techniques are sensitive to different aspects of *cluster morphology* – some detect boundaries, some detect outliers, some detect circular clusters, some detect elliptical clusters, and so on. Areas of high value can take many shapes, yet most cluster-detection techniques employ geographic “templates” such as circular scanning windows. These include the GAM (Openshaw et al. 1988), the scan statistic (Kulldorff and Nagarwalla 1995), Turnbull’s test (Turnbull et al. 1990), Besag and Newell’s (1991) test, the score test of Lawson and Waller that uses a circular risk function, first-order adjacencies such as LISA statistics, and nearest-neighbors relationships such as used in Cuzick and Edward’s (1990) test. These tests are most sensitive to cluster shapes that correspond to the geographic templates they employ. But spatial variation and hence cluster morphology in geographic systems is highly complex, and cannot be well described by a single geographic template or clustering approach.

This observation recently motivated the development of an integrative approach that provides a more complete description of cluster morphology (Jacquez and Greiling 2003a, b). This integrative framework employs a battery of ESDA techniques including geographic

boundary analysis, spatial agglomerative clustering, local Moran tests, and scan statistics (Table 1). Jacquez and Greiling employed it to more fully describe multi-scalar and multivariate patterns in the incidence of breast, colorectal and lung cancers on Long Island, New York. They used global, local and focused tests to explore the spatial scale of clustering, LISA statistics to identify spatial outliers, hot spots and cool spots, boundary analysis to find zones of rapid change, boundary overlap to evaluate possible associations between lung cancer and airborne carcinogens, and spatial agglomerative clustering to identify multivariate clusters that were homogeneous in lung, breast and colorectal cancer incidence (Figure 1). This integrative approach yielded a detailed description of the morphology of statistically significant geographic variation patterns for breast, lung and colorectal cancer incidence on Long Island.

**[Table 1 near here]**

**[Figure 1 near here]**

An alternative approach is to use techniques for which the geographic template is flexible and can assume any shape. The first method, called the Upper Level Set scan statistic (Patil and Taillie 2004), involves estimation of cluster morphology (e.g. shape, extent and configuration) from the data itself. The second method involves the “growth” of clusters by grouping adjacent areas that have similar (high or low) rates (see Urban 2004). While techniques for spatially agglomerative clustering have been available for some time (e.g. Legendre 1987), they often do not assign probabilities to the resulting clusters. A new approach called B-statistics was recently proposed that simultaneously detects agglomerative clusters of arbitrary shape as well as edges (borders where two adjacent areas of significantly different rates abut), and that provides cluster probabilities under realistic null hypotheses (Jacquez et al 2006). Finally, kernel density

estimation methods result in spatially continuous maps of the probability of a disease outcome (Rushton 1997; Rushton et al 2004) and appear capable of circumscribing clusters of variable shape, but the impact of kernel-based smoothing on the type I and type II error has yet to be fully quantified.

## 8 Cluster Change and Persistence

With the advent of routine remote sensing and improved environmental monitoring and health surveillance it now is possible to analyze data that are spatially and temporally referenced. In particular, there now are Space-Time Intelligence Systems (STIS) designed specifically to deal with georeferenced data through time. Analysis of how spatial patterns change through time is quite straightforward in such systems. We now consider two aspects of cluster change and persistence: temporal change in the spatial distribution of clusters, and clustering of attributes from two different time periods. In this discussion we use the LISA statistic, but the approach is general and can apply to most cluster tests.

- **Temporal change in the spatial distribution of clusters:** An obvious first-step in the exploration of cluster change and persistence is to cluster an attribute at time  $t$  and compare the spatial distribution of those clusters to those obtained for that attribute at time  $t+1$  (Figures 2 and 3). Boundary and area overlap statistics such as those summarized earlier may be used to determine the amount of association between the clusters at the different time periods. The local overlap statistics are used to distinguish those clusters that significantly overlap from those that do not. This approach is useful



for identifying where clusters existence changes through time and where clusters are persistent.

[Figure 2 near here]

[Figure 3 near here]

- **Clustering of attributes at two different time periods:** Bivariate LISA statistics are useful for identifying those areas with high values at time  $t$  that are surrounded by areas with high values at time  $t+1$ . This tool is useful for gaining insights into cluster persistence and spread (Figure 3).
- **Clustering temporal difference:** Rather than working with maps of the attribute at times  $t$  and  $t+1$ , one can first calculate difference maps that subtract the value at time  $t+1$  from the value at time  $t$ . These clusters (Figure 4) identify areas where the difference is high, and thus are useful for pinpointing those localities where the attribute value is uncertain, unstable and/or changing dramatically.

[Figure 4 near here]

### 8.1 Disparity Clusters

When faced with different classes of an item (say males and females) the question often arises as to whether spatial clusters of disparity exist. This is an important problem in the health sciences, as substantial disparities in disease incidence and mortality are observed for different race, gender and ethnic groups. Consider cancer. According to NCI's planning and budget proposal for 2004 (National Cancer Institute 2003):

“The unequal burden of cancer in our society is more than a scientific and medical challenge. It is a moral and ethical dilemma for our Nation. Certain

populations experience significant disparities in cancer incidence, the care they receive, and the outcomes of their disease. These differences have been recognized, or at least suspected, for some time. They now are being documented with increasing frequency and clarity.”

The identification of locations of high disparity in a health outcome – a disparity cluster – is an important step that allows one to target interventions and to address inequities in access to health care and provision of screening services. How can health disparities be identified?

The approach employed involves three steps similar to those employed for evaluating cluster change and persistence. Comparison of cluster maps, bivariate cluster analysis, and clustering of difference maps.

Consider an example. Pancreatic cancer incidence and mortality have changed little over the last three decades. Mortality rates in this period have been relatively stable for black and white males, have decreased for white females, and increased for black females. In each racial/ethnic group males have higher incidence and mortality than women. Blacks have incidence and mortality rates nearly 50% higher than whites. Rates for native Hawaiians are higher than for whites, while Hispanic and Asian-American rates are lower (Miller et al. 1996). Risk is highest in the older population, and pancreatic cancer is rare among those 30-54 years old. Incidence for blacks 55-69 years of age is 60% higher than for whites of the same age, although this difference diminishes in ages 70 years and older. Age-based racial mortality patterns are similar to those observed in the incidence rates. Does the disparity in cancer mortality between white and black males cluster geographically?

