Anomalous Spatial Cluster Detection

Daniel B. Neill
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
neill@cs.cmu.edu

Andrew W. Moore
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
awm@cs.cmu.edu

ABSTRACT
We describe a general statistical and computational framework for the detection of anomalous spatial clusters, based on the *spatial scan statistic* [1]. Much of this material has been adapted from [2], to which we refer the reader for a more detailed discussion. We focus here on the purely spatial cluster detection task; for extensions to space-time cluster detection, the reader is referred to [3] and the references contained therein.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Apps—Data Mining

General Terms
Algorithms

Keywords
Cluster detection, anomaly detection, spatial scan statistics.

1. INTRODUCTION
Spatial cluster detection has two main goals: to identify the locations, shapes, and sizes of potentially anomalous spatial regions, and to determine whether each of these potential clusters is more likely to be a "true" cluster or simply a chance occurrence. In other words, we wish to answer the questions, is anything unexpected going on, and if so, where? This task can be broken down into two parts: first figuring out what we expect to see, and then determining which regions deviate significantly from our expectations. For example, in the application of disease surveillance, we examine the spatial distribution of disease cases (or some related quantity, such as the number of emergency department visits or over-the-counter drug sales of a specific type), and our goal is to determine whether any regions have sufficiently high case counts to be indicative of an emerging disease epidemic in that area. Thus we first infer the expected case count for each spatial location (e.g. zip code), typically based on historical data (though simpler approaches, such as assuming that number of cases is proportional to census population, can also be used). Then the next step is to determine which (if any) regions have significantly more cases than expected. One simple possibility would be to perform a separate statistical test for each spatial location under consideration, and report all locations that are significant at some level α. However, there are two main problems with this simple approach. First, we cannot use information about the spatial proximity of locations: for example, while a single zip code with count two standard deviations higher than expected might not be sufficiently interesting to trigger an alarm, we would probably be interested in a cluster of adjacent zip codes where each zip code's count is two standard deviations higher than expected. Second, *multiple hypothesis testing* is a problem: because we are performing a separate hypothesis test for each spatial location, where each hypothesis test has some fixed false positive rate α, the total number of false positives that we expect is $Y\alpha$, where Y is the total number of locations tested. For large Y, we are almost certain to get huge numbers of false alarms; alternatively, we would have to use a threshold α so low that the power of the test would be drastically reduced.

To deal with these problems, Kulldorff [1] proposed the *spatial scan statistic*. This method searches over a given set of spatial regions (where each region consists of a set of locations), finding those regions which are most likely to be generated under the "alternative hypothesis" of clustering rather than the "null hypothesis" of no clustering. A likelihood ratio test is used to compare these hypotheses, and randomization testing is used to compute the p-value of each detected region, correctly adjusting for multiple hypothesis testing. Thus, we can both identify potential clusters and determine whether each is significant. Our recent work on spatial scanning has two main emphases: first, to generalize the statistical framework to a larger class of underlying models, making the spatial scan applicable and useful for a wide variety of application domains; and second, to make these methods computationally tractable, even for massive real-world datasets. In this paper, we present an outline of our *generalized spatial scan* framework. We then consider each of the steps in more detail, giving some idea of the relevant decisions that need to be made when applying the spatial scan to a new domain. In [8], we present our experiences in one such domain (outbreak detection using over-the-counter drug sales data); here we discuss the method more generally, considering those issues which apply to any domain.

2. THE GENERALIZED SPATIAL SCAN
Our *generalized spatial scan* framework consists of the following six steps:

1) Obtain data for a set of spatial locations s_i.

AD-KDD'05, August 21, 2005, Chicago, Illinois, USA.

2) Choose a set of spatial regions to search over, where each spatial region S consists of a set of spatial locations s_i.

3) Choose models of the data under H_0 (the null hypothesis of no clusters) and $H_1(S)$ (the alternative hypothesis assuming a cluster in region S).

4) Derive a "score function" $F(S)$ based on $H_1(S)$ and H_0.

5) Find the "most interesting" regions, i.e. those regions S with the highest values of $F(S)$.

6) Determine whether each of these regions is "interesting," either by performing significance testing or calculating posterior probabilities.

We now consider each step of this framework in detail.

1) Obtain data for a set of spatial locations s_i.

