Learning Outbreak Regions in Bayesian Spatial Scan Statistics

Maxim Makatchev MAXIM.MAKATCHEV@CS.CMU.EDU
Daniel B. Neill NEILL@CS.CMU.EDU
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213 USA

Abstract

The problem of anomaly detection for biosurveillance is typically approached in an unsupervised setting, due to the small amount of labeled training data with positive examples of disease outbreaks. On the other hand, such model-based methods as the Bayesian scan statistic (BSS) naturally allow for adaptation to the supervised learning setting, provided that the models can be learned from a small number of training examples. We propose modeling the spatial characteristics of outbreaks from a small amount of training data using a generative model of outbreaks with latent center. We present the model and the EM-based learning of its parameters, and we compare its performance to the standard BSS method on simulated outbreaks injected into real-world Emergency Department visits data from Allegheny County, Pennsylvania.

1. Introduction

The spatial scan statistic (Kulldorff, 1997) has emerged as a technique of choice for cluster detection in such applications as disease outbreak detection and anomalous pattern detection in spatio-temporal data streams. The spatial scan maximizes a likelihood ratio statistic over a given set of space-time regions in order to find the most likely clusters. The choice of search regions is critical for timely and accurate event detection. However, most recent spatial scan methods make simple prior assumptions on the spatial and temporal distribution of events, and suffer from reduced detection power when these simplifying assumptions are incorrect.

The Bayesian scan statistic (BSS) method, proposed in (Neill et al., 2006), addresses some of the drawbacks of the original frequentist scan statistic and enables incorporation of prior knowledge of the effects and spatio-temporal characteristics of an outbreak. However, estimation of the parameters of a Bayesian model in the biosurveillance domain is hampered by the fact that there are only small amounts of labeled data available, rarely including the data from true outbreaks. One way to overcome this problem is to parameterize the model in a way that captures common properties of outbreaks, thus reducing the number of model parameters to learn. In this paper we extend the BSS method by developing a generative model for the spatial region affected by outbreaks. The model assumes a latent center of an outbreak, and the probability that each location is affected by the outbreak depends on its spatial distance from the center. The learned model can be used to compute the prior probability $p(H_1(S))$ that each spatial region S will be affected by an outbreak. By combining these region priors with the likelihood of the data in a Bayesian framework, we can compute the posterior probability that any given space-time region has been affected by an outbreak. Our new detection method was evaluated on simulated disease outbreaks injected into real-world Emergency Department data, demonstrating that learning of the region model substantially improves the timeliness and accuracy of outbreak detection.

2. Bayesian scan statistic

In the spatial event detection framework (Neill et al., 2006), we are given an observed count c_i and expected count b_i for each of a set of spatial locations s_i, and wish to detect spatial regions (sets of locations) where the observed counts are significantly higher than expected. The Bayesian scan statistic searches over a set of spatial regions S, computing the posterior probability that each region has been affected by an outbreak: $p(H_1(S) \mid D) \propto p(D \mid H_1(S))p(H_1(S))$, where D is the observed data. Similarly, we can compute the posterior probability of the null hypothesis of no outbreaks

Appearing in the Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

as $p(H_0 \mid D) \propto p(D \mid H_0)p(H_0)$. The likelihood of the data given the outbreak in a region $S \subseteq \{s_1 \ldots s_n\}$ is defined according to the Gamma-Poisson model conventional for disease modeling (Mollié, 1999): $c_i \sim \text{Poisson}(q_i b_i)$, and $q_i \sim \text{Gamma}(\delta\alpha, \beta)$. Here q_i represents the relative risk for location s_i, δ represents the multiplicative effect of the outbreak on the affected locations, and $\delta = 1$ for locations not affected by the outbreak. Expected counts (baselines) b_i are estimated from historical data using a 28-day moving average.

Typically, in an unsupervised setting the prior probabilities $p(H_1(S))$ are unknown and set to an uninformative distribution, for example uniform over all regions S under consideration. However, we expect some outbreak regions to be much more likely than others, depending on features such as region size, shape, and the set of locations affected (e.g. urban vs. rural areas). When labeled outbreak data is available, the prior probability of each possible outbreak region can be learned to improve the timeliness and accuracy of detection. Computing these region priors is especially relevant to the problem that arises in practical spatial scan implementations, considering only a restricted search space (e.g. circular or rectangular regions) for computational efficiency. In this case, there would likely be no scanned regions that match the outbreak region exactly, thus diluting the effect of the increased counts over a number of overlapping scanned regions and reducing the detection power of the algorithm. Estimating the prior probability of each scanned region based on its spatial proximity to likely outbreak regions has the potential to alleviate this problem.

