Detecting Space-Time Clusters: Prior Work and New Directions

Daniel B. Neill Andrew W. Moore

November 2004
CMU-CS-05-115

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

The problem of space-time cluster detection arises in a variety of applications, including disease surveillance and brain imaging. In this work, we briefly review the state of the art in space-time cluster detection, focusing on space-time scan statistics, and we derive a number of new statistics. First, we distinguish between tests for clusters with higher disease rates inside the cluster than outside (as in the traditional spatial scan statistics framework) and tests for clusters with higher counts than expected (as is appropriate when inferring the expected counts in a region from the time series of past counts). Second, we distinguish between tests for "persistent" clusters (where the disease rate remains constant throughout the duration of a cluster) and tests for "emerging" clusters (where the disease rate increases monotonically through the duration of a cluster). These new statistics for spatio-temporal cluster detection will serve as the basis for our future work in detection of emerging space-time clusters.

Keywords: algorithms, biosurveillance, cluster detection, space-time scan statistics

1 Introduction

The problem of detecting space-time clusters arises in a variety of applications, including disease surveillance and brain imaging. In general, spatio-temporal methods can be divided into three classes: spatial modeling techniques such as "disease mapping," where observed values are spatially smoothed to infer the distribution of values in space-time (e.g. Clayton and Kaldor, 1987; Besag et al., 1991); tests for a general tendency of the data to cluster (e.g. Knox, 1964; Mantel, 1967); and tests which attempt to infer the location of clusters (e.g. Kulldorff et al., 1998; Kulldorff, 2001; Kulldorff et al., 2004). We focus on the latter class of methods, since these are the only methods which allow us to both answer whether any significant clusters exist, and if so, identify these clusters.

Let us assume that we have a set of data collected at a set of discrete time steps $k = 1 \dots k_{base}$, and at a set of discrete spatial locations s_i. For each s_i at each time step k, we are given a *count* c_i^k and (optionally) a *baseline* b_i^k. For example, in epidemiology, the counts may be the number of disease cases in a given spatial region over a given time interval, or some related observable quantity such as the number of Emergency Department visits or OTC drug sales. The baselines may be *given* (based on results from a control group, or an at-risk population derived from census data), or may be *inferred* based on the time series of counts. In all cases, we assume that counts c_i^k are generated by some distribution with mean proportional to $b_i^k q_i^k$, where q_i^k is the *rate* (or expected ratio of count to baseline). Our goal, then, is to find whether there is any region (set of locations s_i) and time interval ($k = k_{min} \dots k_{max}$) for which the rates are significantly higher than expected; in epidemiology, this may correspond to a disease outbreak. Within this very general framework, there are a number of questions we can ask:

1. Which spatial regions to search over? We typically search over the set of all regions of some given shape and variable size. For simplicity, we assume here that the spatial locations s_i are aggregated to a d-dimensional grid, and search over the set of all d-dimensional hyper-rectangular regions on the grid.

2. Which temporal intervals to search over? For *prospective* analysis, we search only over intervals ending at the present time, while for *retrospective* analysis we search over all intervals including those ending before the present time.

3. What distributions are assumed? For simplicity, we assume that c_i^k are generated independently from Poisson distributions with mean $q_i^k b_i^k$. We could also take other factors such as extra-Poisson variation (overdispersion) and spatial correlation into account; we do this somewhat in the CATS and RATS methods discussed below, since these perform aggregation of counts at the level of grid cells and regions respectively. In the BATS method discussed below, which considers a separate time series for each building, we do not account for correlation. We can also use Normal distributions instead of Poisson to model distributions with dispersion different from the mean and spatially varying.

4. Do we want to infer baselines from the time series of previous counts, or are the baselines given? For the time being, we assume that baselines are given; we discuss methods of inferring baselines from previous counts in Section 3.

In any case, the value of the space-time statistic D_{max} is taken to be the maximum over all spatial regions $S \subseteq G$ of $D(S)$, where $D(S)$ is the maximum $D_{k_{min}}^{k_{max}}(S)$ for all temporal intervals $k = k_{min} \ldots k_{max}$. For retrospective analysis, we have $1 \le k_{min} \le k_{max} \le k_{base}$; for prospective analysis, we have $1 \le k_{min} \le k_{max} = k_{base}$.

Now, in order to decide which statistic $D_{k_{min}}^{k_{max}}(S)$ to use, we must first decide what sort of regions we are looking for. In particular:

1. Do we want to detect regions such that the rates c_i^k / b_i^k are significantly higher than some prior expectation q_0, or such that they are significantly higher inside the region than outside? We call the former "globally sensitive" tests, since they are sensitive to global increases in rate. For the latter, we must decide whether to adjust for the overall global rate ("globally adaptive" tests) or to adjust separately for each day's rate ("daily adaptive" tests).

2. Do we expect the rate to be *constant* over the time duration of a cluster, or do we expect the rate to be *increasing* over the time duration of a cluster? In the first case, we have a test for *persistent* clusters, while in the second case, we have a test for *emerging* clusters. We can also make several other assumptions, such as a rate increasing according to some parametrized distribution (ex. linear increase, exponential increase).

Based on our answers to these two questions, we may define a number of different statistics, as defined in Section 2.

2 Space-time statistics

We first consider the case of a simple prospective space-time scan statistic, where we want to detect only if there are any space-time clusters on the present day $k = k_{base}$. In this case, we have the same statistics whether we assume that the cluster is persistent, emerging, etc.; the only relevant question is whether our test is globally sensitive, globally adaptive, or daily adaptive. After that, we present the more general space-time scan statistics for persistent clusters, emerging clusters, and parametrized clusters in turn.

2.1 1-day clusters

2.1.1 Globally sensitive

In this case, we compare the null hypothesis H_0: the rate equals q_0 over all locations and times, to the alternative hypothesis $H_1(S)$: the rate is higher than q_0 at the present time k in region S, and equals q_0 over all other locations and times. The likelihood ratio is:

$$D(S) = \frac{\max_{q_{in} \ge q_0} \prod_{s_i \in S} \Pr(c_i^k \sim \text{Po}(b_i^k q_{in}))}{\prod_{s_i \in S} \Pr(c_i^k \sim \text{Po}(b_i^k q_0))}$$

$$= \frac{\max_{q_{in} \ge q_0} \prod_{s_i \in S} e^{-b_i^k q_{in}} {q_{in}}^{c_i^k}}{\prod_{s_i \in S} e^{-b_i^k q_0} {q_0}^{c_i^k}}$$

$$= \frac{\max_{q_{in} \geq q_0} e^{-B_{in} q_{in}} {q_{in}}^{C_{in}}}{e^{-B_{in} q_0} {q_0}^{C_{in}}}$$

where "in" are sums over region S at time k. This quantity is maximized at $q_{in} = \frac{C_{in}}{B_{in}}$, assuming this quantity is greater than q_0; otherwise we have $D(S) = 1$. In the former case, we have:

$$D(S) = \frac{e^{-C_{in}} \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}}}{e^{-B_{in} q_0} {q_0}^{C_{in}}} = \left(\frac{C_{in}}{q_0 B_{in}}\right)^{C_{in}} e^{q_0 B_{in} - C_{in}}$$

Note that prior days are not accounted for, except possibly in computing the baselines; this statistic is exactly identical to the globally sensitive, purely spatial scan statistic.