The difference in standardized rates can be used to create cancer mortality disparity maps (e.g. for BM-WM), and clusters of high values on these maps (hot spots) then identify locations

of elevated mortality for black males. In addition to univariate clustering of the difference, bivariate LISA's for detecting clusters and anomalies of disparities in cancer mortality can be constructed of the form:

$$L_{i,BM \times WM} = z_{i,BM} \sum_j w_{ij} z_{j,WM} \quad (12)$$

Here  $L_{i,BM \times WM}$  is the bivariate LISA at location  $i$  for the disparity in pancreatic cancer mortality between black and white males at the local spatial scale.  $z_{i,BM}$  is the standardized mortality rate for black males at that location and  $z_{j,WM}$  is the local component for white males at location  $j$  adjacent to  $i$ . The term  $\sum_j w_{ij} z_{j,WM}$  is the average white male pancreatic cancer mortality for locations (e.g. counties) adjacent to  $i$ . The disparity statistic  $L_{i,BM \times WM}$  is positive when pancreatic cancer mortality at neighboring locations is similar for both races, and negative if there is a disparity. Significance of the statistic is evaluated under conditional randomization, and the Moran scatter plot identifies clusters and hotspots of racial disparity in pancreatic cancer mortality. P-values under randomization are then used to construct pancreas cancer mortality disparity maps (Figure 5).

**[Figure 5 near here]**

The two approaches (univariate on difference maps and bivariate on standardized rates) inform two different aspects of the geography of disparity. The univariate difference clusters detect significant spatial clusters and outliers in the difference between standardized rates (e.g. BM-WM). The bivariate LISA statistic identifies spatial clusters and outliers in the standardized rates for one race-gender combination (e.g. BM) relative to the average of the standardized rates for the second race-gender combination (e.g. WM) in surrounding areas. Spatial cluster analysis

has obvious utility for geographically pinpointing locations of statistically significant disparities in pancreatic cancer mortality.

An alternative approach to evaluating statistical significance of health disparities was recently proposed by Goovaerts (2005), who recognized that tests for differences in means, such as that based on Student's *t*-distribution, would be useful for detecting health disparities provided differences in population sizes could be accounted for. His disparity statistic is an adaptation of the classical test for inference of two population proportions (Devore, 2000) to the comparison of rates measured in two sub-populations, labeled as reference and target populations. For a given region, the disparity statistic is calculated as the standardized difference between the target and reference rates, weighted by the population proportions, and has been demonstrated to detect regions with statistically significant health disparities that account for the population sizes of the reference and target populations (see Goovaerts 2005).

## **9 Multiple Testing**

ESDA and the description of cluster morphology may involve iterations of visualization and statistical analyses to elucidate different aspects of spatial patterns and to successively refine the alternative hypotheses explored in pattern recognition procedures. Hence a typical analysis, say of prostate cancer incidence, may begin with the creation of maps using appropriately adjusted rates (e.g. to stabilize the rates and to adjust for covariates such as age), and may then involve the use of global, local and focused tests to determine whether the rates are spatially autocorrelated (using global tests such as those of Moran 1950, Oden 1995, Tango 1995 and others), to identify the locations of cold-spots, hot-spots and outliers (using local tests such as Anselin 1995, Getis and Ord 1992, Turnbull et al. 1990, Besag and Newell 1991, among others), and to assess

whether cases tend to cluster near the locations of putative environmental exposures (using focused tests such as Lawson and Waller 1996). To maintain statistical rigor the impact of repeated statistical procedures must be accounted for (Jacquez et al. 1996b), and this may be accomplished within the structure of the test itself (as in Kulldorff's (1997) scan test) or by adjusting P-values or Type I errors using the methods of Bonferroni (Sidak 1967, Simes 1986), Holms (Holland and Copenhaver 1987) or Hochberg (1988). Recently, Tango (2006) proposed using a "min-P" approach in which the test statistic itself is the minimum p-value observed from a group of tests. Because of the exploratory nature of the analyses there is some question as to whether a formal approach to inferential statistics (e.g. comparing a P-value to the alpha level to determine whether the null hypothesis is rejected) is applicable. Most experts now advocate interpretation of P-values within the context of other information, such as the biological plausibility of the cluster, the quality of the data, and the costs associated with false positives and negatives (Waller and Jacquez 1995; Jacquez et al. 1996 a, b). Some software packages such as ClusterSeer (Jacquez et al. 2001, 2002) account for multiple tests automatically and provide appropriately adjusted probability values.

## 10 Tools

*Information Frames.* There are literally dozens of spatial cluster tests, and cogent summaries of the different tests are needed to support method selection and to remind one of the properties of a given test. In a recent study funded by the National Cancer Institute, researchers at BioMedware, Inc. and the University of Michigan School of Public Health developed one-page information frames that give a quick overview of the properties of a test (Figure 6).

**[Figure 6 near here]**

*Software.* Cluster analysis software includes Satscan, which implements a scan-type statistic employing circular scanning windows. The commercial software ClusterSeer (Jacquez et al. 2001, 2002) has dozens of statistical techniques that employ a variety of geographic templates (e.g. circular scanning windows, nearest neighbor relationships, LISA tests, global tests and focused tests). It comes with a cluster advisor and adjusts for multiple testing. BoundarySeer (Jacquez et al. 2001b) employs techniques to detect the edges of clusters, and these clusters can be of any shape. Both univariate and multivariate methods are included. CancerAtlas viewer (<http://www.terraseer.com/atlasviewer.html>) comes with county, State Economic Area, and State level mortality data for 43 site specific cancers, and employs LISA statistics. TerraSeer's Space Time Intelligence System (<http://www.terraseer.com/products/stis.html>) supports linked windows, statistical brushing, spatial and space-time cluster statistics, animation as well as spatio-temporal georeferencing.