Because there are an exponential number of possible outbreak regions (subsets of locations) and the number of training examples is small, we cannot learn a multinomial distribution over all possible regions, but must instead learn a model with fewer parameters. We assume a generative model with latent center; this model has few enough parameters to be learned from a small amount of data, yet can model many possible distributions of outbreak sizes, shapes, and commonly affected areas. We note that latent center models of an outbreak are common in the disease mapping literature (Lawson & Denison, 2002), but we use the model for outbreak detection rather than mapping disease risk. To the best of our knowledge, this is the first work which incorporates such a generative model into a spatial cluster detection framework.

3. Latent center model

Consider the following setting of the outbreak detection problem in metric space:

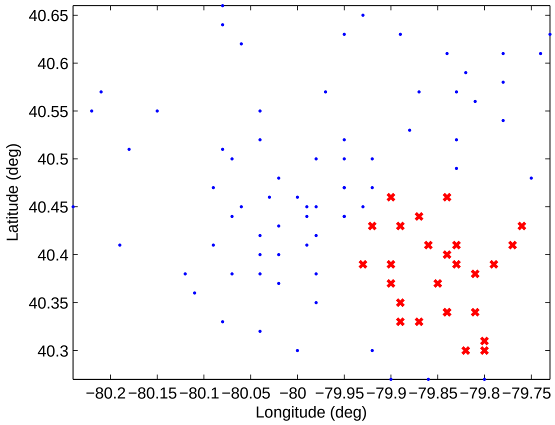


Figure 1. Map of zipcode centroids for Allegheny County, Pennsylvania. Crosses indicate locations affected by a simulated outbreak.

- A set of *locations* $s_1 \ldots s_N$ (e.g. zipcode centroids).
- A distance metric $d(s_i, s_j)$ on the set of locations.
- Each location has a corresponding observed count c_i and expected count b_i for the monitored time interval.
- An outbreak affects some region $S \subseteq \{s_1 \ldots s_N\}$; we assume that the set of locations affected is constant over the outbreak duration.

We represent an outbreak region $S \subseteq \{s_1 \ldots s_N\}$ by the values of N binary random variables $X_1 \ldots X_N$, such that $X_i = 1$ if $s_i \in S$ and $X_i = 0$ otherwise. In this supervised learning setting we treat these variables as observations in our training dataset. For each instance of an outbreak in the dataset, we assume that the X_i are generated by the following process:

- Choose $Y = s_j \sim \text{Multinomial}(1, \theta)$, the *center* of the outbreak region.
- Let r be an unknown parameter controlling the spread of the outbreak. We consider the cases when r is (a) a fixed parameter, (b) uniformly distributed in some interval $[r_1, r_2]$, and (c) drawn from $N(\mu, \sigma)$.
- Given the center of the outbreak $Y = s_j$ and the spread parameter r, $p(X_i \mid Y = s_j, r) \sim K(d(s_i, s_j), r)$, where K is some monotone function of r and the distance from the center. In this work we assume a sigmoid function centered at r and controlled with a steepness parameter h, so that $p(X_i \mid Y = s_j, r) = \left(1 + e^{\frac{d(s_i,s_j)-r}{h}}\right)^{-1}$. We use longitude and latitude as the Cartesian coordinates of each zip code centroid.

To allow full Bayesian estimation of the multinomial parameters θ, we assume a Dirichlet prior on $\theta = (\theta_1, \ldots, \theta_N)$ with parameters $\gamma = (\gamma_1, \ldots, \gamma_N)$. The graphical model representation for the case of a Gaussian distribution on outbreak radius is shown in Figure 2. Note that the variables X_i, denoting whether or not each location s_i is affected, are independent given the center of the outbreak and its radius.

