Rank-based spatial clustering: an algorithm for rapid outbreak detection

Jialan Que, Fu-Chiang Tsui

► An additional appendix is published online only. To view this file please visit the journal online (www.jamia.org).

RODS Laboratory, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Correspondence to

Jialan Que, Parkvale Building M-183, 200 Meyran Avenue, Pittsburgh, PA 15260, USA; jq4@pitt.edu

Received 12 August 2009
Accepted 10 February 2011

ABSTRACT

Objective Public health surveillance requires outbreak detection algorithms with computational efficiency sufficient to handle the increasing volume of disease surveillance data. In response to this need, the authors propose a spatial clustering algorithm, rank-based spatial clustering (RSC), that detects rapidly infectious but non-contagious disease outbreaks.

Design The authors compared the outbreak-detection performance of RSC with that of three well established algorithms—the wavelet anomaly detector (WAD), the spatial scan statistic (KSS), and the Bayesian spatial scan statistic (BSS)—using real disease surveillance data on to which they superimposed simulated disease outbreaks.

Measurements The following outbreak-detection performance metrics were measured: receiver operating characteristic curve, activity monitoring operating curve curve, cluster positive predictive value, cluster sensitivity, and algorithm run time.

Results RSC was computationally efficient. It outperformed the other two spatial algorithms in terms of detection timeliness, and outbreak localization. RSC also had overall better timeliness than the time-series algorithm WAD at low false alarm rates.

Conclusion RSC is an ideal algorithm for analyzing large datasets when the application of other spatial algorithms is not practical. It also allows timely investigation for public health practitioners by providing early detection and well-localized outbreak clusters.

INTRODUCTION

As recent experience with SARS, H1N1, and cholera demonstrate, large outbreaks of infectious disease continue to threaten the world's population. It thus has become a priority of public health practitioners to identify and localize an outbreak in its early stages to allow public health response as early as possible.

Over the past decade, a large number of approaches to early detection and localization based on more rapid collection of disease surveillance data and their automatic algorithmic analysis have been developed.^{1–4} The algorithms, in addition to time-series algorithms that analyze univariate data in the temporal dimension, include many spatial and tempo-spatial algorithms, which allow better detection and localization of outbreaks caused by infectious but non-contagious disease agents (eg, an aerosol release of *Bacillus anthracis*, waterborne outbreaks caused by pathogens such as *Cryptosporidium parvum* and *Giardia lamblia*), which typically affect an aggregated group of geographic areas. These spatial algorithms

have also been applied (either retrospectively or prospectively) to non-infectious public health problems such as infant death⁵ and prostate cancer survival data.⁶

Current spatial algorithms divide into frequentist and the Bayesian approaches. Given a hypothesis, a Bayesian approach assigns a probability to it, whereas a frequentist approach typically tests it. The spatial scan statistic by Kulldorff *et al* (KSS)⁵ is a frequentist approach that exhaustively searches for areas of maximum disease activity (eg, incidence) within a region of interest using circular or elliptic sub-regions of different sizes centered on various locations. Other frequentist methods include the flexible spatial scan statistic (FSS), which relaxes the constraint on cluster shape used in KSS,⁷ and the upper level set scan statistic (ULS), which searches tessellated clusters from some subsets of the study areas (each subset includes the areas with higher elevated rates than a predefined threshold).⁸ All of the frequentist algorithms compute a score of a likelihood ratio of having an outbreak in each considered cluster versus no outbreak and then perform a randomization test to decide its significance.

On the other hand, the Bayesian approaches do not require the randomization test. Neill's Bayesian spatial scan statistic (BSS) employs an $m \times m$ grid covering the area of interest to search for clusters. Each cell in the grid is a geographic unit. BSS identifies a rectangular sub-region, which is composed of the cells with the highest posterior probability of having an outbreak. Other Bayesian methods include the Bayesian-based multilevel spatial clustering algorithm and the z-score-based multilevel spatial clustering algorithm.^{9–10} These algorithms identify clusters from a sub-dataset to achieve computational efficiency; here the challenge of deciding on the proper criteria for data selection is introduced.

The current frequentist and the Bayesian scan statistics face some common limitations. First, they are computationally intensive because of exhaustive searching, randomization testing or both. Therefore, in time-sensitive applications, an algorithm may take too long to complete, rendering its results outdated or delayed for decision makers. Second, certain artificial cluster shapes used by those algorithms may not conform to true outbreak shapes, thus reducing their sensitivity to small outbreaks and the timeliness of detection of other outbreaks.

In an effort to overcome these limitations, we developed a non-parametric methodology for early outbreak detection and localization, a 'rank-based clustering' (RSC) algorithm. RSC uses heuristic

search, based on statistical models that assess the risk rates of having an outbreak occur in each geographic unit (eg, a ZIP code area) to improve the time efficiency of cluster searching. In the following sections, we describe the RSC algorithm in detail and then evaluate the performance of RSC by comparing it with three well-established detection algorithms.

