16
CHAPTER

Methods for Detecting Spatial and Spatio-Temporal Clusters

Daniel B. Neill and Andrew W. Moore
School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania

1. INTRODUCTION

This chapter focuses on the detection of spatial clusters of disease, with the goal of rapidly detecting emerging outbreaks by prospective surveillance. Spatial cluster detection has two main goals: identifying the locations, shapes and sizes of potential disease clusters, and determining whether each of these potential clusters is due to a genuine outbreak or due to chance fluctuations in case counts. In other words, we want to know whether anything unexpected is going on, and if so, where? This task can be broken down into two parts: first, figuring out what we expect to see, based on populations or on expected counts inferred from historical data, and then determining which regions deviate significantly from our expectations. In this chapter, we present an overview of spatial cluster detection, and then discuss a number of cluster detection methods, focusing on the spatial scan statistic. In addition to presenting the standard spatial scan framework, we consider a number of extensions to this framework, including generalization of the scan statistic to situations in which baselines must be inferred from historical data, computational methods for fast spatial scanning, and extensions to spatio-temporal cluster detection. This chapter does three things. First, it provides the basic statistical and computational tools to make the spatial scan applicable and useful for analyzing large real-world data sets. Second, it motivates cluster detection approaches and third, it compares them to other outbreak detection methods.

Epidemiologists have been analyzing biosurveillance data spatially since the seminal work of John Snow on cholera (Snow, 1855). In the 1970s, researchers automated the map creation aspect of spatial analysis. The results of this work—geographic information systems—are now in widespread use in health departments. Over the past decade, researchers have developed spatial scan statistics, which automate the pattern recognition component of spatial analysis. This advance enables a biosurveillance organization to analyze spatial data far more exhaustively than ever before.

2. OVERVIEW OF SPATIAL CLUSTER DETECTION

In this chapter, we focus on the task of *spatial cluster detection*: finding spatial areas where some monitored quantity is significantly higher than expected. For example, we may want to monitor the observed number of cases of influenza, or some other specific type of disease, and find any regions where the number of cases is abnormally high. The spatial cluster detection techniques that we describe are disease independent; that is, they are capable of detecting clusters of any type of disease including those of previously unknown diseases.

At present, health departments are using automatic spatial cluster detection primarily on syndromic data with a goal of detecting regions in a city or even a country with abnormally high case counts of some syndrome (e.g., respiratory), based on observed quantities such as the number of emergency department visits or sales of over-the-counter cough and cold medication. The detected clusters of disease may be indicative of a naturally occurring outbreak, a bioterrorist attack (e.g., anthrax release) or an environmental hazard.

The main goals of spatial scanning are to identify the locations, shapes, and sizes of potential clusters (i.e., pinpointing those areas which are most relevant), and to determine whether each of these potential clusters is more likely to be a "true" cluster (requiring further investigation by public health officials) or simply a chance occurrence (which can safely be ignored).

As mentioned previously, the spatial cluster detection task involves the two questions: Is anything unexpected going on? If so, where? In order to answer these questions, we must first have some idea of what we expect to see. We typically take one of two approaches, illustrated in Figure 16.1. In the *population-based* approach, we estimate an at-risk population for each area (e.g., zip code); this population can either be estimated simply from census data, or can be adjusted for a variety of covariates (patients' age and gender, seasonal and day of week effects, etc.). Then the expected number of cases in an area is assumed to be proportional to its at-risk population, and thus clusters are areas where the disease rate (number of cases per unit population) is significantly higher inside the region than outside. The *expectation-based* approach directly estimates the number of cases we expect to see in each area; typically by fitting a model based on past data (e.g., the number of cases in each area on each previous day). Wagner et al. (2003) describes such an approach. Using an expectation-based method, clusters are areas where the number of cases is significantly greater than its expectation. One important difference between these two approaches is their response to a global increase (i.e., where the number of cases increases in the entire area being monitored). The expectation-based approach

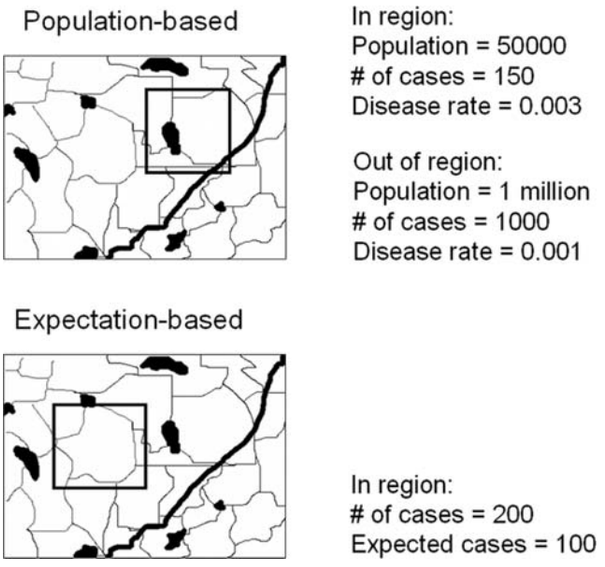

FIGURE 16.1 Population-based versus expectation-based scan statistics.

will find this increase very significant because the counts everywhere are higher than expected. However, the population-based approach will only find the increase significant if there is spatial variation in the amount of increase; otherwise the ratio of disease rate inside the region to disease rate outside the region remains constant, so no significant increase is detected. We discuss these approaches in more detail below.

For now, let us consider the expectation-based approach, assuming that we are given the observed number of cases c, as well as an estimated mean μ and standard deviation σ, for each zip code. How can we tell whether any zip code has a number of cases that is significantly higher than expected? One simple possibility would be to perform a separate statistical test for each zip code: for example, we might want to detect all zip codes with observed count more than three standard deviations above the mean. However, there are two main problems with this simple method. First, treating each zip code separately prevents us from using information about the spatial proximity of adjacent zip codes. For instance, while a single zip code with count two standard deviations above the mean might not be sufficiently surprising to trigger an alarm, we would probably be interested in detecting a cluster of four adjacent zip codes each with count two standard deviations above the mean. Thus, the first problem with performing separate statistical tests for each zip code is reduced power to detect clusters spanning multiple zip codes: we cannot detect such increases unless the amount of increase is so large as to make each zip code individually significant. A second, and somewhat more subtle, problem is that of *multiple hypothesis testing*. We typically perform statistical tests to determine if an area is significant at some fixed level α, such as $\alpha = 0.05$, which means that if there is no abnormality in that area (i.e., the "null hypothesis" of no clusters is true) our probability of a false alarm is at most α. A lower value of α results in less false alarms, but also reduces our chance of detecting a true disease cluster. Now let us imagine that we are searching for clusters in a large area containing 1000 zip codes, and that there happen to be no outbreaks today, so any areas we detect are false alarms. If we perform a separate significance test for each zip code, we expect each test to trigger an alarm with probability $\alpha = 0.05$. But because we are doing 1000 separate tests, our expected number of false alarms is $1000 \times 0.05 = 50$. Moreover, if these 1000 tests were independent, we would expect to get at least one false alarm with probability $1-(1-0.05)^{1000} \approx 1$. Of course, counts of adjacent zip codes are likely to be correlated, so the assumption of independent tests is not usually correct. The main point here, however, is that we are almost certain to get false alarms every day, and the number of such false alarms is proportional to the number of tests performed. One way to correct for multiple tests is the *Bonferroni correction* (Bonferroni, 1935): If we want to ensure that our probability of getting any false alarms is at most α, we report only those regions which are significant at level α/N, where N is the number of tests. The problem with the Bonferroni correction is that it is too conservative, thus reducing the power of the test to detect true outbreaks. In our example, with $\alpha = 0.05$ and $N = 1000$, we only signal an alarm if a region's statistical significance (p-value) is less than 0.00005, and thus only very obvious outbreaks can be detected.