4. Inference

Let $X = \{X^1, \ldots, X^M\}$ be M instances of observed outbreak regions, where the m-th outbreak region X^m is represented by a vector of values of binary random variables X_i: $X^m = (X_1^m, \ldots, X_N^m)$. Since an observation X^m d-separates the lower and upper parts of the graph in Figure 2, the observed counts do not play a role in the estimation of outbreak parameters θ and μ. Similarly, given the set of outbreak locations and their corresponding counts and baselines, the outbreak region model parameters θ and μ do not contribute to the inference about the effect δ of an outbreak. In this paper we use the same MAP estimates of δ across all detection methods; these estimates are described in (Neill et al., 2006).

The posterior probability of the parameters is $p(\theta, \mu \mid X) \propto p(X \mid \mu, \theta) p(\mu) p(\theta)$, and our goal is to maximize the log-posterior:

$$\begin{aligned}\log p(\theta, \mu \mid X) &\propto \log p(X \mid \mu, \theta) + \log p(\mu) + \log p(\theta) \\ &= \sum_{m=1}^{M} \log \sum_{n=1}^{N} p(X^m \mid Y^m = s_n) p(Y^m = s_n) \\ &\quad + \log p(\mu \mid \lambda, \eta) + \log p(\theta \mid \gamma). \qquad (1)\end{aligned}$$

Direct maximization over θ is possible in closed form:

$$\theta_n^{\mathrm{MAP}} = \frac{\sum_{m=1}^{M} z_{mn} + (\gamma_n - 1)}{N + M(\gamma_n - 1)}, \qquad (2)$$

where the responsibilities z_{mn} are

$$\begin{aligned} z_{mn} &= p(Y^m = s_n \mid X^m) \\ &= \frac{p(X^m \mid Y^m = s_n) p(Y^m = s_n)}{p(X^m)} \qquad (3)\end{aligned}$$

and

$$p(X^m) = \sum_{n=1}^{N} \int p(X^m | Y^m = s_n, r) p(r) p(Y^m = s_n) \mathrm{d}r.$$

We maximize over μ by maximizing the expected complete data log likelihood as a lower bound on the log likelihood $\log p(X \mid \mu, \theta)$. Namely,

$$\begin{aligned}\log p(\theta, \mu \mid X) &= \sum_{m=1}^{M} \log \sum_{n=1}^{N} \int p(X^m, Y^m = s_n, r) \mathrm{d}r \\ &\quad + \log p(\mu \mid \lambda, \eta) + \log p(\theta \mid \gamma) \\ &\geq \sum_{m=1}^{M} \sum_{n=1}^{N} \int z_{mnr} \log p(X^m, Y^m = s_n, r) \mathrm{d}r \\ &\quad + \log p(\mu \mid \lambda, \eta) + \log p(\theta \mid \gamma) \qquad (4)\end{aligned}$$

due to Jensen's inequality, where z_{mnr} are responsibilities defined as

$$\begin{aligned} z_{mnr} &= p(Y^m = s_n, r \mid X^m) \\ &= \frac{p(X^m \mid Y^m = s_n, r) p(Y^m = s_n, r)}{p(X^m)}. \qquad (5)\end{aligned}$$

The value of μ that maximizes the lower bound (4) can be obtained in closed form:

$$\begin{aligned}\mu^{\mathrm{MAP}} &= \left(\frac{\eta}{\lambda^2} + \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} \int \gamma_{mnr} r \mathrm{d}r}{\sigma^2} \right) \\ &\quad \cdot \left(\frac{1}{\lambda^2} + \frac{M}{\sigma^2} \right)^{-1} \qquad (6)\end{aligned}$$

The EM algorithm (see for example (Nigam et al., 2000)) alternates between the evaluation of the expected values of the hidden variables (responsibilities) z_{mn} and z_{mnr} (eqs. (3), (5)) and maximization over θ and μ (eqs. (2), (6)). We can find MLE estimates for the fixed and uniform radius models similarly.

In addition to learning the parameters of the distribution of the centers of the outbreaks θ and of the radius of the outbreak region μ we maximize the data log likelihood over the steepness parameter h using a generalized EM step. Gradient descent on h is especially prone to local optima, so we choose the final estimate of h among the candidates generated via generalized EM by comparing the respective average days to detection performance on the training dataset.