For each spatial location s_i, we are given a *count* c_i and optionally a *baseline* b_i. For example, each s_i may represent a zip code, with location (latitude and longitude) assumed to be at the centroid of the zip code; c_i may represent the number of respiratory disease cases in that zip code, and b_i may represent the at-risk population. In any case, the goal of our method is to find regions where the counts are higher than expected, given the baselines. Two typical approaches are the *population-based* method, where b_i represents the underlying population of location s_i, and we expect each count to be proportional to its population under the null hypothesis, and the *expectation-based* method, where b_i represents the expected count of location s_i, and thus we expect each count to be equal to its expectation under the null. In either case, the b_i for each location may either be given (e.g. census population) or may be inferred from the time series of past counts. For example, one simple expectation-based approach would be to estimate today's expected count in a zip code by the mean daily count in that zip code over the past d days. For many datasets, more complicated methods of time series analysis should be used to infer baselines; for example, in the over-the-counter drug sales data, we must account for both seasonal and day-of-week effects. We consider various methods of inferring baselines in [3].

2) Choose a set of spatial regions to search over, where each spatial region S consists of a set of spatial locations s_i.

We want to choose a set of regions that corresponds well with the shape and size of the clusters we are interested in detecting. In general, the set of regions should cover the entire space under consideration (otherwise we will have no power to detect clusters in non-covered areas) and adjacent regions should overlap (otherwise we will have reduced power to detect clusters that lie partly in one region and partly in another). We typically consider the set of all regions of some fixed shape (e.g. circle, ellipse, rectangle) and varying size; what shape to choose depends on both statistical and computational considerations. If we search too few regions, we will have reduced power to detect clusters that do not closely match any of the regions searched; for example, if we search over square or circular regions, we will have low power to detect highly elongated clusters. On the other hand, if we search too many regions, our power to detect any particular subset of these regions is reduced because of multiple hypothesis testing. Additionally, the runtime of the algorithm is proportional to the number of regions searched, and thus choosing too large a set of regions will make the method computationally infeasible.

Our typical approach in epidemiological domains is to map the spatial locations to a grid, and search over the set of all rectangular regions on the grid. Additionally, non-axis-aligned rectangles can be detected by searching over multiple rotations of the data. The two main advantages of this approach are its ability to detect elongated clusters (this is important in epidemiology because disease clusters may be elongated due to wind or water dispersion of pathogens) and also its computational efficiency. Use of a grid structure allows us to evaluate any rectangular region in constant time, independent of the size of the region, using the well-known "cumulative counts" trick [4]. Additionally, we can gain huge computational speedups by applying the "fast spatial scan" algorithm [4-6], as we discuss below.

3) Choose models of the data under H_0 (the null hypothesis of no clusters) and $H_1(S)$ (the alternative hypothesis assuming a cluster in region S).

4) Derive a "score function" $F(S)$ based on $H_1(S)$ and H_0.

These are perhaps the most difficult steps in our method, as we must choose models which are both efficiently computable and relevant to the application domain under consideration. For our models to be *efficiently computable*, the score function $F(S)$ should be computable as a function of some additive sufficient statistics of the region S being considered (typically these statistics are the total count of the region, $C(S) = \sum_S c_i$, and the total baseline of the region, $B(S) = \sum_S b_i$). If this is not the case, the model may still be useful for small datasets, but will not scale well to larger sources of data. For our models to be *relevant*, any simplifying assumptions that we make must not reduce our power to distinguish between the "cluster" and "no cluster" cases, to too great an extent. Of course, any efficiently computable model is very unlikely to capture all of the complexity of the real data, and these unmodeled effects may have either small or large impacts on detection performance. Thus we typically use an iterative design process, beginning with very simple models, and examining their detection power (ability to distinguish between "cluster" and "no cluster") and calibration (number of false positives reported in day-to-day use). If a model has high detection power but poor calibration, then we have a choice between increasing model complexity and artificially recalibrating the model (i.e. based on the empirical distribution of scores); however, if detection power is low, then we have no choice but to figure out which unmodeled effects are harming performance, and deal with these effects one by one. Some such effects (e.g. missing data) can be dealt with by pre-processing, and others (e.g. clusters caused by single locations) can be dealt with by post-processing (filtering the set of discovered regions to remove those caused by known effects), while others must actually be included in the model itself. In [8], we discuss several of these effects present in the over-the-counter sales data, and how we have dealt with each; here we focus on the general framework and then present two simple and efficiently computable models.