To compute the p-value, we compare D_{max} of the original grid to D_{max} of a large number of replica grids, where a replica grid has all counts c_i^k generated from $\mathrm{Po}(q_0 b_i^k)$. Note that prior days need not be regenerated, since these do not impact the score.

2.1.2 Globally adaptive

In this case, we compare the null hypothesis H_0: the rate equals q_{all} over all locations and times, to the alternative hypothesis $H_1(S)$: the rate equals q_{in} at the present time k in region S, and equals q_{out} over all other locations and times, $q_{in} > q_{out}$. The likelihood ratio is:

$$D(S) = \frac{\max_{q_{in} \geq q_{out}} \prod_{s_i^k \in S \times k} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{in})) \prod_{s_i^k \in G \times (1 \ldots k_{base}) - S \times k} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{out}))}{\max_{q_{all}} \prod_{s_i^k \in G \times (1 \ldots k_{base})} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{all}))}$$

$$= \frac{\max_{q_{in} \geq q_{out}} \prod_{s_i^k \in S \times k} e^{-b_i^k q_{in}} {q_{in}}^{c_i^k} \prod_{s_i^k \in G \times (1 \ldots k_{base}) - S \times k} e^{-b_i^k q_{out}} {q_{out}}^{c_i^k}}{\max_{q_{all}} \prod_{s_i^k \in G \times (1 \ldots k_{base})} e^{-b_i^k q_{all}} {q_{all}}^{c_i^k}}$$

$$= \frac{\max_{q_{in} \geq q_{out}} e^{-B_{in} q_{in}} {q_{in}}^{C_{in}} e^{-B_{out} q_{out}} {q_{out}}^{C_{out}}}{\max_{q_{all}} e^{-B_{all} q_{all}} {q_{all}}^{C_{all}}}$$

where "in" are sums over region S at time k, "all" are sums over all space and time, and "out" are sums over all space and time except $S \times k$. This quantity is maximized at $q_{in} = \frac{C_{in}}{B_{in}}$, $q_{out} = \frac{C_{out}}{B_{out}}$, and $q_{all} = \frac{C_{all}}{B_{all}}$, assuming $\frac{C_{in}}{B_{in}} > \frac{C_{out}}{B_{out}}$; otherwise we have $D(S) = 1$. In the former case, we have:

$$D(S) = \frac{e^{-C_{in}} \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} e^{-C_{out}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}}}{e^{-C_{all}} \left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}} = \frac{\left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}}}{\left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}}$$

Note that, even though this is identical in appearance to the globally adaptive, purely spatial scan statistic, it is different since the "out" and "all" are defined including counts/baselines from all previous days in addition to the counts/baselines outside the region on the current day.

To compute the p-value, we compare D_{max} of the original grid to D_{max} of a large number of replica grids, where a replica grid has all counts c_i^k generated from $\mathrm{Po}(q_{all} b_i^k)$. Again, prior days need not be regenerated, though regenerating them may impact the score. We can also generate replicas by permuting the counts in space-time.

2.1.3 Daily adaptive

In this case, we compare the null hypothesis H_0: for each day k, the rate equals $q_{all,k}$ for all locations, to the alternative hypothesis $H_1(S)$: for the present day k, the rate equals q_{in} inside the region and q_{out} outside, $q_{in} > q_{out}$, and for all other days k the rate equals $q_{all,k}$ for all locations. The likelihood ratio is:

$$D(S) = \frac{\max_{q_{in} \geq q_{out}} \prod_{s_i^k \in S \times k} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{in})) \prod_{s_i^k \in (G-S) \times k} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{out}))}{\max_{q_{all}} \prod_{s_i^k \in G \times k} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{all}))}$$

$$= \frac{\max_{q_{in} \geq q_{out}} \prod_{s_i^k \in S \times k} e^{-b_i^k q_{in}} {q_{in}}^{c_i^k} \prod_{s_i^k \in (G-S) \times k} e^{-b_i^k q_{out}} {q_{out}}^{c_i^k}}{\max_{q_{all}} \prod_{s_i^k \in G \times k} e^{-b_i^k q_{all}} {q_{all}}^{c_i^k}}$$

$$= \frac{\max_{q_{in} \geq q_{out}} e^{-B_{in} q_{in}} {q_{in}}^{C_{in}} e^{-B_{out} q_{out}} {q_{out}}^{C_{out}}}{\max_{q_{all}} e^{-B_{all} q_{all}} {q_{all}}^{C_{all}}}$$

where "in" are sums over region S at time k, "all" are sums over all space at time k, and "out" are sums over $G-S$ at time k. This quantity is maximized at $q_{in} = \frac{C_{in}}{B_{in}}$, $q_{out} = \frac{C_{out}}{B_{out}}$, and $q_{all} = \frac{C_{all}}{B_{all}}$, assuming $\frac{C_{in}}{B_{in}} > \frac{C_{out}}{B_{out}}$; otherwise we have $D(S) = 1$. In the former case, we have:

$$D(S) = \frac{e^{-C_{in}} \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} e^{-C_{out}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}}}{e^{-C_{all}} \left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}} = \frac{\left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}}}{\left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}}$$

In this case, the "out" and "all" are defined only using counts/baselines from the present day, so this is identical to the globally adaptive, purely spatial scan statistic.

To compute the p-value, we compare D_{max} of the original grid to D_{max} of a large number of replica grids, where a replica grid has all counts c_i^k generated from $\mathrm{Po}(q_{all,k} b_i^k)$. Again, prior days' counts need not be regenerated, since regenerating them will not impact the score. We can also generate replicas by permuting the current day's counts in space.

2.2 Persistent clusters

The tests for persistent clusters assume that the rate of a cluster remains constant over time; as a result, the derivations are almost identical to the 1-day cluster tests, with sums taken over the entire duration of a cluster rather than only a single day.

2.2.1 Globally sensitive

In this case, we compare the null hypothesis H_0: the rate equals q_0 over all locations and times, to the alternative hypothesis $H_1(S)$: the rate is $q_{in} > q_0$ at times $k_{min} \ldots k_{max}$ in region S, and equals q_0 over all other locations and times. The likelihood ratio is:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{q_{in} \geq q_0} \prod_{s_i^k \in S \times (k_{min} \ldots k_{max})} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{in}))}{\prod_{s_i^k \in S \times (k_{min} \ldots k_{max})} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_0))}$$

$$= \frac{\max_{q_{in} \geq q_0} \prod_{s_i^k \in S \times (k_{min} \ldots k_{max})} e^{-b_i^k q_{in}} {q_{in}}^{c_i^k}}{\prod_{s_i^k \in S \times (k_{min} \ldots k_{max})} e^{-b_i^k q_0} {q_0}^{c_i^k}}$$

$$= \frac{\max_{q_{in} \geq q_0} e^{-B_{in} q_{in}} {q_{in}}^{C_{in}}}{e^{-B_{in} q_0} {q_0}^{C_{in}}}$$

where "in" are sums over region S at times $k_{min} \ldots k_{max}$. This quantity is maximized at $q_{in} = \frac{C_{in}}{B_{in}}$, assuming this quantity is greater than q_0; otherwise we have $D_{k_{min}}^{k_{max}}(S) = 1$. In the former case, we have:

$$D_{k_{min}}^{k_{max}}(S) = \frac{e^{-C_{in}} \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}}}{e^{-B_{in} q_0} {q_0}^{C_{in}}} = \left(\frac{C_{in}}{q_0 B_{in}}\right)^{C_{in}} e^{q_0 B_{in} - C_{in}}$$

To compute the p-value, we compare D_{max} of the original grid to D_{max} of a large number of replica grids, where a replica grid has all counts c_i^k generated from $\mathrm{Po}(q_0 b_i^k)$.