As an alternative to this simple method, we can choose a set of regions to search over, where each region consists of a set of one or more zip codes. We can define the set of regions based on what we know about the size and shape of potential outbreaks; we can either fix the region shape and size, or let these vary as desired. We can then do a separate test for each region rather than for each zip code. This resolves the first problem of the previous method: assuming we have chosen the set of regions well, we can now detect attacks whether they affect a single zip code, a large number of zip codes, or anything in between. However, the disadvantage of this method is that it makes the multiple hypothesis testing problem even worse: the number of regions searched, and thus the number of tests performed, is typically much larger than the number of zip codes. In principle, the number of regions could be as high as 2^Z, where Z is the number of zip codes, but in practice the number of regions searched is much smaller (because we want to enforce constraints on the connectedness, size, and shape of regions). For example, if we consider circular regions centered at the centroid of some zip code, with continually varying radius (assuming that a region contains all zip codes with centroids inside the circle), the number of distinct regions is proportional to Z^2. For the

example above, this would give us one million regions to search, creating a huge multiple hypothesis testing problem; less restrictive constraints (such as testing ellipses rather than circles) would require testing an even larger number of regions.

This method of searching over regions, without adjusting for multiple hypothesis testing, was first used by Openshaw et al. (1988) in their geographical analysis machine (GAM). Openshaw et al. test a large number of overlapping circles of fixed radius, and draw all of the significant circles on a map; Figure 16.2 gives an example of what the output of the GAM might look like. Because we expect a large number of circles to be drawn even if there are no outbreaks present, the presence of detected clusters is not sufficient to conclude that there is an outbreak. Instead, the GAM can be used as a descriptive tool for outbreak detection: whether any outbreaks are present, and the location of such outbreaks, must be inferred manually from the number and spatial distribution of detected clusters. For example, in Figure 16.2 the large number of overlapping circles in the upper right of the figure may indicate an outbreak, while the other circles might be due to chance. The problem is that we have no way of determining whether any given circle or set of circles is statistically significant, or whether they are due to chance and multiple testing; it is also difficult to precisely locate those clusters which are most likely to correspond to true outbreaks. Besag and Newell (1991) propose a related approach, where the search is performed over circles containing a fixed number of cases; this approach also suffers from the multiple hypothesis testing problem, but again is valuable as a descriptive method for visualizing potential clusters.

The *scan statistic* was first proposed by Naus (1965) as a solution to the multiple hypothesis testing problem. Let us assume we have a score of some sort for each region, such as the *Z-score*, $Z = (c - \mu)/\sigma$. The Z-score is the number of standard deviations that the observed count c is higher than the expected count μ; a large Z-score indicates that the observed number of cases is much higher than expected. Rather than triggering an alarm if *any* region has Z-score higher than some fixed threshold, we instead find the distribution of the *maximum* score of all regions under the null hypothesis of no outbreaks. This distribution tells us what we should expect the most alarming score to be when the system is executed on data in which there is no outbreak. Then we compare the score of the highest-scoring (most significant) region on our data against this distribution to determine its statistical significance (or *p*-value). In other words, the scan statistic attempts to answer the question, "If there were no outbreaks, and we searched over all of these regions, how likely would we be to find *any* regions that score at least this high?" If the analysis shows that we would be very unlikely to find any such regions under the null hypothesis, we can conclude that the discovered region is a significant cluster. The main advantage of the scan statistics approach is that we can adjust correctly for multiple hypothesis testing: we can fix a significance level α, and ensure that the probability of having any false alarms on a given day is at most α, regardless of the number of regions searched. Moreover, because the scan statistic accounts for the fact that our tests are not independent, it will typically have much higher power to detect than a Bonferroni-corrected method. In some applications, the scan statistic results in a *most powerful* statistical test (see Kulldorff, 1997 for more details).

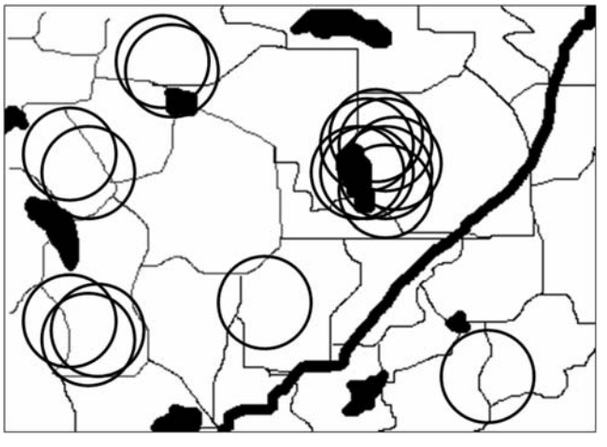

FIGURE 16.2 Example of a set of significant regions.

Although the scan statistic focuses on finding the single most significant region, it can also be used to find multiple regions: secondary clusters can be examined, and their significance found, though the test is typically somewhat conservative for these. The technical difficulty, however, is finding the distribution of the maximum region score under the null hypothesis. Turnbull et al. (1990) solved this problem for circular regions of fixed population, using the maximum number of cases in a circle as the test statistic, and using the method of randomization testing (discussed below) to find the statistical significance of discovered regions. The disadvantage of this approach is that it requires a fixed population size circle, and thus a multiple hypothesis testing problem still exists if we want to search over regions of multiple sizes or shapes. Kulldorff and Nagarwalla (1995) and Kulldorff (1997) solved the problem for variable size regions using a *likelihood ratio test*: The test statistic is the maximum of the likelihood ratio under the alternative and null hypotheses, where the alternative hypothesis represents clustering in that region and the null hypothesis assumes no clusters. We discuss their method, the "spatial scan statistic," in the following section.

3. THE SPATIAL SCAN STATISTIC

The *spatial scan statistic* (Kulldorff and Nagarwalla, 1995, Kulldorff, 1997) is a powerful and general method for

spatial cluster detection. It is in common use by epidemiologists for finding significant spatial clusters of disease cases, which may be indicative of an outbreak. In this section, we present the spatial scan statistic as originally described by Kulldorff, along with a number of generalizations, extensions, and variants which extend the scope and applicability of this method.

In its original formulation, Kulldorff's statistic assumes that we have a set of spatial locations s_i, and are given a count c_i and a population p_i corresponding to each location. For example, each s_i may represent the centroid of a census tract, the corresponding count c_i may represent the number of respiratory emergency department visits in that census tract, and the corresponding population p_i may represent the "at-risk population" of that census tract, derived from census population and possibly adjusted for covariates. The statistic makes the assumption that each observed count c_i is drawn randomly from a Poisson distribution with mean $q_i p_i$, where p_i is the (known) at-risk population of that area, and q_i is the (unknown) *risk*, or *underlying disease rate*, of that area. The risk is the expected number of cases per unit population; that is, we expect to see a number of cases equal to the product of the population and the risk, but the observed number of cases may be more or less than this expectation due to chance. Thus, our goal is to determine whether observed increases in count in a region are due to increased risk, or chance fluctuations. The Poisson distribution is commonly used in epidemiology to model the underlying randomness of observed case counts, making the assumption that the variance is equal to the mean. If this assumption is not reasonable (i.e., counts are "overdispersed" with variance greater than the mean, or "underdispersed" with variance less than the mean), we should instead use a distribution which separately models mean and variance, such as the normal or negative binomial distributions. We also assume that each count c_i is drawn independently, although the model can be extended to account for spatial correlations between nearby locations.