5. Evaluation

5.1. Simulated outbreaks

The dataset consists of the daily counts of respiratory Emergency Department visits in Allegheny County for the year 2002, aggregated by zipcode, with $N = 97$ zipcodes spread geographically as shown in Figure 1. The training data consists of 50 simulated outbreaks injected in the first 6 months of data, and the test set consists of 1000 simulated outbreaks injected in the last 6 months.

The simulation first generates an outbreak region and then augments the background (presumably outbreak-free) counts of Emergency Department visits in that region, by injecting counts corresponding to the cases of the outbreak. In this paper we use a simple linear model of the total number of injected cases, by drawing from Poisson(Δt), where $\Delta = 2$ is the rate of the outbreak's increase in severity and t is number of days from the beginning of the outbreak (this is similar to the FLOO model in (Neill et al., 2005)). Since our main goal is to evaluate the performance of the different outbreak region models, our simulations differ only in the way that the outbreak regions are generated.

5.1.1. Outbreaks satisfying modeling assumptions

We start with simulated outbreaks that follow the assumptions of one of the models: in the experiment shown in Figure 4(a) the outbreak is generated by selecting its center uniformly at random (i.e. each zipcode has equal probability of being the outbreak center) and then selecting the radius r uniformly at random from the interval $[0, 0.2]$. A location (zipcode) is affected if and only if it is within distance r of the outbreak center. Similarly, in the outbreaks corresponding to Figure 4(b) the radius r is drawn from the Gaussian distribution $N(0.15, 0.05)$.

5.1.2. Outbreaks violating the assumption on the distribution of the radius

In the subsequent simulations we increasingly deviate from the assumptions used in our generative models of the outbreak. In the simulation corresponding to Figure 4(c) we still select the outbreak center uniformly at random but instead of specifying the outbreak by a radius, we select k nearest neighbors of the outbreak center, where $k \sim U\{1, \ldots, 25\}$, which is equivalent to selecting an outbreak radius from a distribution that depends on the outbreak center and the density of the surrounding zipcode locations.

In the following simulation (results are shown in Figure 4(d)) we introduce noise to the outbreak boundary by performing an additional coin toss for each location, such that locations that are further from the radius are more likely to be affected if they are inside and less likely to be affected if they are outside of the circle of the radius r centered at the outbreak center. To make a fair comparison with the non-noisy boundary outbreaks we attempt to closely replicate the distribution on the outbreak radius induced by the k nearest neighbor generating procedure of the previous experiment. In particular we first select the k nearest neighbors as in the previous experiment and then set the radius of the outbreak r as the distance between the k-th nearest neighbor and the center of the outbreak. Once the r is found, we toss a biased coin for each location, with probability of a location being affected determined by the sigmoid function $\left(1 + e^{\frac{d-r}{h}}\right)^{-1}$, where $h = 0.1$ and d is the distance from the center of the outbreak.

5.1.3. Outbreaks with centers concentrated in the downtown area vs. the suburbs

In the experiments shown in Figures 4(e) and 4(f) we explore the influence of the density of the zipcode locations. In the outbreak simulation corresponding to Figure 4(e) we draw outbreak centers from a bivariate normal distribution discretized over the zipcode locations and centered at the geographical and density center of the zipcode map, with standard deviations equal to 0.1 along both axes. For the experiment shown in Figure 4(f) we invert the discretized normal distribution on centers of the outbreaks used in the previous experiment so that the outbreaks are more likely to be centered away from the geographical and density center of the zipcode map. Once the center is selected, the noisy boundary outbreak region is generated similarly to the outbreak corresponding to Figure 4(d).

5.1.4. Outbreaks violating the single-center assumption

In Figure 5(a) we show results of detection of outbreaks generated from two different centers. The resultant outbreak region is the union of the two regions each generated according to the model corresponding to Figure 4(c).

5.2. Results

We evaluate our models by comparing the performance of the original BSS method (using an uninformative region prior) with BSS variants that learn each of the three models of the outbreak region with different assumptions on the distribution of the region spread parameter r. In particular we consider the cases when r is modeled (1) as a fixed parameter, (2) as a random variable with uniform distribution $U[a, b]$, where a and b are estimated from the training data, and (3) as a random variable with normal distribution $N(\mu, \sigma)$, where μ is estimated from the training data and $\sigma = 0.1$. In each of the three models of the outbreak region, the latent distribution on centers θ is estimated from the training data as the maximum a-posteriori model assuming a multinomial distribution with Dirichlet prior. In a final experiment we compare the performance of the three models with their simpler versions that restrict the distribution on centers to the uniform distribution over zipcodes.