## **11 Conclusions**

This chapter provided a quick overview of some of the issues and approaches in spatial cluster analysis. The reader should now have some appreciation of the role spatial cluster analysis plays in ESDA, and of global, local and focused techniques. It should now be apparent that the “one size fits all” approach to cluster analysis yields an incomplete picture of cluster morphology. The integrated approach conveys a far more complete quantification of cluster morphology descriptors and ultimately leads to a better understanding of spatial variation. Spatial cluster analysis can yield substantial benefits in documenting cluster change and persistence, and for identifying disparities in two spatially referenced variables. The field is evolving rapidly, and as

the volume of spatially and temporally referenced data increases cluster analysis will play an increasing role in pattern recognition and data reduction.

### **Acknowledgements**

Some of the findings reported in this publication were developed under grants CA92669 from the National Cancer Institute and ES10749 from the National Institute of Environmental and Health Sciences. The perspectives of this publication are solely those of the author and do not necessarily represent the official views of the funding organizations.



## References

- Anselin L 1995 Local indicators of spatial association: LISA. *Geographical Analysis* 27: 93-115
- Bailey T C and Gatrell A C 1995 *Interactive Spatial Data Analysis*. Harlow, Addison Wesley Longman
- Besag J and Newell J 1991 The detection of clusters in rare diseases. *Journal of the Royal Statistical Society Series A* 154: 143-55
- Cuzick J and Edwards R 1990 Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society Series B* 52: 73-104
- Devore, JL 2000 Probability and Statistics for Engineering and the Sciences. Duxbury Press.
- Fisher R A 1935 *The Design of Experiment First Edition*. London, Oliver and Boyd
- Fortin M J 1997 Effects of data types on vegetation boundary delineation. *Canadian Journal of Forest Research* 27: 1851-8
- Fortin M J, Drapeau P, and Jacquez G M 1996 Statistics to assess spatial relationships between ecological boundaries. *Oikos* 77: 51-60
- Gelman A, Price P N, and Lin C 2000 A method for quantifying artefacts in mapping methods illustrated by application to headbanging. *Statistics in Medicine* 19: 2309-20
- Goovaerts, P 2005 Analysis and detection of health disparities using geostatistics and a space-time information system: The case of prostate cancer mortality in the United States, 1970-1994. In Proceedings of GIS Planet 2005, Estoril, May 30-June 2.
- Goovaerts, P and G. Jacquez. 2004 Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *International Journal of Health Geographics* 3:14.

- Goovaerts, P and G. M. Jacquez 2005 Detection of temporal changes in the spatial distribution of cancer rates using LISA statistics and geostatistically simulated spatial neutral models *Journal of Geographical Systems* 7:137-159.
- Gustafson E J 1998 Quantifying landscape spatial pattern: What is the state of the art? *Ecosystems* 1: 143-56
- Hall K R and Maruca S L 2001 Mapping a forest mosaic: A comparison of vegetation and bird distributions using geographic boundary analysis. *Plant Ecology* 156: 105-20
- Hochberg Y 1988 A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75: 800-2
- Holland B S and Copenhaver M D 1987 An improved sequentially rejective Bonferroni test procedure. *Biometrics* 43: 417-23
- Jacquez G M 1995 The map comparison problem: Tests for the overlap of geographic boundaries. *Statistics in Medicine* 14: 2343-61
- Jacquez G M 1998 GIS as an enabling technology. In Gatrell A and Loytonen M (eds) *GIS and Health*. London, Taylor and Francis: 17-28
- Jacquez G M 2000 Spatial epidemiology: Nascent science or a failure of GIS? *Journal of Geographical Systems* 2: 91-7
- Jacquez, G M 2004 Current practices in the spatial analysis of cancer: flies in the ointment. *International Journal of Health Geographics* 3:22.
- Jacquez G M, Grimson R, Waller L, and Wartenberg D 1996a The analysis of disease clusters: Part 2, Introduction to techniques. *Infection Control and Hospital Epidemiology* 17: 385-97
- Jacquez G M, Waller L, Grimson R, and Wartenberg D 1996b The analysis of disease clusters: Part I, State of the art. *Infection Control and Hospital Epidemiology* 17: 319-27

- Jacquez G M, Maruca S L, and Fortin M J 2000 From fields to objects: A review of geographic boundary analysis. *Journal of Geographical Systems* 2: 221-41
- Jacquez G M, Greiling D, Estberg L, Do E, Long A, and Rommel B 2001 *ClusterSeer User Guide: Software for Identifying Disease Clusters*. Ann Arbor, MI, TerraSeer Press
- Jacquez G M, Maruca S L, Greiling D A, Kaufmann A, Muller L, Rommel B, Sengupta S, Agarwal P, and Hall K 2001 *BoundarySeer User Guide: Software for Geographic Boundary Analysis*. Ann Arbor, MI, TerraSeer Press
- Jacquez G M, Greiling D A, Durbeck H, Estberg L, Do E, Long A, and Rommel B 2002 *ClusterSeer User Guide 2: Software for Identifying Disease Clusters*. Ann Arbor, MI, TerraSeer Press
- Jacquez G M and Greiling D A 2003a Local clustering in breast, lung, and colorectal cancers in Long Island, New York. 2: *International Journal of Health Geographics* 2: 3 (available at <http://www.ij-healthgeographics.com/content/2/1/3>)
- Jacquez G M and Greiling D A 2003b Geographic boundaries in breast, lung, and colorectal cancers in relation to exposure to air toxics in Long Island, New York. 2: *International Journal of Health Geographics* 2: 4 (available at <http://www.ij-healthgeographics.com/content/2/1/4>)
- Jacquez, GM, Kaufmann A and Goovaerts P 2006 Boundaries, ladders and clusters: A new paradigm in spatial analysis? *Environmental and Ecological Statistics* (In Press)
- Kafidar K 1996 Smoothing geographical data, particularly rates of disease. *Statistics in Medicine* 15: 2539-60
- Kulldorff M and 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-