3.1. Detailed Description of the Spatial Scan Statistic

As discussed above, Kulldorff's spatial scan statistic attempts to detect spatial regions where the underlying disease rates q_i are significantly higher inside the region than outside the region. Thus, we wish to test the null hypothesis H_0 ("the underlying disease rate is spatially uniform") against the set of alternative hypotheses $H_1(S)$: "The underlying disease rate is higher inside region S than outside region S." More precisely, we have:

- $H_0: c_i \sim \text{Poisson}(q_{all}\, p_i)$ for all locations s_i, for some constant q_{all}.
- $H_1(S): c_i \sim \text{Poisson}(q_{in}\, p_i)$ for all locations s_i in S, and $c_i \sim \text{Poisson}(q_{out}\, p_i)$ for all locations s_i outside S, for some constants $q_{in} > q_{out}$.

The test statistic that we use is the likelihood ratio, that is, the likelihood (denoted by Pr) of the data under the alternative hypothesis $H_1(S)$ divided by the likelihood of the data under the null hypothesis H_0. This gives us, for any region S, a *score function*:

$$D(S) = \frac{\Pr(Data \mid H_1(S))}{\Pr(Data \mid H_0)}.$$

For Kulldorff's statistic, we obtain

$$D(S) = \left(\frac{C_{in}}{P_{in}}\right)^{C_{in}} \left(\frac{C_{out}}{P_{out}}\right)^{C_{out}} \left(\frac{C_{in}+C_{out}}{P_{in}+P_{out}}\right)^{-(C_{in}+C_{out})}, \text{ if } \frac{C_{in}}{P_{in}} > \frac{C_{out}}{P_{out}},$$

and

$$D(S) = 1 \quad \text{otherwise;}$$

see Kulldorff (1997) for a derivation. In this equation, C_{in} and C_{out} represent the aggregate count $\sum c_i$ inside and outside region S, and P_{in} and P_{out} represent the aggregate population $\sum p_i$ inside and outside region S, respectively. See Figure 16.3 for an example of the evaluation of $D(S)$ for a region. Kulldorff (1997) proved that this likelihood ratio statistic is *individually most powerful* for finding a single region of elevated disease rate: For the given model assumptions (H_0 and H_1), for a fixed false alarm rate, and for a given set of regions searched, it is more likely to detect the cluster than any other test statistic.

Given the above test statistic $D(S)$, the spatial scan statistic method can be easily applied by choosing a set of regions S, calculating the score function $D(S)$ for each of these regions, and obtaining the highest scoring region S^* and its score $D^* = D(S^*)$. We can imagine this procedure as moving a "spatial window" (like the rectangle drawn in Figure 16.3) all around

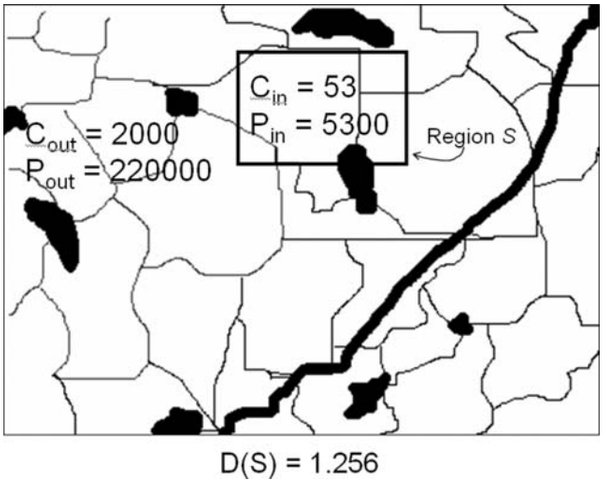

FIGURE 16.3 Counts inside and outside rectangular regions.

the search area, changing the size and shape of the window as desired, and finding the window which gives the highest score *D(S)*. Even though there are an infinite number of possible window positions, sizes, and shapes, we only need to evaluate the score function a finite number of times, since any two regions containing the same set of spatial locations s_i will have the same score. The region with the highest score *D(S)* is the "most significant region," that is, the region that is most likely to have been generated under the alternative hypothesis rather than the null hypothesis, and thus the region most likely to be a cluster. We typically search over the set of all "spatial windows" of a given shape and varying size; for example, Kulldorff et al. (1997) search over circular regions, Neill and Moore (2004a) search over square regions, and Neill and Moore (2004b) search over rectangular regions. Searching over a set of regions which includes both compact and elongated regions (e.g., rectangles or ellipses) has the advantage of higher power to detect elongated clusters resulting from wind dispersal of pathogens, but because the number of regions to search is increased, this also makes the scan statistic more difficult to compute. We discuss computational issues in more detail below. Chapter 19 describes more accurate modeling of windborne dispersion patterns.

Once we have found the regions with the highest scores *D(S)*, we must still determine which of these "potential clusters" are likely to be "true clusters" resulting from a disease outbreak, and which are likely to be due to chance. To do so, we calculate the statistical significance (*p*-value) of each potential cluster, and all clusters with *p*-value less than some fixed significance level α are reported. Because of the multiple hypothesis testing problem discussed above, we cannot simply compute separately whether each region score *D(S)* is significant, because we would obtain a large number of false positives, proportional to the number of regions searched. Instead, for each region *S*, we ask the question, "If this data set were generated under the null hypothesis H_0, how likely would we be to find any regions with scores higher than *D(S)*?" To answer this question, we use the method known as *randomization testing*: we randomly generate a large number of "replicas" under the null hypothesis, and compute the maximum score $D^* = \max_s D(S)$ of each replica. More precisely, each replica is a copy of the original search area that has the same population values p_i as the original, but has each value c_i randomly drawn from a Poisson distribution with mean $\frac{C_{all}}{P_{all}} p_i$, where C_{all} and P_{all} are respectively the total number of cases and the total population for the original search area. Once we have obtained D^* for each replica, we can compute the statistical significance of any region *S* by comparing *D(S)* to these replica values of D^*, as shown in Figure 16.4. The *p*-value of region *S* can be computed as $\frac{R_{beat} + 1}{R + 1}$, where *R* is the total number of replicas created, and R_{beat} is the number of replicas with D^* greater than *D(S)*. If this *p*-value is less than our significance level α, we conclude that the region is significant (likely to be a true cluster); if the *p*-value is greater than α, we conclude that the region is not significant (likely to be due to chance). We typically start from the most significant region S^* and test regions in order of decreasing *D(S)*, since if a region *S* is not significant, no region with lower *D(S)* will be significant. We note that the randomization testing approach given here has the benefit of bounding the overall false positive rate: regardless of the number of regions searched, the probability of any false alarms is bounded by the significance level α. Also, the more replications performed (i.e., the larger the value of *R*), the more precise the *p*-value we obtain; a typical value would be $R = 1000$. However, since the run time is proportional to the number of replications performed, this dramatically increases the amount of computation necessary. Finally, we note that spatial scan software is available at *www.satscan.org* and *www.autonlab.org*. The former is the very widely used SaTScan software of (Kulldorff and Information Management Services Inc., 2002). The latter is prototype software that is discussed below.