2.2.2 Globally adaptive

In this case, we compare the null hypothesis H_0: the rate equals q_{all} over all locations and times, to the alternative hypothesis $H_1(S)$: the rate equals q_{in} at the times $k_{min} \ldots k_{max}$ in region S, and equals q_{out} over all other locations and times, $q_{in} > q_{out}$. The likelihood ratio is:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{q_{in} \geq q_{out}} \prod_{s_i^k \in S \times (k_{min} \ldots k_{max})} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{in})) \prod_{s_i^k \in out} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{out}))}{\max_{q_{all}} \prod_{s_i^k \in G \times (1 \ldots k_{base})} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{all}))}$$

$$= \frac{\max_{q_{in} \geq q_{out}} \prod_{s_i^k \in S \times (k_{min} \ldots k_{max})} e^{-b_i^k q_{in}} {q_{in}}^{c_i^k} \prod_{s_i^k \in out} e^{-b_i^k q_{out}} {q_{out}}^{c_i^k}}{\max_{q_{all}} \prod_{s_i^k \in G \times (1 \ldots k_{base})} e^{-b_i^k q_{all}} {q_{all}}^{c_i^k}}$$

$$= \frac{\max_{q_{in} \geq q_{out}} e^{-B_{in} q_{in}} {q_{in}}^{C_{in}} e^{-B_{out} q_{out}} {q_{out}}^{C_{out}}}{\max_{q_{all}} e^{-B_{all} q_{all}} {q_{all}}^{C_{all}}}$$

where "in" are sums over region S at times $k_{min} \ldots k_{max}$, "all" are sums over all space and time, and "out" are sums over all space and time except $S \times (k_{min} \ldots k_{max})$. This quantity is maximized at $q_{in} = \frac{C_{in}}{B_{in}}$, $q_{out} = \frac{C_{out}}{B_{out}}$, and $q_{all} = \frac{C_{all}}{B_{all}}$, assuming $\frac{C_{in}}{B_{in}} > \frac{C_{out}}{B_{out}}$; otherwise we have $D_{k_{min}}^{k_{max}}(S) = 1$. In the former case, we have:

$$D_{k_{min}}^{k_{max}}(S) = \frac{e^{-C_{in}} \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} e^{-C_{out}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}}}{e^{-C_{all}} \left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}} = \frac{\left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}}}{\left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}}$$

To compute the p-value, we compare D_{max} of the original grid to D_{max} of a large number of replica grids, where a replica grid has all counts c_i^k generated from $\mathrm{Po}(q_{all} b_i^k)$. We can also generate replicas by permuting the counts in space-time.

2.2.3 Daily adaptive

In this case, we compare the null hypothesis H_0: for each day k, the rate equals $q_{all,k}$ for all locations, to the alternative hypothesis $H_1(S)$: for days $k_{min} \ldots k_{max}$, the rate equals $q_{in,k}$ inside the region and $q_{out,k}$ outside, $q_{in,k}/q_{out,k} = \theta \geq 1$, and for all other days k, the rate equals $q_{all,k}$ for all locations. The likelihood ratio is:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{\theta \geq 1} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} \prod_{s_i^k \in S \times k} \Pr(c_i^k \sim \mathrm{Po}(b_i^k \theta q_{out,k})) \prod_{s_i^k \in (G-S) \times k} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{out,k}))}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} \prod_{s_i^k \in G \times k} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{all,k}))}$$

$$= \frac{\max_{\theta \geq 1} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} \prod_{s_i^k \in S \times k} e^{-b_i^k \theta q_{out,k}} \left(\theta q_{out,k}\right)^{c_i^k} \prod_{s_i^k \in (G-S) \times k} e^{-b_i^k q_{out,k}} q_{out,k}{}^{c_i^k}}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} \prod_{s_i^k \in G \times k} e^{-b_i^k q_{all,k}} q_{all,k}{}^{c_i^k}}$$

$$= \frac{\max_{\theta \geq 1} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} e^{-B_{in,k} \theta q_{out,k}} \left(\theta q_{out,k}\right)^{C_{in,k}} e^{-B_{out,k} q_{out,k}} q_{out,k}{}^{C_{out,k}}}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} e^{-B_{all,k} q_{all,k}} q_{all,k}{}^{C_{all,k}}}$$

where "in,k" are sums over region S at time k, "all,k" are sums over all space at time k, and "out,k" are sums over $G - S$ at time k. The numerator and denominator are maximized at $q_{out,k} = \frac{C_{all,k}}{\theta B_{in,k} + B_{out,k}}$ and $q_{all,k} = \frac{C_{all,k}}{B_{all,k}}$ respectively, giving us:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{\theta \geq 1} \prod_{k=k_{min}}^{k_{max}} e^{-C_{all,k}} \theta^{C_{in,k}} \left(\frac{C_{all,k}}{\theta B_{in,k} + B_{out,k}}\right)^{C_{all,k}}}{\prod_{k=k_{min}}^{k_{max}} e^{-C_{all,k}} \left(\frac{C_{all,k}}{B_{all,k}}\right)^{C_{all,k}}}$$

$$= \max_{\theta \geq 1} \prod_{k=k_{min}}^{k_{max}} \theta^{C_{in,k}} \left(\frac{B_{all,k}}{\theta B_{in,k} + B_{out,k}}\right)^{C_{all,k}}$$

Maximizing with respect to θ requires finding the root of a polynomial of degree $k_{max} - k_{min} + 1$; approximate (gradient) methods may also be used.

To compute the p-value, we compare D_{max} of the original grid to D_{max} of a large number of replica grids, where a replica grid has all counts c_i^k generated from $\mathrm{Po}(q_{all,k} b_i^k)$. We can also generate replicas by permuting each day's counts in space.

2.3 Emerging clusters

While the tests for persistent clusters assume that the rate of a cluster remains constant over time, this is typically not true in domains such as epidemiology: when a disease outbreak occurs, disease rate will typically rise continually over the duration of the outbreak until the outbreak reaches its peak, at which point it will level off or decrease. Our main goal in the epidemiological domain is to detect *emerging* outbreaks (i.e. those that have not yet reached their peak), so we focus on finding clusters where the disease rate is monotonically increasing (i.e. non-decreasing) over the duration of the cluster.