- Lawson A B 1989 *Score Tests for Detection of Spatial Trend in Morbidity Data*. Dundee, Dundee Institute of Technology
- Lawson A B and Kulldorff M 1999 A review of cluster detection methods. In Lawson A B, Biggeri A, Böhning D, Lesaffre E, Viel J-F, and Bertollin R (eds) *Advanced Methods of Disease Mapping and Risk Assessment for Public Health Decision Making*. London, John Wiley and Sons: 99-110
- Lawson A B and Waller L A 1996 A review of point pattern methods for spatial modelling of events around sources of pollution. *Environmetrics* 7: 471-87
- Legendre, P 1987 Developments in Numerical Ecology. P. Legendre and L. Legendre, Eds. Springer-Verlag, Berlin. pp 289-307.
- Liebisch N, Jacquez G M, Goovaerts P, and Kaufmann A 2002 New methods to generate neutral images for spatial pattern recognition. In Egenhofer M J and Mark D M (eds) *GIScience2002: The Second International Conference on Geographic Information Science*. Berlin, Springer-Verlag Lecture Notes in Computer Science No. 2478: 181-95
- Maruca S L and Jacquez G M 2002 Area-based tests for association between spatial patterns. *Journal of Geographical Systems* 4: 69-84
- Miller B A, Kolonel L N, Bernstein L, Young Jr. J L, Swanson G M, West D, Key, C R, Liff J M, Glover C S, Alexander G A, Coyle L, Hankey B F, Gloeckler Ries L A, Kosary C L, Harras A, Percy C, and Edwards B K 1996 *Racial/Ethnic Patterns of Cancer in the United States 1988-1992*. Bethesda, MD, National Cancer Institute Publication No. 96-4104
- Moore D A and Carpenter T E 1999 Spatial analytical methods and Geographic Information Systems: Use in health research and epidemiology. *Epidemiologic Reviews* 21: 143-61
- Moran P A P 1950 Notes on continuous stochastic phenomena. *Biometrika* 37: 17-23

- National Cancer Institute 2003 *The National Cancer Institute's (NCI) Planning and Budget Proposal for Fiscal Year 2004: The Nation's Investment in Cancer Research*. WWW document, <http://plan.cancer.gov/discovery/index.html>
- Oden N 1995 Adjusting Moran's I for population density. *Statistics in Medicine* 14: 17-26
- Openshaw S, Charlton M, Craft A W, and Birch J M 1988 Investigation of leukaemia clusters by use of a geographical analysis machine. *Lancet* 1: 272-3
- Ord J K and Getis A 1995 Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27: 286-306
- Pandya A S and Macey R B 1996 *Pattern Recognition with Neural Networks in C++*. Boca Raton, FL, CRC Press
- Patil, GP and Taillie, C 2004 Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* 11:183-97
- Rushton G 1997 Improving public health through geographical information systems: an instructional guide to major concepts and their implementation [CD-ROM]. Version 2.0. Iowa City: University of Iowa, Department of Geography; 1997 Dec. Available at: URL:<http://www.uiowa.edu/~geog/>
- Rushton G and Elmes G 2000 Considerations for improving Geographic Information System research in public health. *Journal of the Urban and Regional Information Systems Association* 12: 31-49
- Rushton, G., Peleg I, Banerjee A, Smith G, and West M 2004 Analyzing geographic patterns of disease incidence: Rates of late-stage colorectal cancer in Iowa. *Journal of Medical Systems* 28: 223-236
- Sidak Z 1967 Rectangular confidence regions for the means of multivariate normal

distributions. *Journal of the American Statistical Association* 62: 626 -33

Simes R J 1986 An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73: 751-4

Smith, G. H. 2003. Disease Cluster Detection Methods: The Impact of Choice of Shape on the Power of Statistical Tests. Department of Geography University of Iowa unpublished report. Email: geoffrey-smith@uiowa.edu

Sun, Y. 2002. "Determining the size of spatial clusters in focused tests: Comparing two methods by means of simulation in a GIS". *Journal of Geographical Systems* 4:359-370.

Tango T 1995 A class of tests for detecting "general" and "focused clustering of rare diseases. *Statistics in Medicine* 14: 2323-34

Tango T, and Takahashi K 2005 A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 4:

Tango T 2006 A test with minimized p-value for spatial clustering applicable to case-control point data. *Biometrics* (In Press)

Turnbull B W, Iwano E J, Burnett W S, Howe H L, and Clark L C 1990 Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. *American Journal of Epidemiology* 132: S136-43.

Urban, DL. 2004. Multivariate Analysis. Nonhierarchical Agglomeration. Spatially Constrained Classification. [http://www.env.duke.edu/landscape/classes/env358/mv\\_pooling.pdf](http://www.env.duke.edu/landscape/classes/env358/mv_pooling.pdf)

Waller L A and Jacquez G M 1995 Disease models implicit in statistical tests of disease clustering. *Epidemiology* 6: 584-90

Waller L A, Turnbull B W, Clark L C, and Nasca P 1992 Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and TCE-

contaminated dumpsites in upstate New York. *Environmetrics* 3: 281-300

### List of Figure Captions

**Figure 1** Cluster morphology of cancer incidence on Long Island. Top: LISA clusters in male lung cancer incidence; Middle: LISA clusters in female lung cancer incidence; Bottom: Boundaries in male and female lung cancer incidence and air toxics from EPA's National Air Toxics Assessment program (from Jacquez and Greiling 2003 a, b)

**Figure 2** Cluster change and persistence. Changes in pancreatic cancer mortality along the lower Mississippi river for white males for all ages from 1950-69 to 1970-94. Top: Mortality 1950-69; Middle: Mortality 1970-94; Bottom: Difference in mortality rates for the two time periods (cancer data from CancerAtlas Viewer; <http://www.terraseer.com/atlasviewer.html>)

**Figure 3** Cluster change and persistence continued. LISA clusters in pancreatic cancer for 1950-69 (top) and 1970-94 (middle). High pancreatic cancer mortality is spreading north in counties along the Mississippi River. Bivariate LISA clusters identify counties high in pancreatic cancer in 1950-69 that are surrounded by counties with high mortality in 1970-94. Analyses conducted in TerraSeer's STIS software

**Figure 4** Cluster change and persistence continued. Difference maps (top) quantify the difference in mortality rates between 1950-69 and 1970-94. LISA clusters of the difference maps (bottom) identify those localities where the change in mortality is high. Analyses conducted in TerraSeer's STIS software