3.2. Generalizing the Spatial Scan Statistic

In this subsection, we consider a general statistical framework for the spatial scan statistic, extending it to allow for a large class of underlying models and thus a wide variety of application domains. As above, we wish to test a null hypothesis H_0 (a model of how the data is generated, assuming there are no clusters of interest) against the set of alternative hypotheses $H_1(S)$, each of which represents a relevant cluster in some region *S* of space. Assuming that the null hypothesis and each alternative hypothesis are point hypotheses (with no free parameters), we can use the likelihood ratio $D(S) = \frac{\Pr(Data \mid H_1(S))}{\Pr(Data \mid H_0)}$ as our test statistic. A more interesting question is what to do when each hypothesis has some parameter space Θ: let $\theta_1(S) \in \Theta_1(S)$ denote parameters for the alternative hypothesis $H_1(S)$, and let $\theta_0 \in \Theta_0$ denote parameters for the null hypothesis H_0. There are two possible answers to this question. In the more typical *maximum likelihood framework*, we use the estimates of each set of parameters that maximize the likelihood of the data: $D(S) = \frac{\max_{\theta_1(S) \in \Theta_1(S)} \Pr(Data \mid H_1(S), \theta_1(S))}{\max_{\theta_0 \in \Theta_0} \Pr(Data \mid H_0, \theta_0)}$. We then perform randomization testing using the maximum likelihood estimates of the parameters under the null hypothesis. In the *marginal likelihood framework*, we instead average over the possible values of each parameter:

$$D(S) = \frac{\int_{\theta_1(S) \in \Theta_1(S)} \Pr(Data \mid H_1(S), \theta_1(S)) \Pr(\theta_1(S))}{\int_{\theta_0 \in \Theta_0} \Pr(Data \mid H_0, \theta_0) \Pr(\theta_0)}.$$

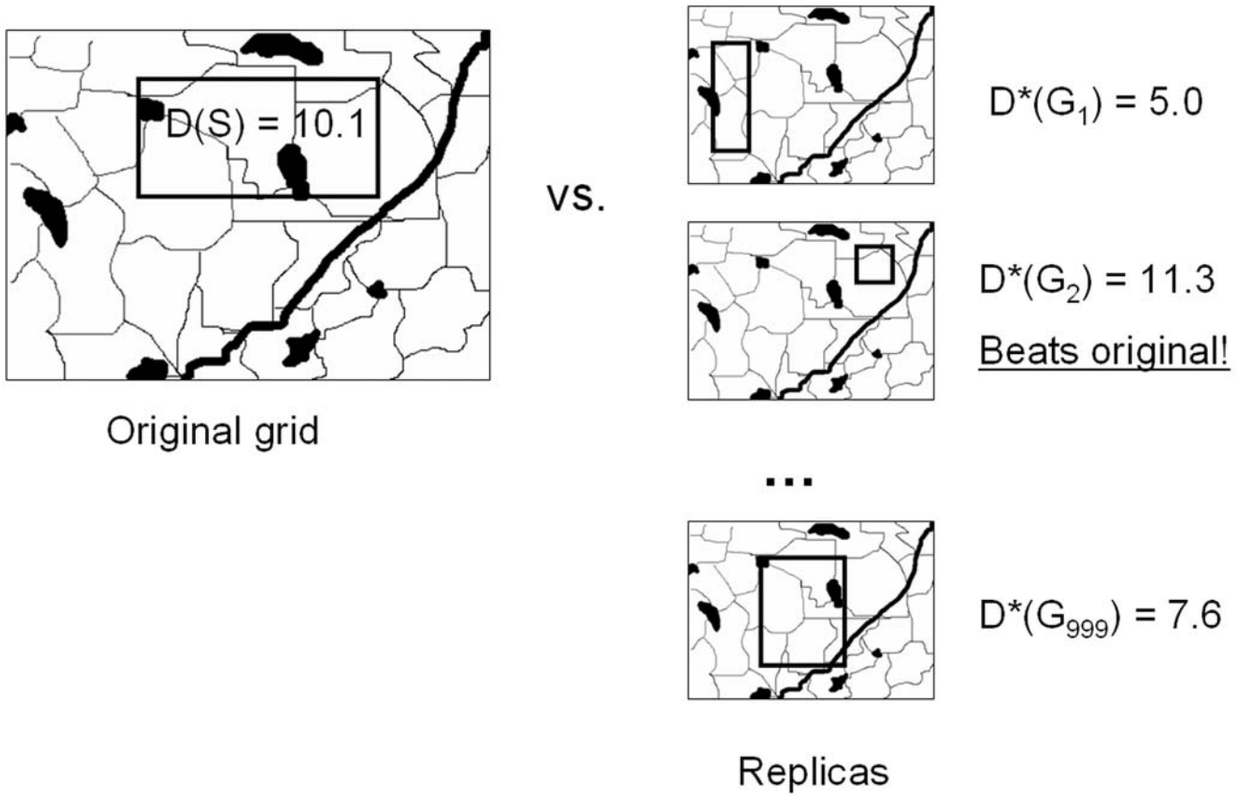

FIGURE 16.4 Calculating the largest region score 1000 times. On the left it is calculated on the real data. On the right it is calculated 999 times on randomized data.

Neill et al. (2005c) present a Bayesian variant of Kulldorff's spatial scan statistic using the marginal likelihood framework; here we focus on the simpler, maximum likelihood approach, and give an example of how new scan statistics can be derived.

Our first step is to choose the null hypothesis and set of alternative hypotheses that we are interested in testing. Here we consider the expectation-based scan statistic discussed above, where we are given the *baseline* (or expected count) b_i and the observed count c_i for each spatial location s_i, and our goal is to determine if any spatial location s_i has c_i significantly greater than b_i. We test the null hypothesis H_0 against the set of alternative hypotheses $H_1(S)$, where:

H_0: $c_i \sim \text{Poisson}(b_i)$ for all spatial locations s_i.

$H_1(S)$: $c_i \sim \text{Poisson}(q\, b_i)$ for all spatial locations s_i in S, and $c_i \sim \text{Poisson}(b_i)$ for all spatial locations s_i outside S, for some constant $q > 1$.

Here, the alternative hypothesis $H_1(S)$ has one parameter, q (the relative risk in region S), and the null hypothesis H_0 has no parameters. Computing the likelihood ratio, and using the maximum likelihood estimate for our parameter q, we obtain the following expression for $D(S)$:

$$D(S) = \frac{\max_{q>1} \prod_{s_i \in S} \Pr(c_i \sim \text{Poisson}(qb_i)) \prod_{s_i \notin S} \Pr(c_i \sim \text{Poisson}(b_i))}{\prod_{s_i} \Pr(c_i \sim \text{Poisson}(b_i))}$$

We find that the value of q that maximizes the numerator is $q = \max(1, C/B)$, where C and B are the total count $\sum c_i$ and total baseline $\sum b_i$ of region S respectively. Plugging in this value of q, and working through some algebra, we obtain: $D(S) = \left(\frac{C}{B}\right)^C \exp(B - C)$, if $C > B$, and $D(S) = 1$ otherwise. Then the most significant region S^* is the one with the highest value of $D(S)$, as above. We can calculate the statistical significance (p-value) of this region by randomization testing as above, where replicas are generated under the null hypothesis $c_i \sim \text{Poisson}(b_i)$.