METHODS

The RSC algorithm

The key steps in the RSC algorithm are: (1) computing the risk rate of each geographic unit having an ongoing outbreak, and ranking each unit by its estimated risk rate; (2) searching for clusters based on the rankings and on geographic adjacency; (3) computing the posterior probabilities for each cluster.

Computing the risk rates

We studied two measures for estimating the risk rate for each geographic unit: standard score and posterior probability.

Standard score (z-score)

The model computes a standard score (also known as z-score), $R(z_i, T) = \frac{c_{i,T} - b_{i,T}}{s_i}$, to measure the risk of each geographic unit z_i on day T , where s_i represents the estimated SD of the residuals. The residuals are computed by subtracting expected values ($b_{i,t}$) from observed ones ($c_{i,t}$) in the time series of z_i at time t , correspondingly.

Bayesian posterior probability

The standard score results in a large number when most of the counts in a time series are close to the mean value (ie, when $s_i \rightarrow 0$ and $c_{i,T} \neq b_{i,T}$, $R(z_i, T) \rightarrow \infty$). In such circumstances, one can use a Bayesian approach to estimate $R(z_i)$ as the posterior probability, $P(H_1(z_i, T) | D_i)$, where T is the most current day.⁹

We assume that the counts for z_i within each period t , $1 \leq t \leq T$, follow a Poisson distribution, which is commonly used to model a certain variable that counts a number of discrete occurrences during a time interval of a given length.¹¹ A gamma distribution, the conjugate prior¹ of a Poisson distribution, is then used to model the ratio q (ie, the ratio of the observed counts to the expected). Expert knowledge can also be introduced by setting different prior probabilities ($P(H_1(z_i, T))$) to different geographic units z_i at different times t . (We applied uniform priors in this study.) The posterior probability of $H_1(z_i, T)$ is computed using Bayes theorem (eqn (1)), where the likelihood of the null hypothesis H_0 (not having an outbreak) and alternative hypothesis H_1 (having an outbreak in some area) are integrals over the ratio q (eqns (2) and (3)) with different shape parameters α and α' , respectively. The marginal probability is computed as the sum over the two hypotheses as denoted in eqn (4).

$$P(H_1(z_i, T) | D_{iT}) = \frac{P(D_{iT} | H_1(z_i, T))P(H_1(z_i, T))}{P(D_{iT})} \quad (1)$$

$$P(D_{iT} | H_0(z_i, T)) = \int P(q \sim \text{Gamma}(\alpha_i, \beta_i)) P(c_{i,T} \sim \text{Poisson}(qb_{i,T})) dq = \frac{b_{i,T}^{\alpha_i} \Gamma(\alpha_i + c_{i,T})}{c_{i,T}! (\beta_i + b_{i,T})^{\alpha_i + c_{i,T}} \Gamma(\alpha_i)} \quad (2)$$

$$P(D_{iT} | H_1(z_i, T)) = \int P(q \sim \text{Gamma}(\alpha_i', \beta_i)) P(c_{i,T} \sim \text{Poisson}(qb_{i,T})) dq = \frac{b_{i,T}^{\alpha_i'} \Gamma(\alpha_i' + c_{i,T})}{c_{i,T}! (\beta_i + b_{i,T})^{\alpha_i' + c_{i,T}} \Gamma(\alpha_i')} \quad (3)$$

$$P(D_{iT}) = P(D_{iT} | H_0(z_i, T))P(H_0(z_i, T)) + P(D_{iT} | H_1(z_i, T))P(H_1(z_i, T)) \quad (4)$$

In the above equations, $c_{i,T}$ and $b_{i,T}$ are the observed and the expected values for geographic unit z_i on day T , respectively, and $\Gamma(\cdot)$ represents the γ function. The shape parameter (α_i) and the rate parameter (β_i) of the gamma distribution are learned from the historical data using moment matching methods (eqns (5–8)). As with the alternative hypothesis, we assume that the outbreak will increase α_i by a multiplicative factor $k > 1$, while leaving β_i unchanged. In this study, k is assumed to follow a discrete uniform distribution in the range of [1.2, 3] with 10 values in between with steps 0.2.

$$\alpha_i = \frac{(E[q_i])^2}{\text{Var}[q_i]} \quad (5)$$

$$\beta_i = \frac{E[q_i]}{\text{Var}[q_i]} \quad (6)$$

$$\hat{\alpha}_i = \frac{(E_{\text{sample}}[\frac{c_{i,t}}{b_{i,t}}])^2}{(E_{\text{sample}}[\frac{c_{i,t}}{b_{i,t}}])^2} \quad (7)$$

$$\hat{\beta}_i = \frac{E_{\text{sample}}[\frac{c_{i,t}}{b_{i,t}}]}{\text{Var}[E_{\text{sample}}[\frac{c_{i,t}}{b_{i,t}}]]} \quad (8)$$

Searching for clusters

RSC computes the shortest Euclidean distance d_{ij} between every two geographic units, z_i and z_j . We define an adjacency threshold η if $d_{ij} \leq \eta$, z_i and z_j are considered to be adjacent. The geographical units are considered to fall into the same cluster when each of them is adjacent to at least one of the other members of the cluster.