The most common statistical framework for the spatial scan is a frequentist, hypothesis testing approach. In this approach, assuming that the null hypothesis and each alternative

hypothesis are point hypotheses (with no free parameters), we can use the likelihood ratio $F(S) = \frac{\Pr(Data \mid H_1(S))}{\Pr(Data \mid H_0)}$ 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: $F(S) = \frac{\max_{\theta_1(S) \in \Theta_1(S)} \Pr(Data \mid H_1(S), \theta_1(S))}{\max_{\theta_0 \in \Theta_0} \Pr(Data \mid H_0, \theta_0)}$. In many cases, such as in Kulldorff's statistic [1], this will lead to an *individually most powerful* statistical test under the given model assumptions. We then perform randomization testing using the maximum likelihood estimates of the parameters under the null hypothesis, as discussed below. In the *marginal likelihood* framework, on the other hand, we instead average over the possible values of each parameter:

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

This, however, makes randomization testing very difficult. A third alternative (discussed in detail in [7]) is a Bayesian approach, in which we use the marginal likelihood framework to compute the likelihood of the data under each hypothesis, then combine these likelihoods with the prior probabilities of an cluster in each region S. Thus our test statistic is the posterior probability of a cluster in each region: $F(S) = \frac{\Pr(Data \mid H_1(S)) \Pr(H_1(S))}{\Pr(Data)}$. The marginal likelihood of the data is typically difficult to compute, but in [7], we present an efficiently computable Bayesian statistic using Poisson counts and conjugate Gamma priors. Here we instead focus on the simpler, maximum likelihood frequentist approach, and give an example of how new scan statistics can be derived.

Let us first consider the expectation-based scan statistic discussed above, under the simplifying assumption that counts are independently Poisson distributed (i.e. counts are not spatially correlated, and neither overdispersed nor underdispersed). In this case, 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 region S has counts significantly greater than baselines. Furthermore, let us consider a simple cluster model, where we assume a uniform multiplicative increase in counts inside the cluster (the amount of increase is unknown). Thus 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.

$H_1(S)$: $c_i \sim \text{Poisson}(qb_i)$ 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:

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

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:

$$F(S) = \left(\frac{C}{B}\right)^C \exp(B - C), \text{ if } C > B, \text{ and } F(S) = 1 \text{ otherwise.}$$

Because $F(S)$ is a function only of the sufficient statistics $C(S)$ and $B(S)$, this function is efficiently computable: we can calculate the score of any region S by first calculating the aggregate count and baseline (in constant time, as noted above) and then applying the function F.

Kulldorff's spatial scan statistic [1] is a population-based method commonly used in disease surveillance, which also makes the simplifying assumption of independent, Poisson distributed counts. However, this statistic assumes that counts (i.e. number of disease cases) are distributed as $c_i \sim \text{Poisson}(qb_i)$, where b_i is the (known) census population of s_i and q is the (unknown) underlying disease rate. We then attempt to discover spatial regions where the underlying disease rate q is significantly higher inside the region than outside. 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 S." More precisely, we have:

H_0: $c_i \sim \text{Poisson}(q_{all} b_i)$ for all locations s_i, for some constant q_{all}.

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

In this case, the alternative hypothesis has two free parameters (q_{in} and q_{out}) and the null hypothesis has one free parameter (q_{all}). Computing the likelihood ratio, and using maximum likelihood parameter estimates, we obtain:

$$F(S) = \frac{\max_{q_{in} > q_{out}} \prod_{s_i \in S} \Pr(c_i \sim \text{Poisson}(q_{in} b_i)) \prod_{s_i \notin S} \Pr(c_i \sim \text{Poisson}(q_{out} b_i))}{\max_{q_{all}} \prod_{s_i} \Pr(c_i \sim \text{Poisson}(q_{all} b_i))}$$

We can compute the maximum likelihood estimates $q_{in} = C_{in} / B_{in}$, $q_{out} = C_{out} / B_{out}$, and $q_{all} = C_{all} / B_{all}$, where "in", "out", and "all" represent the aggregates of counts and baselines for s_i inside region S, for s_i outside region S, and for all s_i respectively. Plugging in these values and performing some algebra, we obtain: $F(S) = \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}} \left(\frac{C_{all}}{B_{all}}\right)^{-C_{all}}$ if $\frac{C_{in}}{B_{in}} > \frac{C_{out}}{B_{out}}$, and $F(S) = 1$ otherwise. Again, the score function can be computed efficiently from the sufficient statistics of region S.

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. In this case, the score function is still

efficiently computable, as a function of the sufficient statistics $B = \sum \frac{\mu_i}{\sigma_i^2} \mu_i$ and $C = \sum \frac{\mu_i}{\sigma_i^2} c_i$. Many other likelihood ratio scan statistics are possible, including models with simultaneous attacks in multiple regions and models with spatially varying (rather than uniform) 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 relevant to the application domain under consideration.

5) Find the "most interesting" regions, i.e. those regions *S* with the highest values of *F(S)*.