**Figure 5** Disparity clusters

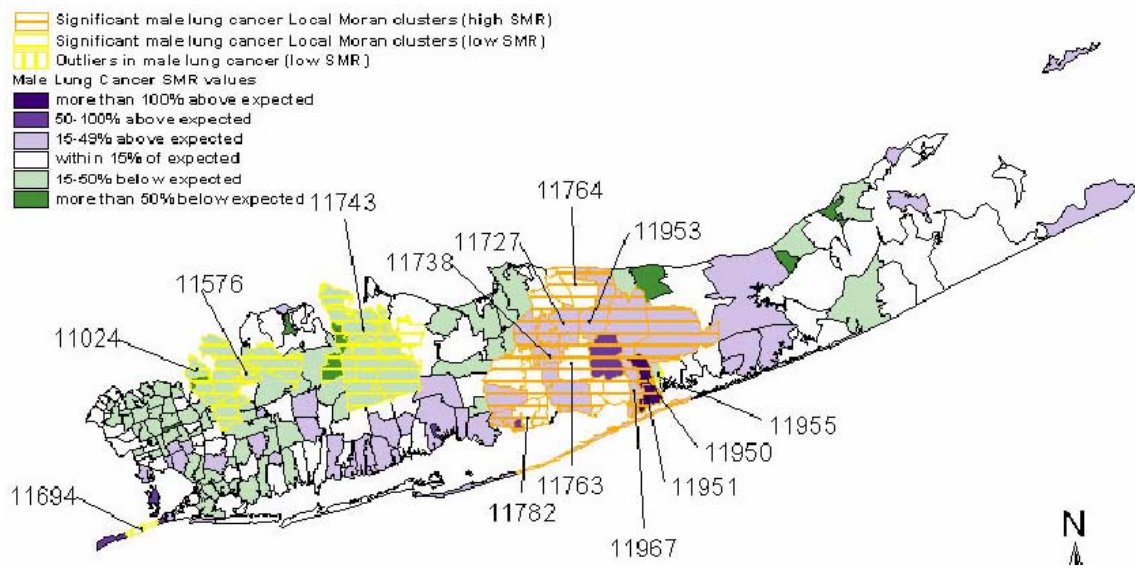
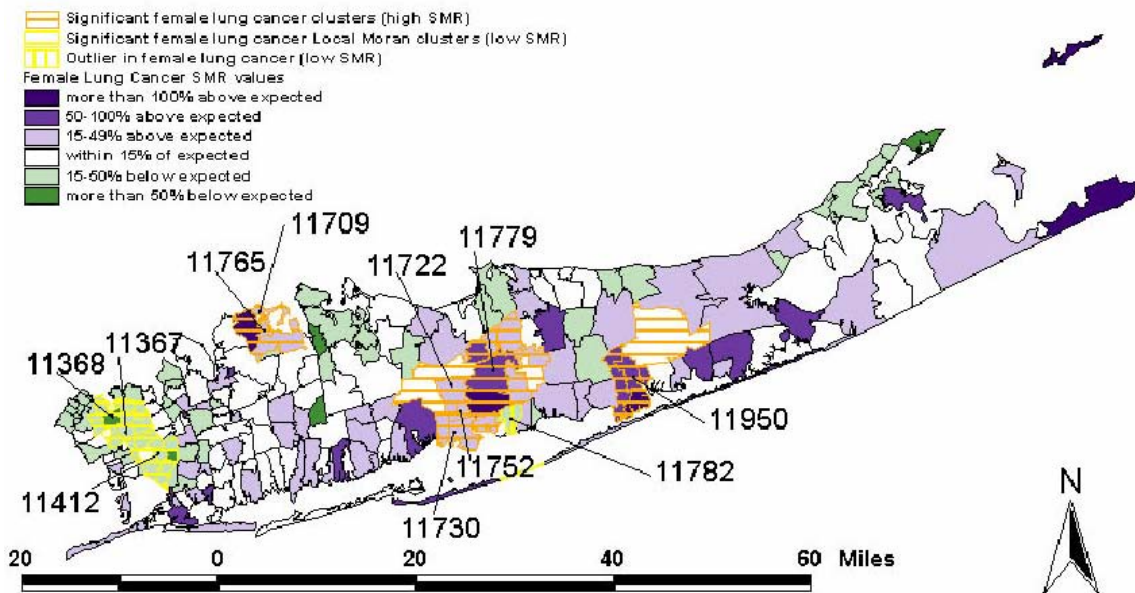
**Figure 6** Information frames provide cogent summaries of the properties of spatial cluster statistics, and are available through the cluster advisor found at <http://zappa.nku.edu/~longa/geomed/stathelp/advisor.html>



**Table 1** Cluster Morphology Descriptors.

<b>Descriptor</b>	<b>Example</b>
Amount of excess or deficit	Relative Risk in a disease cluster, number of cases in the cluster
Extent	Geographic area, number of sub-areas in the cluster
Length	Length of major and minor axes in an elliptical cluster
Boundary	
Length	Length of cluster boundary
Crenellation	Boundary fractal dimension
Fuzziness	Alpha core and surrounding zone of uncertainty
Shape	Ratio of boundary length / cluster area
Bivariate spatial association	Boundary overlap analysis; Cluster overlap analysis
Multivariate spatial structure	Clusters from multivariate spatially agglomerative clustering; boundaries from multivariate boundary analysis

Figure 1.



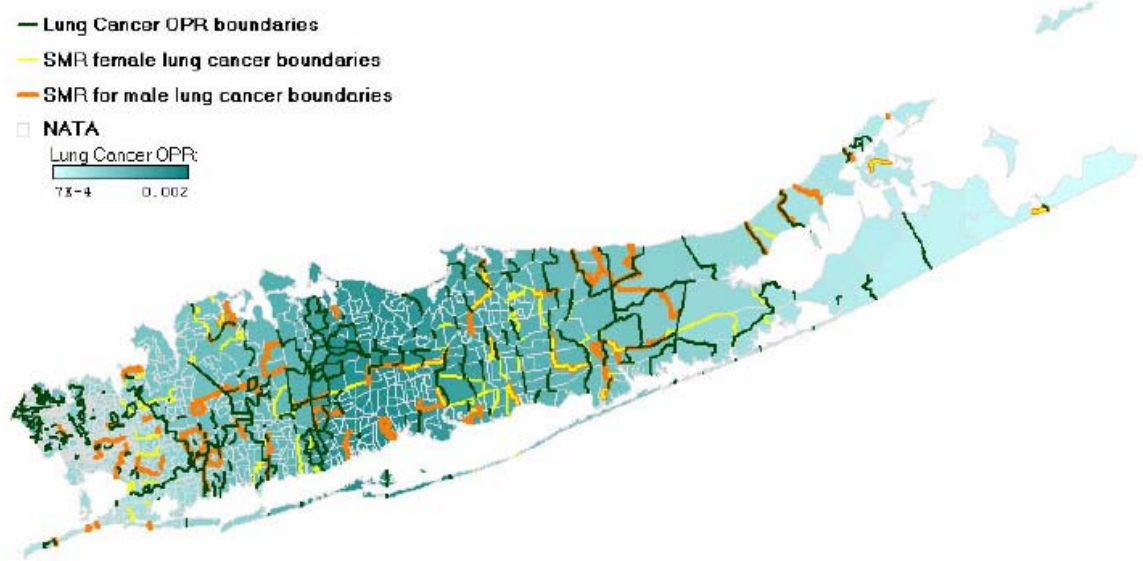


Figure 2.