2.3.1 Globally sensitive

In this case, we compare the null hypothesis H_0: the rate equals q_0 over all locations and times, to the alternative hypothesis $H_1(S)$: the rate is q_k at times $k = k_{min} \ldots k_{max}$ in region S (where the q_k are non-decreasing and at least q_0), and equals q_0 over all other locations and times. The likelihood ratio is:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{q_0 \le q_{k_{min}} \le \ldots \le q_{k_{max}}} \prod_{k=k_{min}}^{k_{max}} \prod_{s_i^k \in S \times k} \Pr(c_i^k \sim \mathrm{Po}(q_k b_i^k))}{\prod_{k=k_{min}}^{k_{max}} \prod_{s_i^k \in S \times k} \Pr(c_i^k \sim \mathrm{Po}(q_0 b_i^k))}$$

$$= \frac{\max_{q_0 \le q_{k_{min}} \le \ldots \le q_{k_{max}}} \prod_{k=k_{min}}^{k_{max}} \prod_{s_i^k \in S \times k} e^{-q_k b_i^k} (q_k)^{c_i^k}}{\prod_{k=k_{min}}^{k_{max}} \prod_{s_i^k \in S \times k} e^{-q_0 b_i^k} (q_0)^{c_i^k}}$$

$$= \frac{\max_{q_0 \le q_{k_{min}} \le \ldots \le q_{k_{max}}} \prod_{k=k_{min}}^{k_{max}} e^{-q_k B_{in,k}} (q_k)^{C_{in,k}}}{\prod_{k=k_{min}}^{k_{max}} e^{-q_0 B_{in,k}} (q_0)^{C_{in,k}}}$$

$$= \frac{\max_{q_0 \le q_{k_{min}} \le \ldots \le q_{k_{max}}} \prod_{k=k_{min}}^{k_{max}} e^{-q_k B_{in,k}} (q_k)^{C_{in,k}}}{e^{-q_0 B_{in}} (q_0)^{C_{in}}}$$

Now, we must maximize the numerator subject to the constraints on the q_k. To do so, let $E = E_1 \ldots E_p$ be a partitioning of $k_{min} \ldots k_{max}$ into sets of consecutive integers, such that 1) for all $k_1, k_2 \in E_i$, $q_{k_1} = q_{k_2} = Q_i$, and 2) for all E_{i_1}, E_{i_2}, where $i_1 < i_2$, $Q_{i_1} < Q_{i_2}$. In other words, the E_i define a partitioning of $k_{min} \ldots k_{max}$ into time periods where the disease rate is constant. Note that $q_{min} \ldots q_{max}$ are uniquely defined by the partitions $E = \{E_i\}$ and the rates $Q = \{Q_i\}$. A pair (E, Q) is *optimal* when the resulting q_k maximize $D_{k_{min}}^{k_{max}}(S)$. We can then write:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{E_1 \ldots E_p} \max_{Q_1 \ldots Q_p} \prod_{E_i} e^{-Q_i B_{in,i}} (Q_i)^{C_{in,i}}}{e^{-q_0 B_{in}} (q_0)^{C_{in}}}$$

where the in, i are the sums of the in, k for all $k \in E_i$.

Lemma 2.1 *A necessary condition for (E, Q) to be optimal is that for all i, $Q_i = \frac{C_{in,i}}{B_{in,i}}$.*

Proof Let us assume a fixed partitioning $E = \{E_i\}$, with strictly increasing Q_i, and ask whether the Q_i are optimal for those E_i. We note that, in the absence of constraints on the Q_i, each expression $e^{-Q_i B_{in,i}} (Q_i)^{C_{in,i}}$ is maximized at $Q_i = \frac{C_{in,i}}{B_{in,i}}$. Moreover, the score is convex with respect to Q_i. Thus, if some $Q_i < \frac{C_{in,i}}{B_{in,i}}$, we can increase the score by raising that Q_i slightly (without changing the ordering of Q_i), so the given Q_i are not optimal. Similarly, if some $Q_i > \frac{C_{in,i}}{B_{in,i}}$, we can increase the score by lowering that Q_i slightly (without changing the ordering of Q_i), so the given Q_i are not optimal. Thus for the Q_i to be optimal, we must have $Q_i = \frac{C_{in,i}}{B_{in,i}}$ for all i. ∎

Lemma 2.2 *A necessary condition for (E, Q) to be optimal is that for all $i_1 < i_2$, $\frac{C_{in,i_1}}{B_{in,i_1}} < \frac{C_{in,i_2}}{B_{in,i_2}}$.*

Proof Otherwise either $Q_{i_1} \neq \frac{C_{in,i_1}}{B_{in,i_1}}$, or $Q_{i_2} \neq \frac{C_{in,i_2}}{B_{in,i_2}}$, or $Q_{i_1} \geq Q_{i_2}$. In the first two cases, the condition of Lemma 2.1 is violated, so the Q_i are not optimal. In the third case, the restriction that the Q_i are strictly increasing is violated, so the Q_i are not legal. ▮

Thus we can write:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{E_1 \ldots E_p} \prod_{E_i} e^{-C_{in,i}} \left(\frac{C_{in,i}}{B_{in,i}}\right)^{C_{in,i}}}{e^{-q_0 B_{in}} (q_0)^{C_{in}}}$$

$$= e^{q_0 B_{in} - C_{in}} (q_0)^{-C_{in}} \max_{E_1 \ldots E_p} \prod_{E_i} \left(\frac{C_{in,i}}{B_{in,i}}\right)^{C_{in,i}}$$

where the partitioning $E = \{E_i\}$ must satisfy the condition of Lemma 2.2, i.e. the ratios $\frac{C_{in,i}}{B_{in,i}}$ are strictly increasing with i.

Lemma 2.3 *A necessary condition for the partitioning E to be optimal is that for each $E_i = k_1 \ldots k_2$, for all k such that $k_1 \leq k < k_2$, we have* $\frac{\sum_{j=k_1}^{k} c_j}{\sum_{j=k_1}^{k} b_j} \geq \frac{\sum_{j=k+1}^{k_2} c_j}{\sum_{j=k+1}^{k_2} b_j}$.

Proof Otherwise there exists some $E_i = k_1 \ldots k_2$, and some k such that $k_1 \leq k < k_2$, where $\frac{\sum_{j=k_1}^{k} c_j}{\sum_{j=k_1}^{k} b_j} < Q_i < \frac{\sum_{j=k+1}^{k_2} c_j}{\sum_{j=k+1}^{k_2} b_j}$ (note Q_i is a weighted average of the two ratios). We can now increase the score by separating E_i into two partitions $E_{i_1} = k_1 \ldots k$ and $E_{i_2} = k+1 \ldots k_2$, where Q_{i_1} is slightly less than Q_i, and Q_{i_2} is slightly more than Q_i (without otherwise changing the order of Q_i). Thus E is not optimal. ▮

Lemma 2.4 *A partitioning E satisfying the conditions of Lemmas 2.2 and 2.3 is unique, and thus that partitioning is optimal.*