The search for the highest probability cluster is a greedy search. The search is seeded with the highest ranked geographic unit and this area itself becomes the first cluster. The second ranked geographic unit is the next to consider. If this area is adjacent to the first, they merge to form a new cluster; otherwise, the second area becomes a cluster. The algorithm processes the remaining geographic units in a risk-rate-descending order by adding them to one cluster if a geographic unit is adjacent to one or more previously constructed clusters or creating a separate single-area cluster. Figure 1 illustrates the clusters produced by the algorithm from a region where some areas are adjacent (connected in this example) and some are not. To constrain the growth of large clusters, one can set an upper bound on cluster size. For example, in this study, a cluster stops growing if it includes more than half of the geographic units. Each created cluster is scored using posterior probability, which we describe in the following section. Figure 2 is the pseudo-code for the algorithm.

⁹In Bayesian probability theory, a conjugate prior is the posterior probability distribution which is in the same family as the prior probability distribution, which provides a closed-form expression for the posterior probability.

Figure 1 Clusters created within a region of eight areas (A). Each figure in (B)–(I) shows a newly created cluster. The number in each cell is the ranking of its risk rate, which decides its order to be considered by the algorithm.

Computing cluster posterior probabilities

We compute the posterior probability of a cluster S using a gamma-Poisson model as used in the BSS algorithm.¹² eqns (9 and 10) compute the likelihood that there is no outbreak in the entire region (H_0) and the likelihood that there is an outbreak in cluster S (H_1), respectively.

$$P(D|H_0) = \frac{(\beta_{all})^{\alpha_{all}} \Gamma(\alpha_{all} + C_{all})}{(\beta_{all} + B_{all})^{\alpha_{all} + C_{all}} \Gamma(\alpha_{all})} \times \prod_{all} \frac{b_i^{c_i}}{c_i!} \quad (9)$$

$$P(D|H_1(S)) = \frac{(\beta_{in})^{\alpha_{in}} \Gamma(\alpha_{in} + C_{in})}{(\beta_{in} + B_{in})^{\alpha_{in} + C_{in}} \Gamma(\alpha_{in})} \times \frac{(\beta_{out})^{\alpha_{out}} \Gamma(\alpha_{out} + C_{out})}{(\beta_{out} + B_{out})^{\alpha_{out} + C_{out}} \Gamma(\alpha_{out})} \times \prod_{all} \frac{b_i^{c_i}}{c_i!} \quad (10)$$

In the null hypothesis, H_0 , the number of observed cases summed over all the areas C_{all} follow a Poisson distribution—that is, $C_{all} \sim \text{Poisson}(q_{all}B_{all})$, where B_{all} is the number of expected cases summed over all these areas and the disease rate q_{all} follows a gamma distribution, denoted as $q_{all} \sim \text{gamma}(\alpha_{all}, \beta_{all})$. Similarly, in the alternative hypothesis $H_1(S)$, the number of observed cases in the areas within cluster S , C_{in} , follows a Poisson distribution $C_{in} \sim \text{Poisson}(q_{in}B_{in})$, where B_{in} is the number of expected cases summed over all the areas inside the cluster S and the disease rate has a gamma distribution $q_{in} \sim \text{gamma}(\alpha_{in}, \beta_{in})$. For the study areas outside cluster S , C_{out} , B_{out} , α_{out} and β_{out} represent the corresponding parameters. The prior variables in gamma distributions can be estimated using

1. $S = \{\}$
2. compute $R(z_i)$, $i = 1, \dots, n$
3. sort $(z_1, z_2, \dots, z_n)$ into $(z_{r_1}, z_{r_2}, \dots, z_{r_n})$ where $R(z_{r_j}) \geq R(z_{r_{j+1}})$, $j = 1, \dots, n-1$
4. for $k = 1$ to n
5. add z_{r_k} into a new cluster S_{new}
6. for each cluster S in S
7. if z_{r_k} is adjacent to any area in S
8. add all the areas in S into S_{new}
9. if $|S_{new}| < \frac{n}{2}$
10. compute score $F(S_{new})$ for cluster S_{new}
11. add S_{new} into S

Figure 2 The pseudo-code for cluster searching.

the moment matching approach (eqns (5–8)). The posterior probability is computed using Bayes Theorem (eqn (11)), and the marginal probability can be computed (eqn (12)). The prior of each cluster S is assumed to follow a uniform distribution, $P(H_1(S)) = \frac{P_1}{M}$, where P_1 represents the probability of there being an outbreak somewhere in the entire study region, and M is the total number of the clusters created by the algorithm.