The results of each simulation are presented as an AMOC plot of average days to detection versus the number of false positives per year. These plots are generated by setting the threshold on the total posterior probability of an outbreak over the regions scanned by BSS at the specific levels of false positives per year (fp/year), and averaging the number of days since the beginning of an outbreak until the threshold is first exceeded. Since the discretization over the levels of false positives implies that the results will not vary for some adjacent levels of fp/year, for the sake of presentation in Figure 4 we group such fp/year levels together. In cases when the threshold is not exceeded within 14 days, the outbreak is effectively missed; we count such outbreaks as requiring 14 days to detection.

5.2.1. Outbreaks satisfying modeling assumptions

We expect each model to perform best when the outbreaks satisfy its modeling assumptions. For example, in the case of the uniform radius outbreaks shown in Figure 4(a), detection with the uniform model outperforms other models in the range of 5–25 fp/year. However, in the case of a Gaussian distribution on the outbreak radius (Figure 4(b)), the Gaussian model $r \sim N(\mu, \sigma)$ does not outperform the other methods. This is because the Gaussian model is sensitive to the mean μ estimated from the training data. The Gaussian mean estimate has high variance due to the small amount of training data. The performance of the fixed radius model suffers from a similar problem. The uniform model, however, appears to be robust to such estimation errors, and outperforms the other models despite the imprecision in the learned parameter values.

5.2.2. Outbreaks violating the assumption on the distribution of the radius

For the outbreaks that do not follow our modeling assumptions, results vary with the severity of the violation of the assumptions. For the case of the k-nearest neighbor outbreak region (Figure 4(c)) the uniform radius model performs better or comparably with the Gaussian radius model for the range of 5–25 fp/year, while it starts lagging behind the Gaussian radius model as noise in the region boundary is introduced (Figure 4(d)). Both the uniform and Gaussian radius models consistently outperform the fixed radius model, and in the ranges of 0–16 fp/year both models outperform the original BSS model (without region learning).

5.2.3. Outbreaks with centers concentrated in the downtown versus suburbs

When the outbreaks with noisy boundary are centered in the dense regions of the zipcode map, the methods perform worse than when the regions are centered in the periphery (Figures 4(e) and 4(f)). We explain this by the fact that the noise model is not scaled with the radius, so it has a larger effect on outbreaks of smaller radius. Incidentally the k-nearest neighbor outbreaks that are centered in the dense region of zipcodes tend to have smaller radius for the same distribution over k, than the outbreaks centered in the sparse zipcode regions (Figure 3). The Gaussian radius model demonstrated robustness to deviations in the radius distribution and consistently outperformed or matched the performance of the BSS without region learning, while the uniform radius model fell behind the BSS model for the dense zipcode outbreaks at levels of 17–25 fp/year.

5.2.4. Outbreaks violating the single-center assumption

For the two-center outbreak model with uniform distribution of the outbreak centers, while methods that model the outbreak region outperform the original BSS method at low levels of false positives (0–4 fp/year), the region modeling does not improve, or even reduces, performance at higher levels of false positives (Figure 5(a)). However, the differences in performance were small for the uniform and Gaussian radius models, while the fixed radius model performed poorly.

5.2.5. Multinomial versus uniform model of outbreak centers

Evaluating $N-1$ parameters of the multinomial distribution of outbreak centers from $M < N$ training samples without further restrictions is bound to overfit, hence in our last experiment we evaluate the benefits gained by using this complex model as opposed to using a uniform distribution over the zipcode centroids. The results of the three radius models using the uniform model versus the same three models using the multinomial model of centers detecting the outbreaks with centers concentrated in the downtown area are shown in Figure 5(b). For this particular distribution of outbreak centers that is correlated with the density of the zipcode locations, there are essentially no benefits of using the more complex multinomial model of outbreak centers. More complicated outbreak patterns, however, may require the richer model, and we will investigate this more fully in future work.