Proof Assume two partitionings E^1 and E^2 satisfying the conditions of Lemmas 2.2 and 2.3. Consider the first i such that $E_i^1 \neq E_i^2$. Let $E_i^1 = k_0 \ldots k_1$ and $E_i^2 = k_0 \ldots k_2$, assuming without loss of generality that $k_1 > k_2$. Now consider the first $j > i$ such that $E_j^2 = k_3 \ldots k_4$ and $k_4 \geq k_1$. Thus we have $k_0 \leq k_2 < k_3 \leq k_1 \leq k_4$. Let us write $\mu(k_0 \ldots k_2) = \frac{\sum_{j=k_0}^{k_2} c_j}{\sum_{j=k_0}^{k_2} b_j}$ and define the other $\mu(\cdot)$ similarly. Applying the condition of Lemma 2.2 to E^2, we know $\mu(k_0 \ldots k_2) < \mu(k_3 \ldots k_4)$. Also, if $k_2 + 1 < k_3$, we know $\mu(k_0 \ldots k_2) < \mu(k_2 + 1 \ldots k_3 - 1) < \mu(k_3 \ldots k_4)$. Applying the condition of Lemma 2.3 to E^2, we know that if $k_1 < k_4$, we have $\mu(k_3 \ldots k_1) \geq \mu(k_3 \ldots k_4) \geq \mu(k_1 + 1 \ldots k_4)$. From these inequalities, we know $\mu(k_3 \ldots k_1) > \mu(k_0 \ldots k_3 - 1)$. But applying the condition of Lemma 2.3 to E^1, we know $\mu(k_0 \ldots k_3 - 1) \geq \mu(k_3 \ldots k_1)$, which is a contradiction. Thus the partitioning satisfying the conditions of Lemmas 2.2 and 2.3 is unique. Since these are necessary conditions for optimality, and a unique partitioning satisfies these conditions, we know that the partitioning is optimal. ▮

Finally, we give an algorithm which produces the optimal partitioning E. This "step method" uses a stack data structure, where each element of the stack represents a partition E_i by a 5-tuple (start, end, C, B, Q). The algorithm starts by pushing (k_{max}, k_{max}, $C_{k_{max}}$, $B_{k_{max}}$, $\max(q_0, \frac{C_{k_{max}}}{B_{k_{max}}})$) onto the stack. Then for each k, from $k_{max}-1$ down to k_{min}, we do the following:

```
temp = (k, k, C_k, B_k, max(q_0, C_k / B_k))
while (temp.Q >= stack.top.Q)
   temp2 = stack.pop
   temp = (temp.start, temp2.end, temp.C+temp2.C, temp.B+temp2.B,
      max(q_0, (temp.C+temp2.C)/(temp.B+temp2.B)))
stack.push(temp)
```

Theorem 2.5 *The step method produces the optimal partitioning E.*

Proof We first note that the method satisfies the conditions of Lemma 2.1 (since $Q_i = \frac{C_i}{B_i}$ for each partition E_i), and Lemma 2.2 (since the while loop ensures the ordering of Q_i). To show that the method satisfies the condition of Lemma 2.3, we show that each new partition created by the "merge step" temp = (temp.start, temp2.end, ...) maintains this condition as an invariant. Let $E_{temp} = k_0 \ldots k_1$, and $E_{temp2} = k_1+1 \ldots k_2$. We know that E_{temp} and E_{temp2} satisfy the condition of Lemma 2.3, and we must show that the merged partition E_{new} also satisfies this condition. In other words, we are given $\mu(k_0 \ldots i) \geq \mu(i+1 \ldots k_1)$ for all i ($k_0 \leq i < k_1$), and $\mu(k_1+1 \ldots i) \geq \mu(i+1 \ldots k_2)$ for all i ($k_1+1 \leq i < k_2$). We also know that temp.Q is at least temp2.Q, since the merge step only takes place if this condition holds, so $\mu(k_0 \ldots k_1) \geq \mu(k_1+1 \ldots k_2)$. To show that the merged partition satisfies the condition of Lemma 2.3, we must show that $\mu(k_0 \ldots i) \geq \mu(i+1 \ldots k_2)$ for all i ($k_0 \leq i < k_2$). We know this is true for $i = k_1$, but must also prove it for $i < k_1$ and $i > k_1$. For $i < k_1$, we have $\mu(k_0 \ldots i) \geq \mu(k_0 \ldots k_1) \geq \mu(i+1 \ldots k_1)$ and $\mu(k_0 \ldots k_1) \geq \mu(k_1+1 \ldots k_2)$. Thus $\mu(k_0 \ldots i) \geq \mu(i+1 \ldots k_1)$ and $\mu(k_0 \ldots i) \geq \mu(k_1+1 \ldots k_2)$, so $\mu(k_0 \ldots i) \geq \mu(i+1 \ldots k_2)$ as desired. For $i > k_1$, we have $\mu(k_1+1 \ldots i) \geq \mu(k_1+1 \ldots k_2) \geq \mu(i+1 \ldots k_2)$ and $\mu(k_0 \ldots k_1) \geq \mu(k_1+1 \ldots k_2)$. Thus $\mu(k_0 \ldots k_1) \geq \mu(i+1 \ldots k_2)$ and $\mu(k_1+1 \ldots i) \geq \mu(i+1 \ldots k_2)$, so $\mu(k_0 \ldots i) \geq \mu(i+1 \ldots k_2)$ as desired. ∎

2.3.2 Globally adaptive

In this case, we compare the null hypothesis H_0: the rate equals q_{all} over all locations and times, to the alternative hypothesis $H_1(S)$: the rate is q_k at times $k_{min} \ldots k_{max}$ in region S (where the q_k are non-decreasing and at least q_{out}), and equals q_{out} over all other locations and times. The likelihood ratio is:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{q_{out} \leq q_{k_{min}} \leq \ldots \leq q_{k_{max}}} \prod_{s_i^k \in S \times (k_{min} \ldots k_{max})} \Pr(c_i^k \sim \mathrm{Po}(q_k b_i^k)) \prod_{s_i^k \in out} \Pr(c_i^k \sim \mathrm{Po}(q_{out} b_i^k))}{\max_{q_{all}} \prod_{s_i^k \in G \times (1 \ldots k_{base})} \Pr(c_i^k \sim \mathrm{Po}(q_{all} b_i^k))}$$

$$= \frac{\max_{q_{out} \leq q_{k_{min}} \leq \ldots \leq q_{k_{max}}} \prod_{s_i^k \in S \times (k_{min} \ldots k_{max})} e^{-q_k b_i^k} (q_k)^{c_i^k} \prod_{s_i^k \in out} e^{-q_{out} b_i^k} (q_{out})^{c_i^k}}{\max_{q_{all}} \prod_{s_i^k \in G \times (1 \ldots k_{base})} e^{-q_{all} b_i^k} (q_{all})^{c_i^k}}$$

$$= \frac{\max_{q_{out} \le q_{k_{min}} \le \dots \le q_{k_{max}}} \prod_{k=k_{in}}^{k_{max}} e^{-q_k B_{in,k}} \left(q_k\right)^{C_{in,k}} \times e^{-q_{out} B_{out}} \left(q_{out}\right)^{C_{out}}}{\max_{q_{all}} e^{-q_{all} B_{all}} \left(q_{all}\right)^{C_{all}}}$$

$$= \frac{\max_{q_{out} \le q_{k_{min}} \le \dots \le q_{k_{max}}} \prod_{k=k_{in}}^{k_{max}} e^{-q_k B_{in,k}} \left(q_k\right)^{C_{in,k}} \times e^{-q_{out} B_{out}} \left(q_{out}\right)^{C_{out}}}{e^{-C_{all}} \left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}}$$