$$P(H_1(S)|D) = \frac{P(D|H_1(S))P(H_1(S))}{P(D)} \quad (11)$$

$$P(D) = P(D|H_0)P(H_0) + \sum_S P(D|H_1(S))P(H_1(S)) \quad (12)$$

Algorithm evaluation

We compared the performance of RSC with that of KSS, the most widely known frequentist approach, and BSS, a well-established Bayesian approach.¹² We also compared RSC with the wavelet anomaly detector (WAD), in order to benchmark the detection-performance characteristics of RSC against a member of the time-series algorithms by considering WAD as a representative.^{13–14} We applied each algorithm to semi-synthetic datasets, which were generated by superimposing outbreak cases into real over-the-counter healthcare product sales data assumed to have no outbreaks.

Over-the-counter healthcare product sales data

We created a semi-synthetic dataset for evaluation by injecting simulated outbreak data into real daily sales counts for cough and cold products in Pennsylvania between January 1, 2008 and December 31, 2008.^{13–15}

We excluded stores with unreliable reporting from the dataset. We defined unreliable reporting as failure to send any records in any of the 23 over-the-counter categories for more than 27 days during the period of the dataset (27 days was 5% of the time period, plus 3 federal holidays per year). This filtering of stores removed 498 of 1502 (33%) stores.

We aggregated the daily sales by ZIP code (471 ZIP code areas).

We used the data from January 1, 2007 through December 31, 2007 as a training dataset from which to compute the threshold values for given false alarm rates by assuming there were no known outbreak signals in this period. Each day within the training period was analyzed using an algorithm, and the greatest score (ie, the posterior probability) was recorded. A set of scores were then used as threshold values to control for

different false alarm rates. It is worth noting that the threshold values computed in this way are likely to be overestimated because of possibly existing but unveiled outbreak signals.

Semi-synthetic outbreaks

We used a simple outbreak model to inject outbreak cases into the dataset of cough and cold products. The size and shape of the injected signal is adjustable by parameters that set outbreak size (K), slope (δ), and duration (D) (eqn (13))

$$O(T, \delta, z_i) = \begin{cases} \delta \cdot \omega \cdot \bar{c}_i & 1 \leq \omega \leq \lfloor \frac{D}{2} \rfloor \\ \delta \cdot (D - \omega) \cdot \bar{c}_i & \lfloor \frac{D}{2} \rfloor + 1 \leq \omega \leq D \end{cases} \quad (13)$$

where ω is the number of days after outbreak onset and $\bar{c}_i$ is the mean value of the daily counts in geographic unit z_i . Figure 3 shows a simulated outbreak curve, which has a simple triangular shape with an upward phase simulating disease diffusion. Although this simplified simulation is clearly not a very realistic outbreak, it does have an advantage: it allows us to precisely control the slope of the outbreak curve and examine how this affects the detection ability of our method; in addition, the slowly elevated curve (slower than the log normal curves often observed in real outbreaks) extends an outbreak onset period to some extent, which allows us to better distinguish the timeliness performances of detection algorithms.^{12 19}

The geographical shapes of the simulated outbreaks created in this study were designed to be flexible and independent of any detection algorithm except that the simulator assumes that an outbreak cluster has contiguous areas. To generate an outbreak, we randomly selected a geographic unit as an outbreak area. Second, we randomly chose the rest of the outbreak areas, with each new one required to be adjacent to at least one of the previously selected. In this way, the outbreak simulator was mostly simplified and generalized. It conformed to the exhibiting characteristics of most outbreaks generated by many well-known simulation models, which also distribute infected cases into clustered (oftentimes connected) regions.^{2 20–22}

We injected the simulated outbreak counts into the real cough and cold sales data for the period January 1, 2008 to December 31, 2008. We injected each simulated outbreak into a randomly selected interval of length D days during this period. We used a time-series algorithm to compute each day's expected value during the evaluation period based on the last 12 months of data previous to that day.

We generated six groups of datasets with K and δ chosen from $\{4, 8, 12\}$ and $\{0.2, 0.3\}$, respectively. We used a fixed duration D of 10 days. Each group included 100 different (differing by start data or seed area) outbreaks. More elevated outbreaks (ie, $\delta > 0.3$) are not discussed here because of the indistinguishable detection performances among the different algorithms applied.

Figure 3 Illustration of the temporal shape of an artificial outbreak curve from day 1 to day D ($D=10$).

Evaluation metrics

The evaluation metrics included (1) the receiver operating characteristic (ROC) curve,²³ (2) the activity monitoring operating characteristic (AMOC) curve,²⁴ (3) the areas under the ROC and AMOC curves, (4) computation time, (5) cluster sensitivity (the proportion of the number of outbreak areas correctly detected), and (6) cluster positive predictive value (PPV, the proportion of the number of true outbreak areas in the detected cluster).^{6 7 25}

Given a threshold value that is computed from the training dataset controlling for a particular false alarm rate, we define a true positive output cluster for all the spatial algorithms as a cluster that satisfies the three conditions: (1) having a score that is greater than the threshold, (2) having one or more outbreak areas identified, and (3) identified within the upward phase of the outbreak (ie, within the first 5 days). For the temporal algorithm, WAD, a true positive output is a geographic unit that (1) has a score that is greater than the threshold and (2) is one of the outbreak areas.