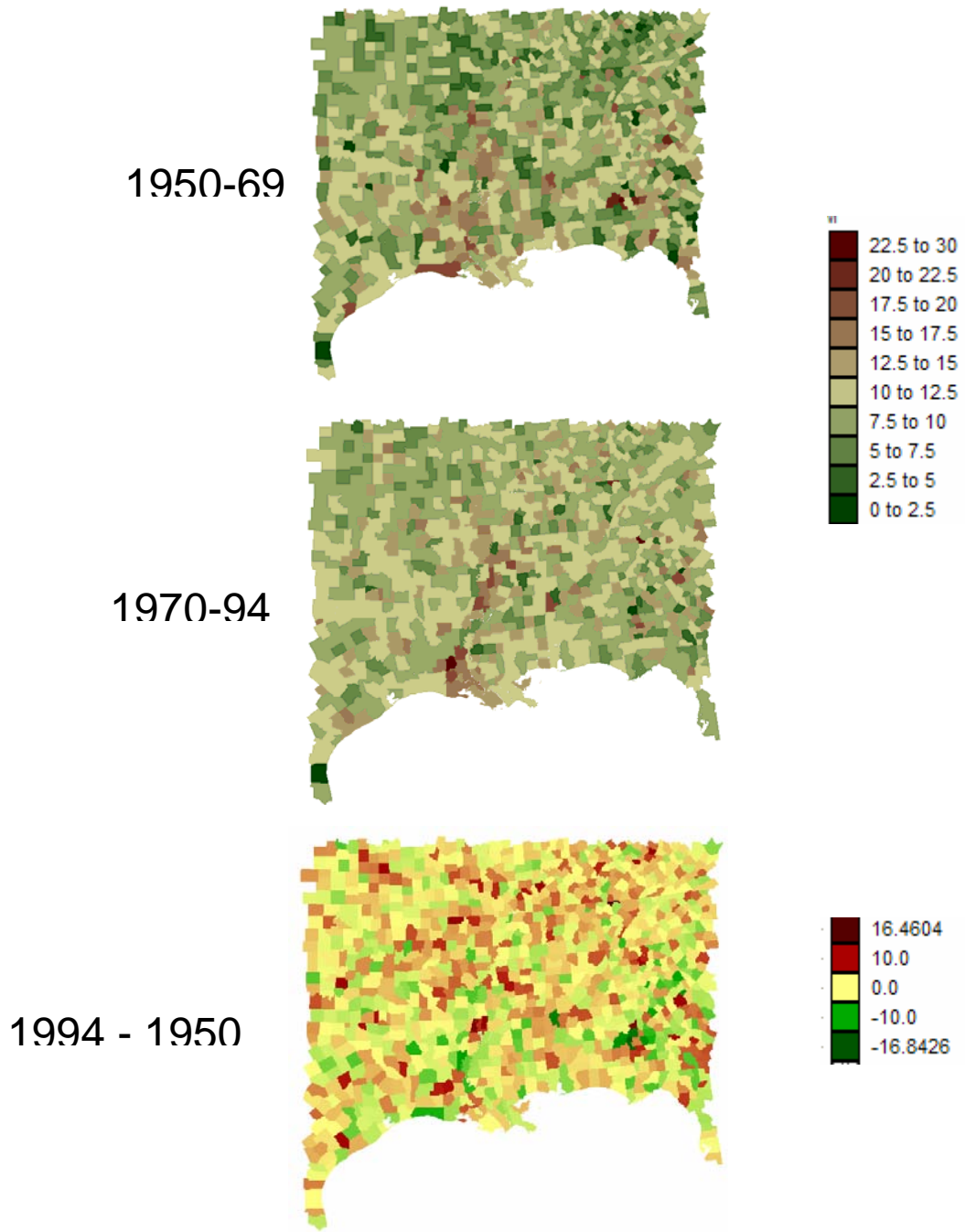


Figure 3.