Now, we must again maximize the numerator subject to the constraints on the q_k. We can use almost the same step method as before, except that we do not need to enforce the constraint $q_k \ge q_0$, but we do need to set q_{out} such that $q_{out} \le q_{k_{min}}$. Thus we let $E = E_1 \dots E_p$ be a partitioning of $out \cup (k_{min} \dots k_{max})$ into sets of consecutive integers, such that 1) for all $k_1, k_2 \in E_i$, $q_{k_1} = q_{k_2} = Q_i$, and 2) for all E_{i_1}, E_{i_2}, where $i_1 < i_2$, $Q_{i_1} < Q_{i_2}$. We can then write:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{E_1 \dots E_p} \max_{Q_1 \dots Q_p} \prod_{E_i} e^{-Q_i B_i} \left(Q_i\right)^{C_i}}{e^{-C_{all}} \left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}}$$

$$= \frac{\max_{E_1 \dots E_p} \prod_{E_i} e^{-C_i} \left(\frac{C_i}{B_i}\right)^{C_i}}{e^{-C_{all}} \left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}}$$

$$= \frac{\max_{E_1 \dots E_p} \prod_{E_i} \left(\frac{C_i}{B_i}\right)^{C_i}}{\left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}}$$

where the C_i are the sums of the $C_{in,k}$ for all $k \in E_i$, plus C_{out} if $out \in E_i$, and similarly for the B_i.

To find the optimal partitioning E of $out \cup (k_{min} \dots k_{max})$, we use the following step method. First, we push $(k_{max}, k_{max}, C_{k_{max}}, B_{k_{max}}, \frac{C_{k_{max}}}{B_{k_{max}}})$ onto the stack. Then for each k, from $k_{max} - 1$ down to k_{min}, we do the following:

```
temp = (k, k, C_k, B_k, C_k / B_k)
while (temp.Q >= stack.top.Q)
   temp2 = stack.pop
   temp = (temp.start, temp2.end, temp.C+temp2.C, temp.B+temp2.B,
      (temp.C+temp2.C)/(temp.B+temp2.B))
stack.push(temp)
```

Finally, we do the same for the "out" partition, treating "out" as an arbitrary integer less than k_{min}:

```
temp = (out, out, C_out, B_out, C_out / B_out)
while (temp.Q >= stack.top.Q)
   temp2 = stack.pop
   temp = (temp.start, temp2.end, temp.C+temp2.C, temp.B+temp2.B,
      (temp.C+temp2.C)/(temp.B+temp2.B))
stack.push(temp)
```

Correctness of the step method follows from the same argument as above.

2.3.3 Daily adaptive

In this case, we compare the null hypothesis H_0: for each day k, the rate equals $q_{all,k}$ for all locations, to the alternative hypothesis $H_1(S)$: for days $k_{min} \ldots k_{max}$, the rate equals $q_{in,k}$ inside the region and $q_{out,k}$ outside, $q_{in,k}/q_{out,k} = \theta_k$ (θ_k non-decreasing and at least 1) and for all other days k, the rate equals $q_{all,k}$ for all locations. The likelihood ratio is:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{1 \leq \theta_{k_{min}} \leq \ldots \leq \theta_{k_{max}}} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} \prod_{s_i^k \in S} \Pr(c_i^k \sim \mathrm{Po}(\theta_k q_{out,k} b_i^k)) \prod_{s_i^k \in G-S} \Pr(c_i^k \sim \mathrm{Po}(q_{out,k} b_i^k))}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} \prod_{s_i^k \in G} \Pr(c_i^k \sim \mathrm{Po}(q_{all,k} b_i^k))}$$

$$= \frac{\max_{1 \leq \theta_{k_{min}} \leq \ldots \leq \theta_{k_{max}}} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} \prod_{s_i^k \in S} e^{-\theta_k q_{out,k} b_i^k} \left(\theta_k q_{out,k}\right)^{c_i^k} \prod_{s_i^k \in G-S} e^{-q_{out,k} b_i^k} \left(q_{out,k}\right)^{c_i^k}}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} \prod_{s_i^k \in G} e^{-q_{all,k} b_i^k} \left(q_{all,k}\right)^{c_i^k}}$$

$$= \frac{\max_{1 \leq \theta_{k_{min}} \leq \ldots \leq \theta_{k_{max}}} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} e^{-\theta_k q_{out,k} B_{in,k}} \left(\theta_k q_{out,k}\right)^{C_{in,k}} e^{-q_{out,k} B_{out,k}} \left(q_{out,k}\right)^{C_{out,k}}}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} e^{-q_{all,k} B_{all,k}} \left(q_{all,k}\right)^{C_{all,k}}}$$

$$= \frac{\max_{1 \leq \theta_{k_{min}} \leq \ldots \leq \theta_{k_{max}}} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} e^{-q_{out,k}(\theta_k B_{in,k} + B_{out,k})} \left(\theta_k\right)^{C_{in,k}} \left(q_{out,k}\right)^{C_{all,k}}}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} e^{-q_{all,k} B_{all,k}} \left(q_{all,k}\right)^{C_{all,k}}}$$

$$= \frac{\max_{1 \leq \theta_{k_{min}} \leq \ldots \leq \theta_{k_{max}}} \prod_{k=k_{min}}^{k_{max}} e^{-C_{all,k}} \left(\theta_k\right)^{C_{in,k}} \left(\frac{C_{all,k}}{\theta_k B_{in,k} + B_{out,k}}\right)^{C_{all,k}}}{\prod_{k=k_{min}}^{k_{max}} e^{-C_{all,k}} \left(\frac{C_{all,k}}{B_{all,k}}\right)^{C_{all,k}}}$$

$$= \max_{1 \leq \theta_{k_{min}} \leq \ldots \leq \theta_{k_{max}}} \prod_{k=k_{min}}^{k_{max}} \left(\theta_k\right)^{C_{in,k}} \left(\frac{B_{all,k}}{\theta_k B_{in,k} + B_{out,k}}\right)^{C_{all,k}}$$

Now, we must maximize this expression subject to the constraints on the θ_k, but it is not immediately clear how to accomplish this. One possibility would be to use a step method on the θ values $\left(\frac{C_{in,k} B_{out,k}}{C_{out,k} B_{in,k}}\right)$, but we have not yet been able to prove the optimality of this method.

2.4 Parametrized clusters

Here we assume that the rate increases over the duration of the cluster according to some known, parametrized distribution. We focus here on the case where the rate is exponentially increasing (multiplied by ϕ on every time step). Similar expressions may be derived for the case of a linear increase in rate (i.e. rate is increased by Δ on every time step).