Algorithm settings

We compared the performance of RSC with WAD,²⁶ KSS,^{5 27 28} and BSS.¹² Note that WAD was not only compared as an individual detection algorithm, but also used to compute expected values required by the three spatial algorithms.

We applied the discrete Poisson model proposed in KSS for a purely spatial analysis performed by SaTScan V.8.0. As suggested by the author of SaTScan, the size of a searched for cluster was limited to $<3\%$ of the population in favor of relatively small and focused outbreaks. The input parameter file for SaTScan used in this study is shown in appendix 1. For BSS, we used a 24-by-24 grid structure to cover the entire study region (ie, state of Pennsylvania), with the area of each grid cell being ~ 80 square miles. It would make a more compelling case to compare RSC with BSS with a finer grid size. However, to have a finer grid so that each cell can cover a single ZIP code area (assuming each ZIP code area covers about 10 square miles on average), we would need a 64-by-64 grid structure. One of our pilot studies showed that a BSS with a grid size of 32 required 70+ min. Theoretically, considering the time complexity of BSS, which is $O(m^4)$, a 64-by-64 grid size algorithm might take about $2^4 \times 70$ min which is about 19 h.

The experiments were executed on a Linux server with a 2 GHz Intel CPU and 4 GB memory. All the algorithms except KSS were implemented in Java 1.5.

RESULTS

Detection power and timeliness

The ROC curves of the RSC algorithms using the standard score model and the Bayesian model are shown in figure 4, as well as the curves for WAD and BSS. Figure 5 provides the corresponding AMOC curves. Because the lowest false alarm rate that KSS could achieve is 0.76,ⁱⁱ it is not shown in the figures, with false alarm rates ranging between 0 and 0.2. We assume that any false alarm rates greater than 0.2 (1 false alarm per 5 days) have no practical advantage to public health practitioners.

ⁱⁱIn order to compute false alarm rates, we ran the algorithms on each day between January 1, 2007 and December 31, 2007 without injecting any outbreak cases. For each day, only the cluster with the lowest p value was considered. The results of KSS showed that 278 out of 365 clusters had the lowest p value, 0.001, which means the lowest false alarm rate that the algorithm was able to achieve was $278/365=0.76$.

Figure 4 Receiver operating characteristic curves of the four algorithms with different outbreak settings. A ($K=4$, $\delta=0.2$); B ($K=4$, $\delta=0.3$); C ($K=8$, $\delta=0.2$); D ($K=8$, $\delta=0.3$); E ($K=12$, $\delta=0.2$) and F ($K=12$, $\delta=0.3$). BSS, Bayesian spatial scan statistic; RSC, rank-based spatial clustering; WAD, wavelet anomaly detector.

Table 1 compares partial areas under the ROC curve for false positive rates within the range $[0, 0.2]$ using trapezoidal approximation. DeLong testing²⁹ showed no significant difference between any pair of the algorithms. RSC, BSS and WAD all demonstrated similar detection sensitivities.

Table 1 also compares the areas under partial AMOC curves, demonstrating the detection timeliness of the algorithms. The two RSC methods exhibited better timeliness than the other algorithms in five out of the six groups of experiments. The average number of days it took the algorithms to detect the outbreaks are shown in table 2. Paired student t tests performed on the variable 'days-to-detect' showed that both RSC methods were able to detect outbreaks significantly earlier than BSS in all six groups at a false alarm rate of 0.1 (ie, 1 false alarm per 10 days). RSC also outperformed WAD when those datasets injected with low intensity outbreaks were analyzed (ie, $\delta=0.2$).

Outbreak localization

The results for average cluster sensitivity for spatial algorithms at a false alarm rate of 0.1 are provided in table 3. In all six groups of experiments, either the $RSC_{z-score}$ or $RSC_{Bayesian}$ showed significantly higher cluster sensitivity than the other algorithms. This result indicates that the RSC is capable of identifying more outbreak areas than other algorithms. Correspondingly, table 3 shows the average measures of cluster PPVs. Both RSC algorithms had significantly better cluster PPV, which suggests that the RSC algorithms made fewer type II (false negative) errors than the other algorithms.