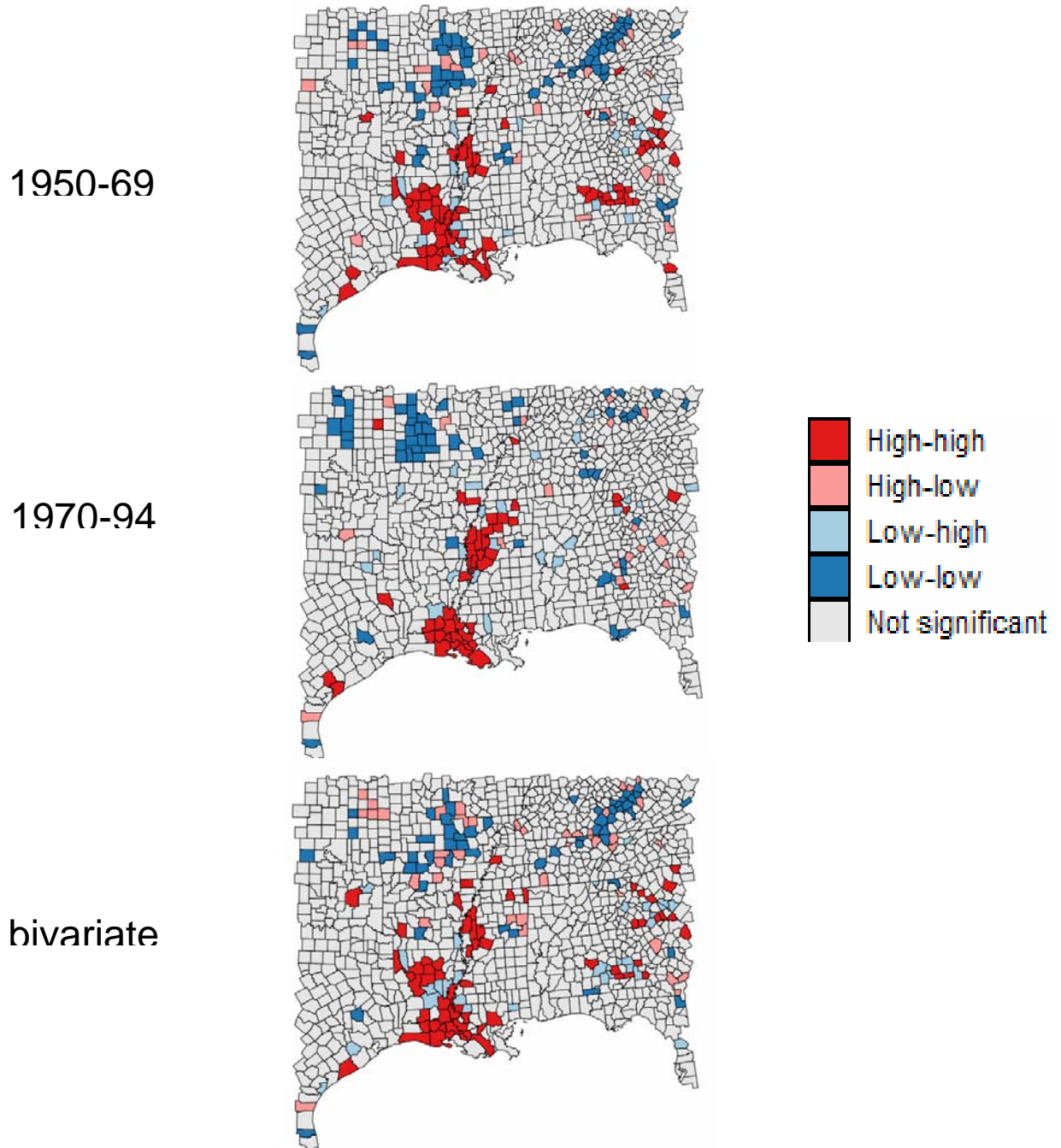


Figure 4.

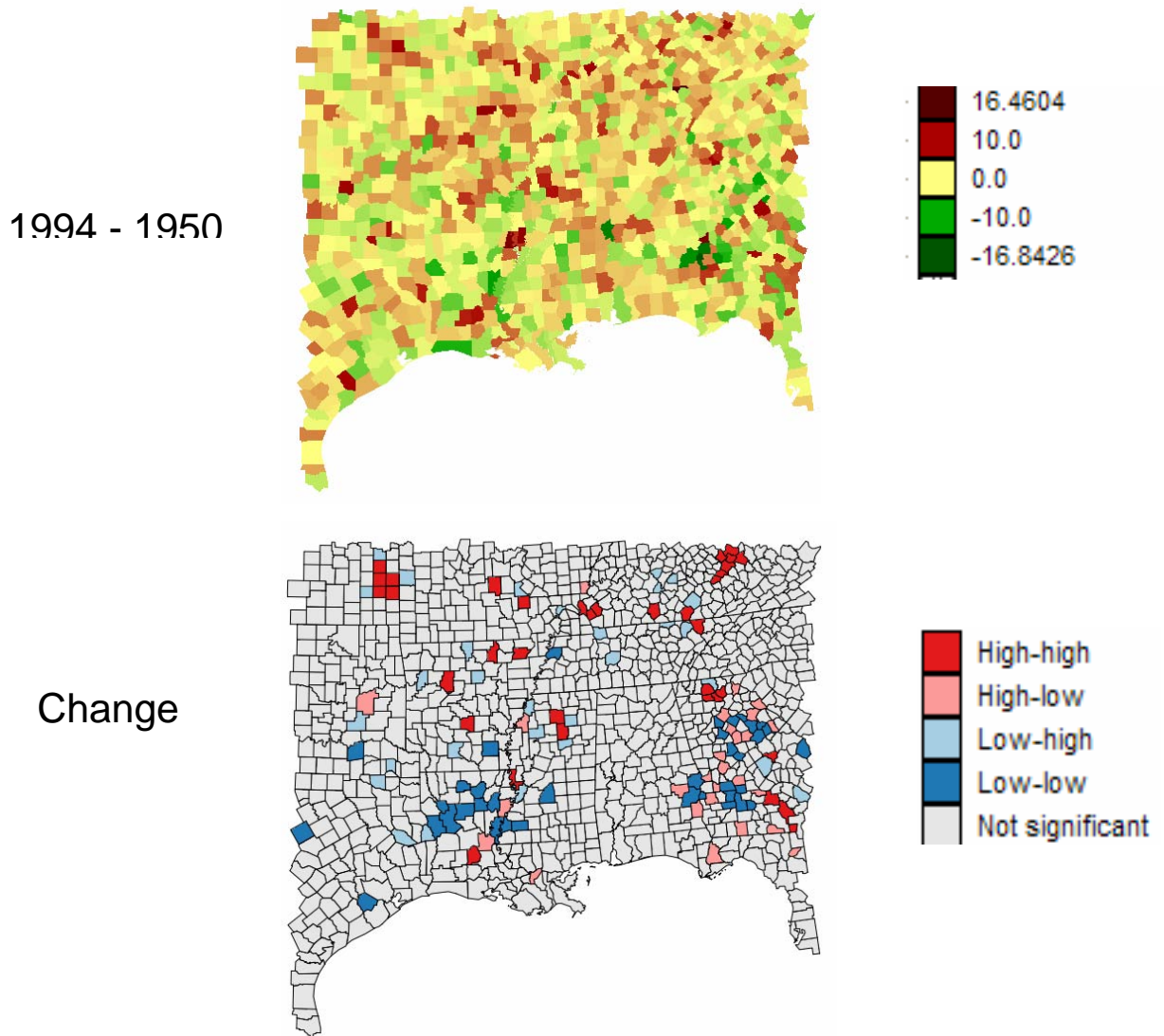


Figure 5.