For the population-based method, a very similar derivation can be used to obtain Kulldorff's statistic: we compute the

maximum likelihood parameter estimates $q_{in} = C_{in}/P_{in}$, $q_{out} = C_{out}/P_{out}$, and $q_{all} = C_{all}/P_{all}$. We have also used this general framework to derive scan statistics assuming that counts c_i are generated from normal distributions with mean (i.e., expected count) μ_i and variance σ_i^2; these statistics are useful if counts might be overdispersed or underdispersed. Many other likelihood ratio scan statistics are possible, including models with simultaneous attacks in multiple regions and models with spatially varying (rather than uniform) disease rates. We believe that some of these more complex model specifications may have more power to detect relevant and interesting clusters, while excluding those potential clusters which are not epidemiologically relevant.

3.3. Computational Considerations

While considering deployment of spatial methods, it may be necessary to account for the computational time of an algorithm if it is to be run over a very large area or use many randomizations. In this section we return to the question of what set of regions to search over, and discuss how to perform this search efficiently. First, we note that the run time of the spatial scan can be approximated by the product of three factors: the number of replications R, the average number of regions searched per replication $|S|$, and the average time to search a region t. The number of replications R is typically fixed in advance, but we can stop early if many replicas beat the original search area (i.e., the maximum region scores D^* of the replicas are higher than the maximum region score D^* of the original). If this happens, it is clear that no significant clusters are present. The other two factors $|S|$ and t depend on both the set of regions to be searched and the algorithm used to search these regions. For a set of M distinct spatial locations in two dimensions, the number of circular or axis-aligned square regions (assuming that the size of the circle or square can vary) is proportional to M^3, while the number of axis-aligned rectangular regions (assuming that both dimensions of the rectangle can vary) is proportional to M^4. For non—axis-aligned squares or rectangles, we must also multiply this number by the number of different orientations searched. However, most algorithms only search a subset of these regions: for example, Kulldorff (1999) algorithm searches only circles centered at one of the M spatial locations, and the number of such regions is proportional to M^2, not M^3. Another possibility is to aggregate the spatial locations to a grid, either uniform or based on the distinct spatial coordinates of the data points. For a two-dimensional $N \times N$ grid, the number of axis-aligned square regions is proportional to N^3, and the number of axis-aligned rectangular regions is proportional to N^4. Whatever set of regions we choose, the simplest possible implementation of the scan statistic is to search each of these regions by stepping through the M spatial locations, determining which locations are inside and outside the region, computing the aggregate populations/baselines and counts, and applying the score function. Thus, in this approach, we have $|S|$ (number of regions searched per replication) equal to the total number of distinct regions, and t (time to search a region) proportional to the number of spatial locations M.

There are several possible ways to improve on the runtime of this naïve approach. First, we can reduce the time to search a region t, making this search time independent of the number of spatial locations M. We consider two possible methods for searching a region in constant time. The first method, which we call "incremental addition," assumes that we want to search over all regions of a given type: for example, in the approach of Kulldorff (1999), we want to search all distinct circular regions centered at one of the spatial locations. To do so, we increase the region's size incrementally, such that one new spatial location at a time enters the region; for each new location, we can add that location's count and population/baseline to the aggregates, and recompute the score function. For example, in Kulldorff's method, for each location s_i we keep a list of the other locations, sorted from closest to furthest away. Then we can search over the M distinct circular regions centered at s_i by adding the other points one at a time in order. Because the sorting only has to be done once (and does not have to be repeated for each replication), this results in constant search time per region. In other words, Kulldorff's method requires time proportional to M^2 to search over all M^2 regions. This must be done for each of the R replications, giving total search time proportional to RM^2. The second method assumes that points have been aggregated to an $N \times N$ grid, and that we are searching over squares or rectangles. We can use the well-known "cumulative counts" technique to search in constant time per region; see Neill and Moore (2004b) for more details. As a result, we can perform the scan statistic for gridded square or rectangular regions in time proportional to R times the number of regions, that is, RN^3 or RN^4 for square or rectangular regions, respectively.

Even if we can search in constant time per region, the spatial scan statistic is still extremely computationally expensive, because of the large number of regions searched. For example, to search over all rectangular regions on a 256×256 grid, and perform randomization testing (assuming $R = 1000$ replications), we must search a total of 1.1 trillion regions, which would take 14 to 45 days on our test systems. This is clearly far too slow for real-time detection of emerging outbreaks. While one option is to simply search fewer regions, this reduces our power to detect disease clusters. A better option is provided by the *fast spatial scan* algorithms of Neill and Moore (2004a, 2004b) and Neill et al. (2005a), which allow us to reduce the number of regions searched, but without losing any accuracy. The idea is that, since we only care about the most significant regions, that is, those with the highest scores $D(S)$, we do not need to search a region S if we know that it will not have a

high score. Thus, we start by examining large regions S, and if we can show that none of the smaller regions contained in S can have high scores, we do not need to actually search each of these regions. Thus, we can achieve the same result as if we had searched all possible regions, but by only searching a small fraction of these. Further speedups are gained by the use of multiresolution data structures, which allow us to efficiently move between searching at coarse and fine resolutions. A detailed description of these structures is beyond the scope of this chapter, but this has been an active area of research in computer science and biomedical informatics. In these algorithms, we apply multiresolution techniques in order to accelerate the spatial scan statistic for searching square regions (Neill et al. 2005a) or rectangular regions (Neill and Moore, 2004b). We have also extended these techniques to rotated rectangular regions and multidimensional data sets (Neill et al., 2005a). These methods are able to search hundreds or thousands of times faster than an exhaustive search, without any loss of accuracy (i.e., the fast spatial scan finds exactly the same region and p-value as exhaustive search). As a result, these methods have enabled us to perform spatial scans on data sets such as nationwide over-the-counter sales data, from over 20,000 stores in near real time, searching for disease clusters in minutes or hours rather than days or weeks. In Figure 16.5, we show a screen shot from the current version of our spatial scan software, available from our websites *www.autonlab.org* (Auton Laboratory) and *rods.health. pitt.edu* (RODS Laboratory).

3.4. Calculation of Populations/Baselines

In the discussion of the spatial scan methods above, we have paid relatively little attention to the question of how the underlying populations or baselines are obtained. In the population-based methods, we often start from census data, which gives an unadjusted population p_i corresponding to each spatial location s_i. This population can then be adjusted for covariates such as the distribution of patient age and gender, giving an estimated "at-risk population" for each spatial location. In a recent paper, Kleinman et al. (2005) suggest two additional model-based adjustments to the population estimates. First, they present a method for temporal adjustment (accounting for day of week, month of the year, and holidays),