2.4.1 Globally sensitive

In this case, we compare the null hypothesis H_0: the rate equals q_0 over all locations and times, to the alternative hypothesis $H_1(S)$: the rate is $\phi^{k-k_{min}+1} q_0$ at times $k = k_{min} \ldots k_{max}$ in region S, and equals q_0 over all other locations and times. The likelihood ratio is:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{\phi \geq 1} \prod_{s_i^k \in S \times (k_{min} \ldots k_{max})} \Pr(c_i^k \sim \mathrm{Po}(b_i^k \phi^{k-k_{min}+1} q_0))}{\prod_{s_i^k \in S \times (k_{min} \ldots k_{max})} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_0))}$$

$$= \frac{\max_{\phi\geq 1} \prod_{s_i^k \in S\times(k_{min}\ldots k_{max})} e^{-\phi^{k-k_{min}+1} q_0 b_i^k} (\phi^{k-k_{min}+1} q_0)^{c_i^k}}{\prod_{s_i^k \in S\times(k_{min}\ldots k_{max})} e^{-q_0 b_i^k} (q_0)^{c_i^k}}$$

$$= \frac{\max_{\phi\geq 1} \prod_{k=k_{min}}^{k_{max}} e^{-\phi^{k-k_{min}+1} q_0 B_{in,k}} (\phi^{k-k_{min}+1} q_0)^{C_{in,k}}}{\prod_{k=k_{min}}^{k_{max}} e^{-q_0 B_{in,k}} (q_0)^{C_{in,k}}}$$

$$= \max_{\phi\geq 1} \prod_{k=k_{min}}^{k_{max}} e^{(1-\phi^{k-k_{min}+1}) q_0 B_{in,k}} \phi^{(k-k_{min}+1) C_{in,k}}$$

Maximizing with respect to ϕ requires finding the root of a polynomial of degree $k_{max} - k_{min} + 1$; approximate (gradient) methods may also be used.

2.4.2 Globally adaptive

In this case, we compare the null hypothesis H_0: the rate equals q_{all} over all locations and times, to the alternative hypothesis $H_1(S)$: the rate is $\phi^{k-k_{min}+1} q_{out}$ at times $k_{min} \ldots k_{max}$ in region S, and equals q_{out} over all other locations and times. The likelihood ratio is:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{\phi\geq 1, q_{out}} \prod_{s_i^k \in S\times(k_{min}\ldots k_{max})} \Pr(c_i^k \sim \mathrm{Po}(b_i^k \phi^{k-k_{min}+1} q_{out})) \prod_{s_i^k \in out} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{out}))}{\max_{q_{all}} \prod_{s_i^k \in G\times(1\ldots k_{base})} \Pr(c_i^k \sim \mathrm{Po}(b_i^k q_{all}))}$$

$$= \frac{\max_{\phi\geq 1, q_{out}} \prod_{s_i^k \in S\times(k_{min}\ldots k_{max})} e^{-b_i^k \phi^{k-k_{min}+1} q_{out}} (\phi^{k-k_{min}+1} q_{out})^{c_i^k} \prod_{s_i^k \in out} e^{-b_i^k q_{out}} (q_{out})^{c_i^k}}{\max_{q_{all}} \prod_{s_i^k \in G\times(1\ldots k_{base})} e^{-b_i^k q_{all}} (q_{all})^{c_i^k}}$$

$$= \frac{\max_{\phi\geq 1, q_{out}} \prod_{k=k_{min}}^{k_{max}} e^{-B_{in,k} \phi^{k-k_{min}+1} q_{out}} (\phi^{k-k_{min}+1} q_{out})^{C_{in,k}} \times e^{-B_{out} q_{out}} (q_{out})^{C_{out}}}{\max_{q_{all}} e^{-B_{all} q_{all}} (q_{all})^{C_{all}}}$$

$$= \frac{\max_{\phi\geq 1, q_{out}} (q_{out})^{C_{all}} e^{-q_{out}(B_{out} + \sum B_{in,k} \phi^{k-k_{min}+1})} \phi^{\sum (k-k_{min}+1) C_{in,k}}}{\max_{q_{all}} e^{-B_{all} q_{all}} (q_{all})^{C_{all}}}$$

$$= \frac{\max_{\phi\geq 1} e^{-C_{all}} \left(\frac{C_{all}}{B_{out} + \sum B_{in,k} \phi^{k-k_{min}+1}} \right)^{C_{all}} \phi^{\sum (k-k_{min}+1) C_{in,k}}}{e^{-C_{all}} \left(\frac{C_{all}}{B_{all}} \right)^{C_{all}}}$$

$$= \max_{\phi\geq 1} \left(\frac{B_{all}}{B_{out} + \sum B_{in,k} \phi^{k-k_{min}+1}} \right)^{C_{all}} \phi^{\sum (k-k_{min}+1) C_{in,k}}$$

where the summations are taken from $k = k_{min} \ldots k_{max}$, "all" are sums over all space and time, and "out" are sums over all space and time except $S \times (k_{min} \ldots k_{max})$. Again, maximizing with respect to ϕ requires finding the root of a polynomial of degree $k_{max} - k_{min} + 1$.

2.4.3 Daily adaptive

In this case, we compare the null hypothesis H_0: for each day k, the rate equals $q_{all,k}$ for all locations, to the alternative hypothesis $H_1(S)$: for days $k_{min} \dots k_{max}$, the rate equals $q_{in,k}$ inside the region and $q_{out,k}$ outside, $q_{in,k}/q_{out,k} = \phi^{k-k_{min}+1}$, $\phi \geq 1$; and for all other days k, the rate equals $q_{all,k}$ for all locations. The likelihood ratio is:

$$D_{k_{min}}^{k_{max}}(S) = \frac{\max_{\phi \geq 1} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} \prod_{s_i^k \in S} \Pr(c_i^k \sim \text{Po}(b_i^k \phi^{k-k_{min}+1} q_{out,k})) \prod_{s_i^k \in G-S} \Pr(c_i^k \sim \text{Po}(b_i^k q_{out,k}))}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} \prod_{s_i^k \in G} \Pr(c_i^k \sim \text{Po}(b_i^k q_{all,k}))}$$

$$= \frac{\max_{\phi \geq 1} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} \prod_{s_i^k \in S} e^{-b_i^k \phi^{k-k_{min}+1} q_{out,k}} (\phi^{k-k_{min}+1} q_{out,k})^{c_i^k} \prod_{s_i^k \in G-S} e^{-b_i^k q_{out,k}} (q_{out,k})^{c_i^k}}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} \prod_{s_i^k \in G} e^{-b_i^k q_{all,k}} (q_{all,k})^{c_i^k}}$$

$$= \frac{\max_{\phi \geq 1} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} e^{-B_{in,k} \phi^{k-k_{min}+1} q_{out,k}} (\phi^{k-k_{min}+1} q_{out,k})^{C_{in,k}} e^{-B_{out,k} q_{out,k}} (q_{out,k})^{C_{out,k}}}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} e^{-B_{all,k} q_{all,k}} (q_{all,k})^{C_{all,k}}}$$

$$= \frac{\max_{\phi \geq 1} \prod_{k=k_{min}}^{k_{max}} \max_{q_{out,k}} e^{-q_{out,k}(B_{out,k} + B_{in,k} \phi^{k-k_{min}+1})} \phi^{(k-k_{min}+1)C_{in,k}} (q_{out,k})^{C_{all,k}}}{\prod_{k=k_{min}}^{k_{max}} \max_{q_{all,k}} e^{-B_{all,k} q_{all,k}} (q_{all,k})^{C_{all,k}}}$$

$$= \frac{\max_{\phi \geq 1} \prod_{k=k_{min}}^{k_{max}} e^{-C_{all,k}} \left(\frac{C_{all,k}}{B_{out,k} + B_{in,k} \phi^{k-k_{min}+1}}\right)^{C_{all,k}} \phi^{(k-k_{min}+1)C_{in,k}}}{\prod_{k=k_{min}}^{k_{max}} e^{-C_{all,k}} \left(\frac{C_{all,k}}{B_{all,k}}\right)^{C_{all,k}}}$$

$$= \max_{\phi \geq 1} \prod_{k=k_{min}}^{k_{max}} \left(\frac{B_{all,k}}{B_{out,k} + B_{in,k} \phi^{k-k_{min}+1}}\right)^{C_{all,k}} \phi^{(k-k_{min}+1)C_{in,k}}$$

Again, maximizing with respect to ϕ requires finding the root of a polynomial of degree $k_{max} - k_{min} + 1$.