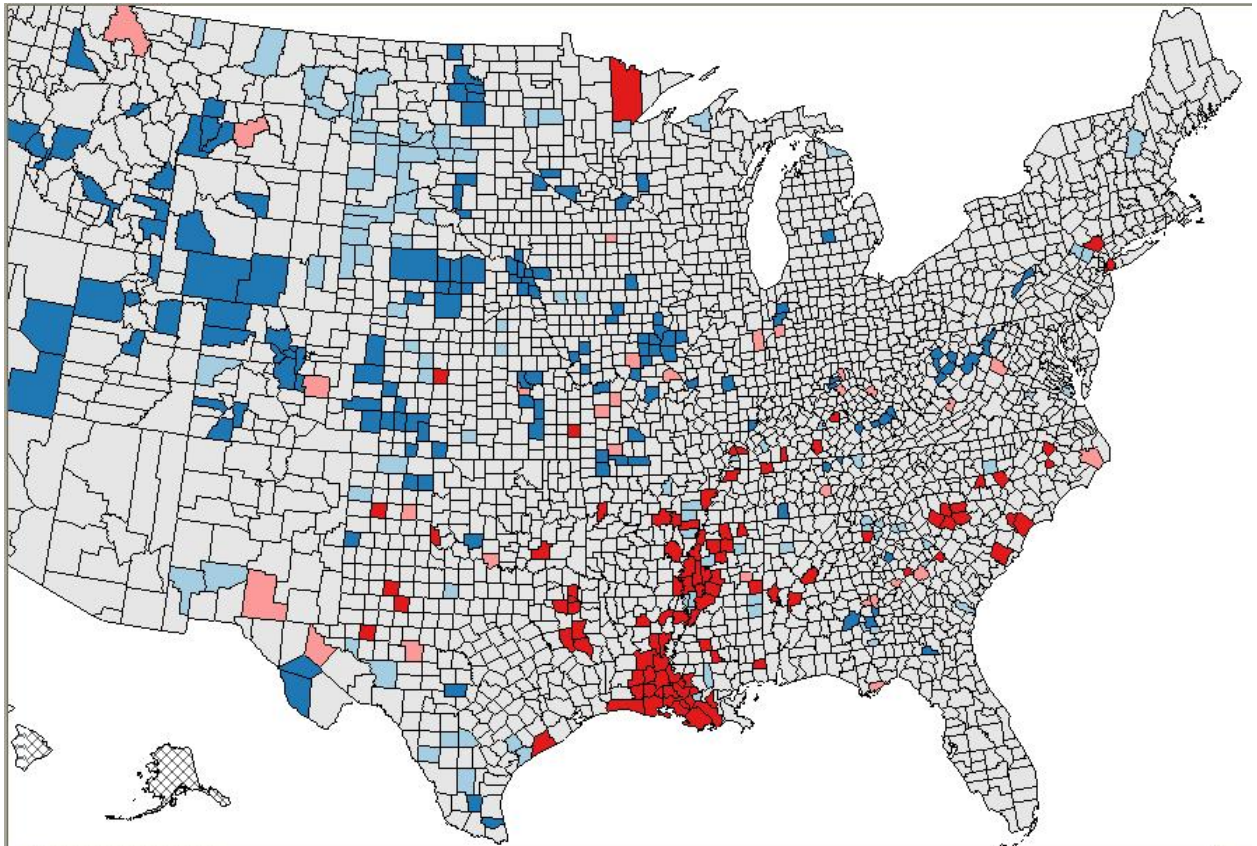




Figure 6

<b>k-Nearest Neighbor test</b>	<ul style="list-style-type: none"> <li>• <a href="#">spatial/temporal</a></li> <li>• <a href="#">global</a></li> <li>• <a href="#">region-based</a></li> </ul>	<ul style="list-style-type: none"> <li>• Similar Tests: <ul style="list-style-type: none"> <li>○ <a href="#">Mantel</a></li> <li>○ <a href="#">Knox</a></li> <li>○ <a href="#">Direction</a></li> <li>○ <a href="#">Grimson</a></li> </ul> </li> <li>• <a href="#">All available tests</a></li> </ul>
<p><b>Indications/Recommendations for use:</b> When space-time interaction is present nearby cases will occur at about the same time, (space nearest neighbors tend to be time nearest neighbors), and the test statistic will be large (<a href="#">more</a>).</p>		
<p><b>Description:</b> A <math>k</math>-nearest neighbor method used to detect space-time interaction.</p>		
<p><b>Test statistic:</b> The number of pairs of cases that are <math>k</math>-nearest neighbors in both space and time.</p>	$J_k = \sum_{i=1}^N \sum_{j=1}^N s_{ijk} t_{ijk}$	
<p><b>Null Hypothesis:</b> Whether cases are nearest neighbors in space is independent of whether they are nearest neighbors in time.</p>		
<p><b>Alternative Hypothesis:</b> Nearest neighbors in space tend to be nearest neighbors in time.</p>		
<p><b>GeoMed Inputs:</b> Space and time distances between pairs of cases.</p>		
<p><b>GeoMed Outputs:</b> Results table includes:</p> <ul style="list-style-type: none"> <li>○ <math>k</math>, the number of nearest neighbors being considered</li> <li>○ <math>J_k</math>, the number of space-time <math>k</math> nearest neighbors</li> <li>○ significance of <math>J_k</math></li> <li>○ <math>\Delta J_k</math>, the number of space-time nearest neighbors added when <math>k</math> increased from <math>k-1</math></li> <li>○ significance of <math>\Delta J_k</math></li> <li>○ p-values for the probability, under <math>H_0</math>, of observing a statistic as large or larger than that observed and combined for the <math>k</math> tests in 3 ways: <ul style="list-style-type: none"> <li>■ Bonferroni</li> <li>■ Simes</li> <li>■ Statistical Distance</li> </ul> </li> <li>○ A linkage map <ul style="list-style-type: none"> <li>■ case locations are mapped</li> <li>■ <math>k</math>-order linkages displayed as selected by user</li> </ul> </li> </ul>		
<p><a href="#">Example Analysis</a></p>	<p><b>Reference:</b></p>	