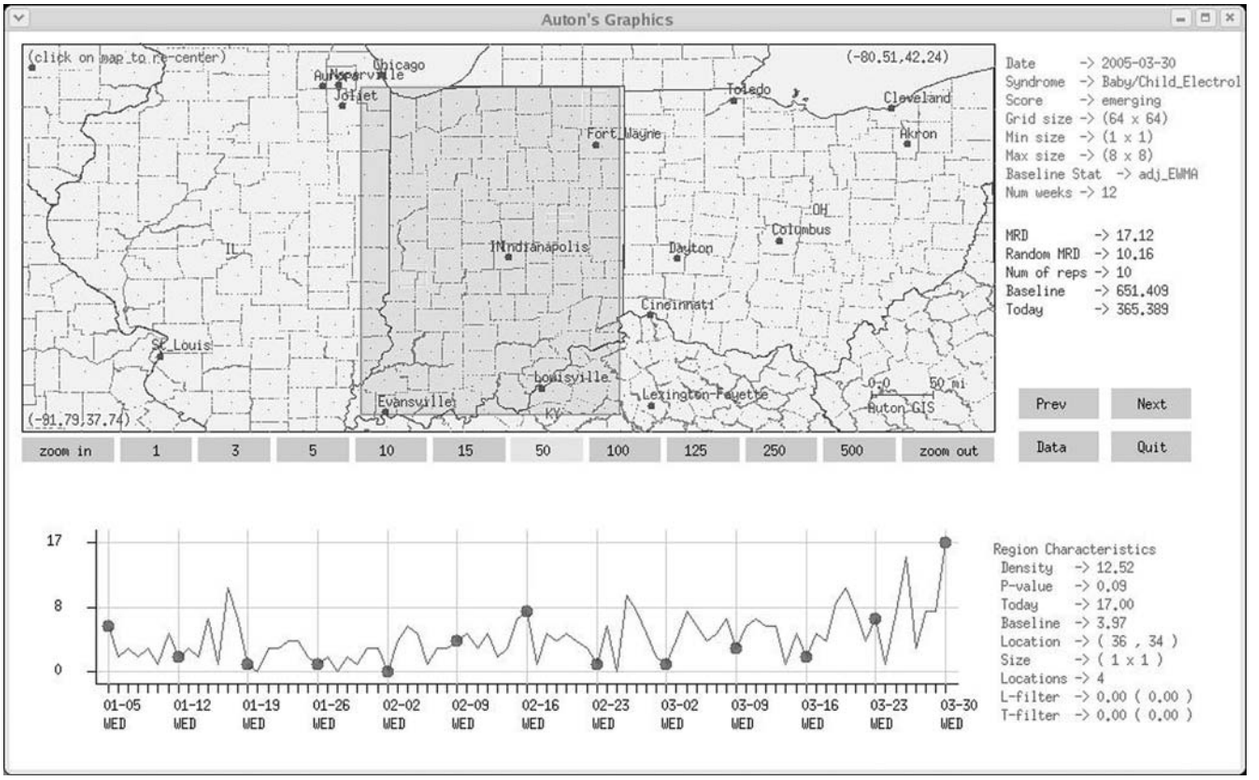

FIGURE 16.5 A screen shot from the current RODS spatial scan software. The bottom time series is the recent history of electrolyte sales in a region north of Indianapolis that was determined as the most significant region in the state on that day (the actual region is hidden in order to avoid providing information that might reveal the data providers).

making the populations larger on days when more visits are likely (e.g., Mondays during influenza season) and smaller on days when fewer visits are likely (e.g., Sundays and holidays). Second, they apply a "generalized linear mixed models" (GLMM) approach, first presented in Kleinman et al. (2004), to adjust for the differing baseline risk in each census tract. This makes the adjusted population larger in tracts that have a larger baseline risk, which makes sense since a given number of observed cases should not be as significant if the observed counts in that region are consistently high. These baseline risks are computed from historical data, that is, the time series of past counts in each census tract, using the GLMM version of logistic regression to fit the model; see Kleinman et al. (2004) for details.

In the expectation-based methods, we also make use of the historical data, but for these methods the goal is to directly estimate the number of cases we expect to see in each area. Thus, we must predict the expected number of cases b_i for each spatial location s_i based on the history of past counts at that location (and optionally, considering spatial correlation of counts at nearby locations). This becomes, in essence, a univariate time series analysis problem, and many of the techniques discussed in Chapter 14 can be used. For example, to adjust for day of week effects, we can either stratify by day of week (i.e., predict Tuesday's expected count by using only prior Tuesdays) or adjust for day of week using the sickness availability method (see Chapter 14). For data sets without strong seasonal effects, simple mean or exponentially weighted moving average methods (e.g., estimating the expected value of today's count as the mean of the counts 7, 14, 21, and 28 days ago) can be sufficient, but for data sets with strong seasonality, these methods will lag behind the seasonal trend, resulting in numerous false positives for increasing trends (e.g., sales of cough and cold medication at the start of winter) or false negatives for decreasing trends (e.g., cough and cold sales at the end of winter). To account for these trends, we recommend the use of regression methods (either weighted linear regression or nonlinear regression depending on the data) to extrapolate the current counts; see Neill et al. (2005b) for more details of this expectation-based approach. Another possibility is to make the assumption of independence of space and time, as in Kulldorff et al. (2005); this means that the expected count in a given region is equal to the total count of the entire area under surveillance, multiplied by the historical proportion of counts in that region. This approach is successful in detecting very localized outbreaks, but loses power to detect more widespread outbreaks (Neill et al., 2005b). The reason for this is that a widespread outbreak will increase the total count significantly, thus increasing the expected count in the outbreak region, and hence making the observed increase in counts seem less significant. In the worst scenario, a massive outbreak which causes a constant, multiplicative increase in counts across the entire area under surveillance would be totally ignored by this approach; this is also true for many of the population-based methods, since they only detect spatial variation in disease rate, not an overall increase in counts. If these methods are used, we recommend using a purely temporal method in parallel to ensure that large-scale outbreaks (as well as localized outbreaks) can be detected. Either way, the accurate inference of expected counts from historical data is still an open problem, with different methods performing well for different data sets and outbreak types. See Neill et al. (2005b) for empirical testing of the various time series methods and further discussion.

3.5. From Space to Space-Time

In this subsection, we briefly consider extensions of the spatial scan statistic to spatio-temporal cluster detection. The space-time scan statistic was first proposed by Kulldorff et al. (1998), and a variant was applied to prospective disease surveillance by Kulldorff (2001). The goal of the space-time scan statistic is a straightforward extension of the purely spatial scan: to detect regions of space-time where the counts are significantly higher than expected. Let us assume that we have a discrete set of time steps $t = 1...T$ (e.g., daily observations for T days), and for each spatial location s_i, we have counts c_i^t representing the observed number of cases in the given area on each time step. There are two very simple ways of extending the spatial scan to space-time: to run a separate spatial scan for each time step t, or to treat time as an extra dimension and thus run a single multidimensional spatial scan in space-time (for example, we could search over three-dimensional "hyper-rectangles" which represent a given rectangular region of space during a given time interval). The problem with the first method is that, by only examining one day of data at a time, we may fail to detect more slowly emerging outbreaks. The problem with the second method is that we tend to find less relevant clusters: for prospective disease surveillance, we want to detect newly emerging clusters of disease, not those that have persisted for a long time. Thus, in order to achieve better methods for space-time cluster detection, we must consider the question, "How is the time dimension different from space?" In Neill et al. (2005b), we argue that there are three main distinctions:

1. The concept of "now." In the time dimension, the present is an important point of reference: we are typically only interested in disease clusters that are still "active" at the present time, and that have emerged within the recent past (e.g., within a few days or a week). We do not want to detect clusters that have persisted for months or years, and we are also not interested in those clusters which have already come and gone. The exception to this, of course, is if we are performing a *retrospective analysis*, attempting to detect all space-time clusters regardless of how long ago they occurred. The space-time scan statistic for retrospective analysis was first presented in Kulldorff et al. (1998),

and the space-time scan statistic for prospective analysis was first presented in Kulldorff (2001). In brief, the retrospective statistic searches over time intervals $t_{min}...t_{max}$, where $1 \leq t_{min} \leq t_{min} \leq T$, while the prospective statistic searches over time intervals $t_{min} \ldots T$, where $1 \leq t_{min} \leq T$, adjusting correctly for multiple hypothesis testing in each case. We focus here on prospective analysis, since this is more relevant for our typical disease surveillance task.