Run time

The average run times for $RSC_{z-score}$, $RSC_{Bayesian}$, WAD, BSS, and KSS were 26 s, 24 s, 0.22 s, 44 min, and 2.58 s, respectively. WAD had a shorter run time than any of the spatial algorithms. Among the spatial algorithms, both the $RSC_{z-score}$ and

Figure 5 Activity monitoring operating characteristic curves of the four algorithms with different outbreak settings. A ($K=4$, $\delta=0.2$); B ($K=4$, $\delta=0.3$); C ($K=8$, $\delta=0.2$); D ($K=8$, $\delta=0.3$); E ($K=12$, $\delta=0.2$) and F ($K=12$, $\delta=0.3$). BSS, Bayesian spatial scan statistic; RSC, rank-based spatial clustering; WAD, wavelet anomaly detector.

Table 1 Algorithm outbreak detection performance

Algorithm	RSC _{z-score}	RSC _{Bayesian}	BSS	WAD
Areas under ROC				
(K=4, $\delta=0.2$)	0.1005	0.0957	0.0599	0.0875
(K=4, $\delta=0.3$)	0.1348	0.1280	0.0794	0.1525
(K=8, $\delta=0.2$)	0.1069	0.0889	0.0792	0.1028
(K=8, $\delta=0.3$)	0.1497	0.1364	0.1093	0.1706
(K=12, $\delta=0.2$)	0.1151	0.1082	0.0935	0.1256
(K=13, $\delta=0.3$)	0.1560	0.1402	0.1224	0.1740
Areas under AMOC				
(K=4, $\delta=0.2$)	0.8617	0.8552	0.9254	0.9199
(K=4, $\delta=0.3$)	0.7814	0.7896	0.8853	0.8021
(K=8, $\delta=0.2$)	0.8318	0.8567	0.9000	0.8988
(K=8, $\delta=0.3$)	0.7269	0.7520	0.8065	0.7543
(K=12, $\delta=0.2$)	0.8387	0.8111	0.8544	0.8751
(K=12, $\delta=0.3$)	0.6965	0.7318	0.7952	0.6929

Areas under the receiver operating characteristic and activity monitoring operating characteristic curves with false positive rates in the range (0, 0.20). K represents outbreak size and δ represents outbreak intensity. The boldfaced numbers show the best performing algorithms.

BSS, Bayesian spatial scan statistic; RSC, rank-based spatial clustering; WAD, wavelet anomaly detector.

RSC_{Bayesian} were more than 100 times faster than BSS. Please note that the measured average run time for KSS is not strictly comparable, since the software SaTScan was implemented in a different programming language, C.

DISCUSSION

We have presented an RSC algorithm and demonstrated several ways in which this approach is preferable to other temporal or spatial scan algorithms. The experiments demonstrated that RSC using both risk estimation models consistently outperforms the other algorithms in terms of detection timeliness while having comparable detection powers.

The results for the run times of RSC and three compared algorithms can be explained by their time complexity theoretically. The time complexity of RSC is $O(n^2)$ plus the time complexity of computing the adjacency relations (ie, the shortest distances) between any two geographic units. The time complexity of computing the shortest distance between two vector polygons defining the borderlines of two geographic units is $O(uv)$, where u and v are the numbers of vertices of the two polygons.³⁰ In fact, most spatial database systems (eg, postGIS) offer querying capabilities on topological relations. The adjacency relationship can also be computed in advance by storing them in a look-up (hash) table for a quick search. Both ways make the run time of adjacency computation less significant. BSS has $O(m^4)$ time complexity, where m is the length of the grid.¹² KSS has $O(Rn^2)$ complexity for the purely spatial

Table 2 Average days to detect one false alarm per 10 days, for each of the six groups of simulations

K	δ	RSC _{z-score}	RSC _{Bayesian}	BSS	WAD
K=4	$\delta=0.2$	<u>4.3</u>	4.2	4.63	4.62
	$\delta=0.3$	<u>3.83</u>	3.79	4.4	<u>3.83</u>
K=8	$\delta=0.2$	4.08	<u>4.29</u>	4.53	4.47
	$\delta=0.3$	<u>3.51</u>	3.61	3.95	3.5
K=12	$\delta=0.2$	<u>4.22</u>	3.98	4.32	4.37
	$\delta=0.3$	<u>3.27</u>	3.52	3.92	3.17

K represents outbreak size and δ represents outbreak intensity. The boldfaced results reveal the best performance. Those with underscores are not significantly different (at $\alpha=0.05$) from the best.

BSS, Bayesian spatial scan statistic; RSC, rank-based spatial clustering; WAD, wavelet anomaly detector.