3 Inferring baselines from previous counts

We now consider various methods of inferring baselines from the time series of previous counts. First, the building-aggregated time series (BATS) method considers the time series for each spatial location separately, computing the baseline for that location from that time series. The resulting counts and baselines are aggregated to a grid, and then one of the above scan statistics is used. For replica grids, each location's count is regenerated independently under the null hypothesis, and these are aggregated to a grid as before. Second, the cell-aggregated time series (CATS) method aggregates the time series for all locations in a grid cell into a single time series. Then the time series for each grid cell is considered separately to generate each grid cell's baseline, and then one of the above scan statistics is used. For replica grids, each cell's count is regenerated independently under the null hypothesis. Third, the region-aggregated time series (RATS) method creates a single time series for each region it searches (by aggregating the time series of the individual cells), computes the baseline for the region from that time series, and applies one of the scan statistics above.

The RATS method attempts to account for spatial correlations between cells, but the lack of a separate baseline per cell makes randomization difficult (since we would have to model correlations between cells and generate counts from this correlated distribution). Instead, we are considering various alternative methods of dealing with the multiple hypothesis testing problem, including use of the False Discovery Rate (FDR) criterion.

For all of these methods, a variety of univariate time series methods may be used to infer the baseline of a location, cell, or region from its time series of past counts. These include simple mean, maximum, and moving average methods; we are also considering methods which allow us to adjust for day of week effects, etc. Missing data is a serious problem for all of these methods. For BATS, we may use time series approaches which adjust for the presence of missing data; for CATS and RATS, we must infer these missing values before aggregating data at the cell or region level. For the over-the-counter drug sales data, our current best approach is an exponentially weighted moving average (EWMA) approach, applied to day-of-week-adjusted counts; the adjustment is made by estimating the proportion of weekly counts falling on each day.

4 Related work

In the spatio-temporal cluster detection literature, three main approaches have been proposed by Martin Kulldorff et al.: the retrospective space-time scan statistic (Kulldorff et al., 1998), the prospective space-time scan statistic (Kulldorff, 2001), and the space-time permutation scan statistic (Kulldorff et al., 2004). The first two of these approaches are very similar in that they use the globally adaptive test for detecting persistent clusters (Section 2.2.2), assuming that baselines are given based on census population estimates. The main difference is that the retrospective statistic searches over all space-time intervals, while the prospective statistic searches over those intervals ending at the present time, as described above. Kulldorff (2001) also gives a method of adjusting the prospective statistic for repeated time-periodic tests, i.e. if we want a false probability of α over an interval of longer than 1 day. This is straightforward, comparing D_{max} of the original grid to D_{max} of replica grids, where clusters in the replica grid can end before the present day. The globally adaptive statistic assumes that counts are proportional to population everywhere, with a constant of proportionality that is fixed through space and time. This assumption is clearly false: as a result, the statistics may pick up purely spatial clusters resulting from spatial variation in the underlying rate (e.g. different parts of the country have different disease rates), or purely temporal clusters based on temporal fluctuations in rate (e.g. seasonal affects or long-term trends). The daily adaptive test could deal with temporal clusters, while spatial clusters are best dealt with by inferring baselines from the time series of counts instead of using census populations.

Kulldorff et al. (2004) attempt to fix both problems by proposing the space-time permutation scan statistic. This statistic again uses the globally adaptive test for detecting persistent clusters, and does a prospective analysis. The main difference from the previous approaches is that baselines are inferred from the time series of counts: the inference is done by assuming that cases are independently distributed in space and time, and thus that $b_i^k = E[c_i^k] = \frac{\left(\sum_k c_i^k\right)\left(\sum_i c_i^k\right)}{\sum_i \sum_k c_i^k}$. Then the globally adaptive likelihood is used as before; note that B_{all} is set equal to C_{all} by construction, and thus the denominator $\left(\frac{C_{all}}{B_{all}}\right)^{C_{all}}$ can be ignored. Randomization is done by permuting the dates and locations of cases. In our view, this method has several disadvantages. First, by making the assumption of independence of space and time, the statistics for the current period are affected by space-time interaction occurring at any time in the past as well; we cannot adjust for known space-time trends and ask whether the current period has higher counts than expected even taking these trends into account. This is why we propose the spatial time series method: by separately examining the time series of past counts at each spatial location (or aggregated set of spatial locations), we attempt to predict the current count for that location or set of locations. The space-time permutation statistic uses the counts of the current period to infer the baselines for the current period, thus losing power (since baselines will be increased by increased counts, reducing the power to detect these increased counts). Additionally, since the expected count at each location is known, a globally sensitive test (with $q_0 = 1$, since we expect count to equal baseline) should be used instead of a globally adaptive test. Finally, the space-time permutation statistic assumes persistent clusters (i.e. constant disease rate over the outbreak period) while real outbreaks typically exhibit a disease rate that increases over the outbreak period; thus tests for emerging clusters should be able to detect an emerging outbreak more quickly than the space-time permutation scan statistic.

Several other spatio-temporal cluster detection methods have also been proposed. Iyengar (2004) searches over "truncated rectangular pyramid" shapes in space-time, thus allowing detection of clusters which move and grow/shrink linearly in space over time. The globally adaptive test for persistent clusters is again used, as in Kulldorff's statistics, and baselines are assumed to be given. Assuncao et al. (2004) assume a spatio-temporal Poisson point process: the exact location of each point in time and space is given, rather than aggregating points to discrete locations and intervals. A test statistic similar to the space-time permutation scan statistic is derived, assuming a Poisson intensity function that is separable in space and time.

References

• R. M. Assuncao, A. I. Tavares, M. Kulldorff (2004). An early warning system for space-time cluster detection. Working paper.

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

• D. Clayton, J. Kaldor (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 43, 671-681.

• V. Iyengar (2004). On detecting space-time clusters. *Proc. 10th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, 587-592.

• E. G. Knox (1964). The detection of space-time interactions. *Applied Statistics* 13, 25-29.

• M. Kulldorff (2001). Prospective time-periodic geographical disease surveillance using a scan statistic. *J. Royal Stat. Soc. A* 164, 61-72.

• M. Kulldorff, W. Athas, E. Feuer, B. Miller, C. Key (1998). Evaluating cluster alarms: a space-time scan statistic and cluster alarms in Los Alamos. *American Journal of Public Health* 88, 1377-1380.

• M. Kulldorff, R. Heffernan, J. Hartman, R. Assuncao, F. Mostashari (2004). A space-time permutation scan statistic for the early detection of disease outbreaks. Submitted for publication.

• N. Mantel (1967). The detection of cancer clustering and the generalized regression approach. *Cancer Research* 27, 209-220.