2. "Learning from the past." In the space-time cluster detection task, we often do not have reliable denominator data (i.e., populations), so we must infer the expected counts b_i^t of recent days from the time series of previous counts c_i^t, taking into account effects such as seasonality and day of week. Some methods for inferring these expected counts were discussed in the previous section; see Neill et al. (2005b) for further discussion.
3. The "arrow of time." Time has a fixed directionality, moving from the past, through the present, to the future. We typically expect disease clusters to *emerge* in time: For example, a disease may start out having only minor impact on the affected population, then increase its impact (and thus the observed symptom counts) either gradually or rapidly until it peaks. Based on this observation, we propose a variant of the scan statistic designed for more rapid detection of emerging outbreaks (Neill et al., 2005b). The idea is that rather than assuming (as in the standard, "persistent" space-time scan statistic) that the disease rate q remains constant over the course of an epidemic, we expect the disease rate to *increase* over time, and thus we fit a model which assumes a monotonically increasing sequence of disease rates q_t at each affected time step t in the affected region. In Neill et al. (2005b), we show that this "emerging cluster" space-time scan statistic often outperforms the standard "persistent cluster" approach. We note that Iyengar (2005) accounts for a different aspect of the arrow of time: this method searches over truncated pyramid shapes in space-time, allowing detection of spatial clusters that move, grow, or shrink linearly with time.

Taking these factors into account, the prospective space-time scan statistic has two main parts: inferring (based on past counts) what we expect the recent counts to be, and finding regions where the observed recent counts are significantly higher than expected. More precisely, given a "temporal window size" W, we wish to know whether any space-time cluster within the last W days has counts c_i^t higher than expected. To do so, we first infer the expected counts $b_i^t = E[c_i^t]$ for all spatial locations on each recent day t, $T - W < t \leq T$. See Neill et al. (2005b), Kulldorff et al. (2005), and Kleinman et al. (2005) for methods of inferring these expected counts; earlier methods such as Kulldorff et al. (1998) and Kulldorff (2001) instead use at-risk populations determined from census data. Next, we choose the models H_0 and $H_1(S, t_{min})$, where the null hypothesis H_0 assumes no clusters and the alternative hypothesis $H_1(S, t_{min})$ represents a cluster in spatial region S starting at time t_{min} and continuing to the present time T. Neill et al. (2005b) gives two such models, one for persistent clusters and one for emerging clusters. From our model, we can derive the corresponding score function $D(S, t_{min})$ using the likelihood ratio statistic, and then find the space-time cluster (S^*, t_{min}^*) which maximizes the score function D. Finally, we can compute the statistical significance (p-value) of this space-time cluster by randomization testing, as above. More details of the space-time method described here, as well as empirical tests on several semi-synthetic outbreak data sets, are given in Neill et al. (2005b). We also refer the reader to Kulldorff et al. (1998, 2005), Kulldorff (2001), and Kleinman et al. (2005) for other useful perspectives on space-time cluster detection.

3.6. When to Apply Spatial Scan Approaches

Within epidemiology, scan statistics are a well-used and thriving analytic method. As a result, they have been incorporated into several experimental biosurveillance systems such as Heffernan et al. (2004), Lombardo et al. (2003), and Yih et al. (2004). We recommend caution when using scan statistics on new kinds of data. For example, in our early experiences of applying scan statistics to over-the-counter retail pharmacy data, it was immediately clear that simplistic assumptions in the underlying model can lead to false alarms: there are dozens of nondisease-related reasons for clusters of over-the-counter medication purchases to occur. Conversely, with the wrong data, even a sophisticated model will fail. For example, if home zipcodes are the only data in an emergency department's records then an attack on a downtown office location might not appear as a spatial cluster (although it is possible that appropriate use of commuting statistics can help in this case) (Buckeridge et al., 2003, Duczmal and Buckeridge, 2005). We believe that careful modeling is needed in order to overcome these effects on novel sources of data, and there is considerable ongoing work in the area, such as Kleinman et al. (2004, 2005).

4. RELATED METHODS

In this chapter, we have discussed those methods that can be used to solve the two problems of cluster detection: determining whether any significant clusters exist, and pinpointing the spatial location and extent of clusters. Many other spatial methods can be found in the literature on spatial epidemiology and spatial statistics, although most such methods either do not find specific clusters, or do not evaluate the statistical significance of discovered clusters. More general overviews of the literature on spatial statistical methods can be found in Lawson (2001) and Elliott et al. (2000). In addition to the spatial cluster detection methods discussed here, these methods include *general* and *focused* clustering methods, *disease mapping* approaches, and *spatial cluster modeling*.

General clustering methods are hypothesis testing methods that test for a general tendency of the data to cluster; in other words, they attempt to answer the question, "Is this data set more spatially clustered than we would expect?" Such methods do not identify specific clusters, but instead give a single result of "spatially clustered" or "not spatially clustered." These methods are useful if we want to know whether anything unexpected is going on, but do not care about the specific locations of unexpected events. Examples of such methods include Whittemore et al. (1987), Cuzick and Edwards (1990), and Tango (1995); see Lawson (2001) and Elliott et al. (2000) for more details. We also refer the interested reader to two general tests for space-time clustering: Knox (1964) and Mantel (1967).

Focused clustering methods are hypothesis testing methods that, given a prespecified spatial location, attempt to answer the question, "Is there an increase in risk in areas near this location?" These methods can be used to examine potential environmental hazards, such as testing for an increased risk of lung cancer near a coal-burning power plant. Since the locations are specified in advance, these methods cannot be used to identify specific cluster locations, but are instead used to test locations that have been identified by other means. Examples of such methods include Stone (1988), Besag and Newell (1991), and Lawson (1993); see Lawson (2001) and Elliott et al. (2000) for more details.

Disease mapping approaches have the goal of producing a spatially smoothed map of the variation in disease risk. For example, a very simple disease mapping approach might plot the observed disease rate (number of observed cases per unit population) in each area; more advanced approaches use a variety of Bayesian models and other spatial smoothing techniques to estimate the underlying risk of disease in each area. These methods do not explicitly identify cluster locations, but disease clusters may be inferred manually by identifying high-risk areas on the resulting map. Nevertheless, no hypothesis testing is typically done, so we cannot draw statistical conclusions as to whether these high-risk areas have resulted from true disease clusters or from chance fluctuations. Examples of such methods include Clayton and Kaldor (1987), Besag et al. (1991), and Clayton and Bernardinelli (1992); see Lawson, (2001) and Elliott et al. (2000) for more details.