Table 3 Comparison of average cluster positive predictive values (PPVs) and cluster sensitivities at a false alarm rate of 0.1

Algorithm	RSC _{z-score}	RSC _{Bayesian}	BSS
Cluster PPV			
(K=4, $\delta=0.2$)	0.89	<u>0.83</u>	0.54
(K=4, $\delta=0.3$)	0.89	<u>0.85</u>	0.55
(K=8, $\delta=0.2$)	<u>0.72</u>	0.75	<u>0.60</u>
(K=8, $\delta=0.3$)	0.84	0.84	0.56
(K=12, $\delta=0.2$)	<u>0.76</u>	<u>0.69</u>	0.49
(K=12, $\delta=0.3$)	0.71	<u>0.70</u>	0.53
Cluster sensitivity			
(K=4, $\delta=0.2$)	<u>0.93</u>	0.94	0.81
(K=4, $\delta=0.3$)	0.95	0.95	0.79
(K=8, $\delta=0.2$)	<u>0.90</u>	0.91	0.76
(K=8, $\delta=0.3$)	0.91	<u>0.89</u>	0.77
(K=12, $\delta=0.2$)	<u>0.86</u>	0.91	0.77
(K=12, $\delta=0.3$)	0.86	0.86	0.71

K represents outbreak size and δ represents outbreak intensity. The boldfaced results reveal the best performance; those with underscores are not significantly different (at $\alpha=0.05$) from the best.

BSS, Bayesian spatial scan statistic; RSC, rank-based spatial clustering.

model, where R represents the number of replicated analyses required by randomization tests. $O(n)$ is the time complexity of the pyramid algorithm implemented in WAD.³¹

The RSC algorithm can be extended to detect outbreaks in the areas not restricted to connected ones. One approach is to adjust the threshold distance inversely proportional to some density measure of a baseline factor (eg, population). Another is to attach non-residential landforms (eg, lakes, valleys, etc) to their nearest census tract and consider them in the analysis as well. In addition, one can use a grid cell as a unit study area that comprises one or more geographic units (eg, ZIP code areas) and search for clusters with adjacent grid cells, where the non-connected geographical areas (eg, ZIP code areas) inside two or more connected grid cells can be clustered together (J Que, unpublished data).

Furthermore, RSC has better detection accuracy and precision than other tested algorithms. A cluster that is inaccurately shaped or much larger than the actual outbreak area may result in barriers for epidemiologists to conduct investigations given the very limited resources they have.

Detecting flexible shaped clusters is another advantage of RSC. The output of RSC is a combination of a set of adjacent areas (eg, ZIP code areas). A cluster identified by RSC may be any irregular and flexible shape, and the identified shape can be informative and indicative for detection of specific types of outbreaks. For example, an elongated cluster may suggest an aerosol release of a disease agent that follows a downwind direction, or a fish-bone-like cluster may imply a waterborne outbreak.

Since RSC is computationally efficient, it may be applied to finer spatial units other than postal code areas while still having an acceptable run time. For example, individual cases can be aggregated into street or block groups for a drill-down analysis, given that a more finely grained dataset is available (eg, street-level address is available for analysis). However, when no predefined demographic unit is appropriate, such as individual cases that are indicative of an environmental hazard (eg, a radiation leak) or a nosocomial infection (ie, hospital-acquired infection), artificial units have to be considered, such as circles centering on a nuclear plant or a grid structure displaying ward locations in a hospital.

Similar to RSC, the ULS algorithm⁸ searches for clusters in a reduced search space. However, the ULS and RSC algorithms

differ in three respects: the risk estimation models used, the cluster search space, and cluster significance testing. More specifically, ULS stratifies the dataset into several subsets using a set of predefined threshold rates looking for tessellated areas from each subset, whereas RSC creates a new cluster each time when taking the next-ranked area into consideration, which makes its search space a super set of ULS. When ULS sets a set of thresholds and each value in the set is equal to the risk rate of each geographic unit in the study, ULS finds sets of clusters identical with those RSC would find if both applied the same risk estimation model. In terms of significance testing, ULS uses frequentist randomization testing, whereas RSC applies Bayesian inference. Our future work may include performance comparison between these two algorithms.

A limitation of this study is that it used a geometrically shaped outbreak curve and therefore may not represent the complexity of real outbreaks. Future work should include evaluations of these algorithms using the most realistic outbreak simulators possible and validation with real outbreak data.

CONCLUSIONS

One of the essential functions of biosurveillance systems is to detect outbreaks at an early stage at local, regional or even national levels. Our results suggest that the RSC algorithm is computationally efficient for outbreak detection and localization. When given limited computation resources or there is a time-sensitive requirement, its lower order of computational complexity makes it preferable for analyzing large datasets, either covering broad geographic regions or including multiple data types. The significantly better detection timeliness suggests that RSC may be able to identify some outbreaks at an earlier stage. Moreover, the compact clusters localized by RSC allow field epidemiologists to focus on investigating a manageable number of high priority regions given limited resources.