Once we have decided on a set of regions *S* to search, and derived a score function *F(S)*, the "most interesting" regions are those that maximize *F(S)*. In the frequentist spatial scan framework, these are the most significant spatial regions; in the Bayesian framework, these are the regions with highest posterior probabilities. The simplest method of finding the most interesting regions is to compute the score function *F(S)* for every region. An alternative to this naïve approach is to use the *fast spatial scan* algorithms of [4-6], 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, i.e. those with the highest scores *F(S)*, we do not need to search a region *S* if we can prove 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; we discuss these methods in detail in [4-6].

6) Determine whether each of these regions is "interesting," either by performing significance testing or calculating posterior probabilities.

For the frequentist approach, once we have found the highest scoring region S^* and its score $F^* = F(S^*)$, we must still determine the statistical significance of this region by randomization testing. To do so, we randomly create a large number R of replica grids by sampling under the null hypothesis, given our maximum likelihood parameter estimates for the null. For example, for the expectation-based approach given above, we generate counts independently from $c_i \sim \text{Poisson}(b_i)$, and for the population-based approach given above, we generate counts independently from $c_i \sim \text{Poisson}(q_{all} b_i)$, using the maximum likelihood estimate $q_{all} = C_{all} / B_{all}$. We then find the highest scoring region and its score for each replica grid: the *p*-value of S^* is $\frac{R_{beat} + 1}{R + 1}$, where R_{beat} is the number of replicas with F^* higher than the original grid. If this *p*-value is less than some threshold (e.g. 0.05), we can conclude that the discovered region is unlikely to have occurred by chance, and is thus a significant spatial cluster; we can then examine secondary clusters. Otherwise, no significant clusters exist.

For the Bayesian approach, on the other hand, no randomization testing is necessary. Instead, we can compute the posterior probability of each potential cluster by dividing its score $\Pr(Data \mid H_1(S)) \Pr(H_1(S))$ by the total probability $\Pr(Data) = \Pr(Data \mid H_0) \Pr(H_0) + \sum_S \Pr(Data \mid H_1(S)) \Pr(H_1(S))$. We can then report all clusters with posterior probability greater than some predetermined threshold, or simply "sound the alarm" if the total posterior probability of all clusters *S* is sufficiently high. Because we do not need to perform randomization testing in the Bayesian method, we need only to search over all regions for the original grid, rather than the original grid and a large number (typically $R = 1000$) of replicas. Thus the Bayesian approach is approximately 1000x faster than the (naïve) frequentist approach, as we show empirically in [7]. However, we can apply the fast spatial scan described above to achieve similar speedups for the frequentist approach: in this case, we still have to search over all replica grids, but can do a much faster search on each. As a result, the fast frequentist approach is faster than the Bayesian approach for sufficiently large grid sizes ($N > 256$) but slower for smaller grids. Either method can search a 256×256 grid, and calculate significance (*p*-values or posteriors respectively) in 10-12 hours, as compared to months for the standard (naïve frequentist) approach. Thus we now have two ways to make the spatial scan computationally feasible for large datasets: to apply the fast spatial scan of [4-6] or to use the Bayesian framework of [7]. For even larger grid sizes, it may be possible to extend the fast spatial scan to the Bayesian framework: this would give us the best of both worlds, searching only a single grid, and using a fast algorithm to do so. We are currently investigating this potentially useful synthesis.

3. REFERENCES

[1] M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods* **26**(6), 1481-1496, 1997.

[2] D.B. Neill and A.W. Moore. Methods for detection of spatial and spatio-temporal clusters. In M. Wagner et al., eds., *Handbook of Biosurveillance*, 2005.

[3] D.B. Neill, A.W. Moore, M. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. Accepted to *11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2005.

[4] D.B. Neill and A.W. Moore. Rapid detection of significant spatial clusters. *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 256-265, 2004.

[5] D.B. Neill, A.W. Moore, F. Pereira, and T. Mitchell. Detecting significant multidimensional spatial clusters. *Advances in Neural Information Processing Systems* **17**, 969-976, 2005.

[6] D.B. Neill and A.W. Moore. A fast multi-resolution method for detection of significant spatial disease clusters. *Advances in Neural Information Processing Systems* **16**, 651-658, 2004.

[7] D.B. Neill, A.W. Moore, and G.F. Cooper. A Bayesian spatial scan statistic. Submitted for publication.

[8] M.R. Sabhnani, D.B. Neill, A.W. Moore, F.-C. Tsui, M.M. Wagner, and J.U. Espino. Detecting anomalous patterns in pharmacy retail data. *KDD Workshop on Data Mining Methods for Anomaly Detection*, 2005.