Finally, *spatial cluster modeling* methods attempt to combine the benefits of disease mapping and spatial cluster detection, by constructing a probabilistic model in which the underlying clusters of disease are explicitly represented. A typical approach is to assume that cases are generated by some underlying process model which depends on a set of cluster centers, where the number and locations of cluster centers are unknown. Then we attempt to simultaneously infer all the parameters of the model, including the cluster centers and the disease risks in each area, using a simulation method such as reversible jump Markov chain Monte Carlo (Green, 1995). Thus, precise cluster locations are inferred, and while no formal significance testing is done, the method is able to compare models with different numbers of cluster centers, giving an indication of both whether there are any clusters and where each cluster is located. One typical disadvantage of such methods is computational: the underlying models rarely have closed-form solutions, and the Markov chain Monte Carlo methods used to approximate the model parameters are often computationally intensive. Examples of such methods include Lawson (1995), Lawson and Clark (1999), and Gangnon and Clayton (2000). For a more detailed discussion of spatial cluster modeling, see Lawson and Denison (2002).

SUMMARY

In this chapter, we have presented an overview of methods for spatial cluster detection, focusing on spatial scan statistics. The statistical and computational techniques described here have extended the spatial scan framework by increasing the generality of the underlying models, as well as making the spatial scan computationally feasible even for very large sources of data.

ADDITIONAL RESOURCES

- *http://www.autonlab.org* (Auton Laboratory website includes Fast Spatial Scan software, available for download, and spatial scan papers by Neill, Moore, and others.)
- *http://www.satscan.org* (SaTScan website includes latest version of Martin Kulldorff's SaTScan software, available for download, and spatial scan papers by Kulldorff and others.)

REFERENCES

Besag, J., Newell, J. (1991). The detection of clusters in rare diseases. *J R Stat Soc A* 154:143–55.

Besag, J., York, J., Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Ann Inst Stat Math* 43:1–59.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. In: *Studi in Onore del Professore Salvatore Ortu Carboni*, 13–60.

Buckeridge, D. L., Musen, M. A., Switzer, P., et al. (2003). An analytic framework for space-time aberrancy detection in public health surveillance data. In: *Proceedings of American Medical Informatics Association Annual Symposium*, 120–4.

Clayton, D. G., Bernardinelli, L. (1992). Bayesian methods for mapping disease risk. In: Elliot, P., Cuzick, J., English, D., et al., eds. *Geographical and Environmental Epidemiology: Methods for Small Area Studies*, 205–20. Oxford: Oxford University Press.

Clayton, D. G., Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 43:671–81.

Cuzick, J., Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *J R Stat Soc B* 52:73–104.

Duczmal, L., Buckeridge, D. (2005). Using modified spatial scan statistics to improve detection of disease out breaks when exposure occurs in the workplace. *MMWR Morb Mortal Wkly Rep* 54(suppl): 187.

Elliott, P., Wakefield, J. C., Best, N. G., et al., eds. (2000). *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford University Press.

Gangnon, R. E., Clayton, M. K. (2000). Bayesian detection and modeling of spatial disease clustering. *Biometrics* 56:922–35.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–32.

Heffernan, R., Mostashari, F., Das, D., et al. (2004). Syndromic surveillance in public health practice. *Emerg Infect Dis* 10:858–64.

Kleinman, K., Abrams, A., Kulldorff, M., et al. (2005). A model-adjusted space-time scan statistic with an application to syndromic surveillance. In: *Epidemiology and Infection*. 133:409–19.

Kleinman, K., Lazarus, R., Platt, R. (2004). A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *Am J Epidemiol* 159:217–24.

Knox, E. G. (1964). The detection of space-time interactions. *Appl Stat* 13:25–29.

Kulldorff, M. (1999). Spatial scan statistics: models, calculations, and applications. In: Glaz, J., Balakrishnan, N., eds. *Scan Statistics and Applications*, 303–22. Boston: Birkhauser.

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *J R Stat Soc* 164:61–72.

Kulldorff, M., Athas, W. F., Feurer, E. J., et al. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *Am J Public Health* 88:1377–80.

Kulldorff, M., Feuer, E. J., Miller, B. A., et al. (1997). Breast cancer clusters in the northeast United States: a geographic analysis. *Am J Epidemiol* 146:161–70.

Kulldorff, M., Heffernan, R., Hartman, J., et al. (2005). A space-time permutation scan statistic for the early detection of disease outbreaks. *PLOS Med* 2:216–24.

Kulldorff, M., Information Management Services Inc. (2002). *SaTScan v. 3.1: Software for the spatial and space-time scan statistics*.

Kulldorff, M., Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Stat Med* 14:799–810.

Lawson, A. B. (1993). On the analysis of mortality events around a prespecified fixed point. *J R Stat Soc A* 156:363–77.

Lawson, A. B. (1995). Markov Chain Monte Carlo techniques for putative pollution source problems in environmental epidemiology. *Stats Med* 14:2473–86.

Lawson, A. B. (2001). *Statistical Methods in Spatial Epidemiology*. Chichester: John Wiley & Sons.

Lawson, A. B., Clark, A. (1999). Markov chain Monte Carlo methods for putative sources of hazard and general clustering. In: Lawson, A.B., ed. *Disease Mapping and Risk Assessment for Public Health*, Chichester: John Wiley & Sons.

Lawson, A. B., Denison, D. G. T., eds. (2002). *Spatial Cluster Modelling*. Boca Raton, FL: Chapman & Hall/CRC.

Lombardo, J., Burkom, H., Elbert, E. (2003). A systems overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II). *J Urb Health* 80(Suppl 1):i32–42.

Mantel, N. (1967). The detection of cancer clustering and the generalized regression approach. *Cancer Res* 27:209–20.

Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on the line. *J Am Stat Assoc* 60:532–8.

Neill, D. B., Cooper, G. F., Moore, A. W. (2005c). A Bayesian scan statistic for spatial cluster detection. In *Advances in Disease Surveillance*, in press.

Neill, D. B., Moore, A. W. (2004a). A fast multi-resolution method for detection of significant spatial disease clusters. *Adv Neural Information Processing Syst* 16:651–8.

Neill, D. B., Moore, A. W. (2004b). Rapid detection of significant spatial clusters. In proc. 10th ACM SI6KOO Conference on Knowledge Discovery and Data Mining, 256–65.

Neill, D. B., Moore, A. W., Pereira, F., et al. (2005a). Detecting significant multidimensional spatial clusters. *Adv Neural Information Processing Syst* 17:969–976.

Neill, D. B., Moore, A. W., Sabhnani, M., et al. (2005b). Detection of emerging space-time clusters. In proc. 11th ACM SI6KOO Conference on Knowledge Discovery and Data Mining, 218–227.

Openshaw, S., Charlton, M., Craft, A., et al. (1988). Investigation of leukemia clusters by use of a geographical analysis machine. *Lancet* 1:272–3.

Snow, J. (1855). *On the Mode of Communication of Cholera*. London: John Churchill.

Stone, R. A. (1988). Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Stats Med* 7:649–660.

Tango, T. (1995). A class of tests for detecting "general" and "focused" clustering of rare diseases. *Stats Med* 14:2323–34.

Turnbull, B. W., Iwano, E. J., Burnett, W. S., et al. (1990). Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *Am J Epidemiol* 132:s136–43.

Wagner, M., Robinson, J., Tsui, F., et al. (2003). Design of a national retail data monitor for public health surveillance. *J Am Med Inform Assoc* 10:409–18.

Whittemore, A., Friend, N., Brown, B., et al. (1987). A test to detect clusters of disease. *Biometrika* 74:631–5.

Yih, W. K., Caldwell, B., Harmon, R. (2004). The National Bioterrorism Syndromic Surveillance Demonstration Program. *MMWR Morb Mortal Wkly Rep* 53(Suppl):43–6.