Funding This work is supported by grants CDC-1U38HK000063-01, CDC-1 R01 PH00026-01, PADOH-ME-01737, and NSF-IIS-0325581.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

- Lombardo J, Burkom H, Elbert E, et al. A systems overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II). *J Urban Health* 2003;**80**(Suppl 1):32–42.
- Buckeridge DL, Burkom H, Moore AM, et al. Evaluation of syndromic surveillance systems—design of an epidemic simulation model. *Morb Mortal Weekly Report* 2004;**53**(Suppl):137–43.
- Tsui F-C, Espino J, Dato V, et al. Technical description of RODS: a real-time public health surveillance system. *J Am Med Inform Assoc* 2003;**10**:399–408.
- Wagner MM, Tsui FC, Espino JU, et al. The emerging science of very early detection of disease outbreaks. *J Public Health Manage Pract* 2001;**7**:51–9.
- Kulldorff M. A spatial scan statistic. *Commun Stat Theory Methods* 1997;**26**:1481–96.
- Huang L, Kulldorff M, Gregorio D. A spatial scan statistic for survival data. *Biometrics* 2007;**63**:109–18.
- Takahashi K, Kulldorff M, Tango T, et al. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *Int J Health Geogr* 2008;**7**:14.
- Patil GP, Taillie C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environ Ecol Stat* 2004;**11**:183–97.
- Que J, Tsui F-C. A multi-level spatial clustering algorithm for detection of disease Outbreaks. *AMIA Annu Symp Proc* 2008;**6**:611–15.
- Que J, Tsui FC, Espino J. A z-score based multi-level spatial clustering algorithm for the detection of disease outbreaks. *Lecture Notes in Computer Science. BioSecure* 2008;**5453**:108–18.
- Johnson NL, Kotz S, Kemp AW. *Univariate Discrete Distributions*. Hoboken, NJ: Wiley, 1993.
- Neill DB, Moore A, Cooper GF. A Bayesian spatial scan statistic. *Adv Neural Inf Process Syst* 2005;**18**:1003–10.
- Siegrist D, Pavlin J. *BioALERT Biosurveillance Testbed Evaluation. Syndromic Surveillance: Reports from A National Conference*. New York, NY; 2004:152–8.
- Buckeridge D, Burkom H, Campbell M, et al. Algorithms for rapid outbreak detection: a research synthesis. *J Biomed Inform* 2005;**38**:99–113.
- Watkins RE, Eagleson S, Beckett S, et al. Using GIS to create synthetic disease outbreaks. *BMC Med Inform Decis Mak* 2007;**7**:4.
- Magruder SF. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. *Johns Hopkins APL Tech Digest* 2003;**24**:349–53.
- Mandl KD, Reis B, Cassa C. Measuring outbreak-detection performance by using controlled feature set simulations. *MMWR Morb Mortal Wkly Rep* 2004;(Suppl 53):130–6.
- Zhang M, Kong X, Wallstrom GL. Simulation of multivariate spatial-temporal outbreak data for detection algorithm evaluation. *Lecture Notes in Computer Science, BioSecure*. 2008;**5354**:155–63.
- Neill DB, Moore AM, Sabhnani M, et al. *Detection of Emerging Space-time Clusters*. Chicago, IL: Proc 11th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining, 2005.
- Buckeridge DL, Owens DK, Switzer P, et al. Evaluating detection of an inhalational anthrax outbreak. *Emerg Infect Dis* 2006;**12**:1942–9.
- Hogan WR, Cooper GF, Wagner MM, et al. *An Inverted Gaussian Plume Model for Estimating the Location and Amount of Release of Airborne Agents from Downwind Atmospheric Concentrations*. Pittsburgh, PA: RODS Laboratory Technical Report, 2005.
- Kulldorff M, Zhang Z, Hartman J, et al. Benchmark data and power calculations for evaluating disease outbreak detection methods. *MMWR Morb Mortal Wkly Rep* 2004;(Suppl 53):144–51.
- Fawcett T. *ROC Graphs: Notes and Practical Considerations for Researchers. Technical Report*. Palo Alto, USA: HP Laboratories, 2004.
- Fawcett T, Provost F. Activity monitoring: noticing interesting changes in behavior. In: Madigan C, ed. *5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1999:53–62.
- Forsberg L, Bonetti M, Jeffery C, et al. *Distance-based Methods for Spatial and Spatio-temporal Surveillance*. Chichester, UK: Wiley, 2005.
- Zhang J, Tsui F-C, Wagner MM. Detection of outbreaks from time series data using wavelet transform. *AMIA* 2003:410–14.
- Kulldorff M, Huang L, Pickle L, et al. An elliptic spatial scan statistic. *Stat Med* 2006;**25**:3929–43.
- Kulldorff M, Heffernan R, Hartman J, et al. A space–time permutation scan statistic for disease outbreak detection. *PLoS Med* 2005;**2**:e59.
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*. 1988;**44**:837–45.
- Jacquenot G. Minimum distance between two polygons. In: Central M, ed. *An open exchange for the MATLAB and Simulink user community*, 2008. <http://www.mathworks.com/matlabcentral/fileexchange/22444-minimum-distance-between-two-polygons> [accessed 22 Mar 2011].
- Burrus CS, Gopinath RA, Guo H. *Introduction to Wavelets and Wavelet Transforms*. Upper Saddle River, NJ: Prentice Hall, 1997.