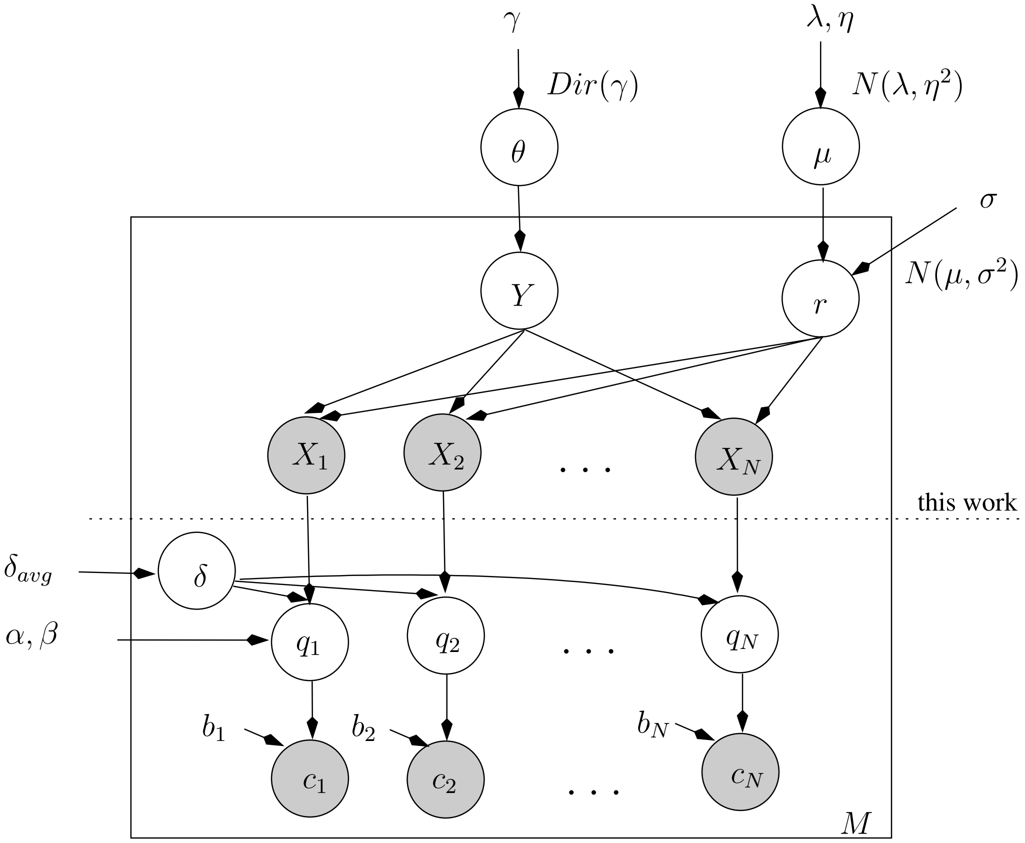
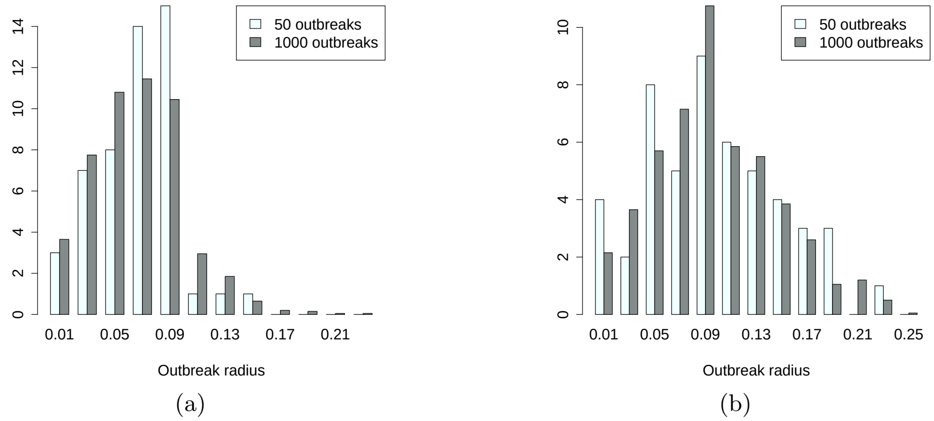


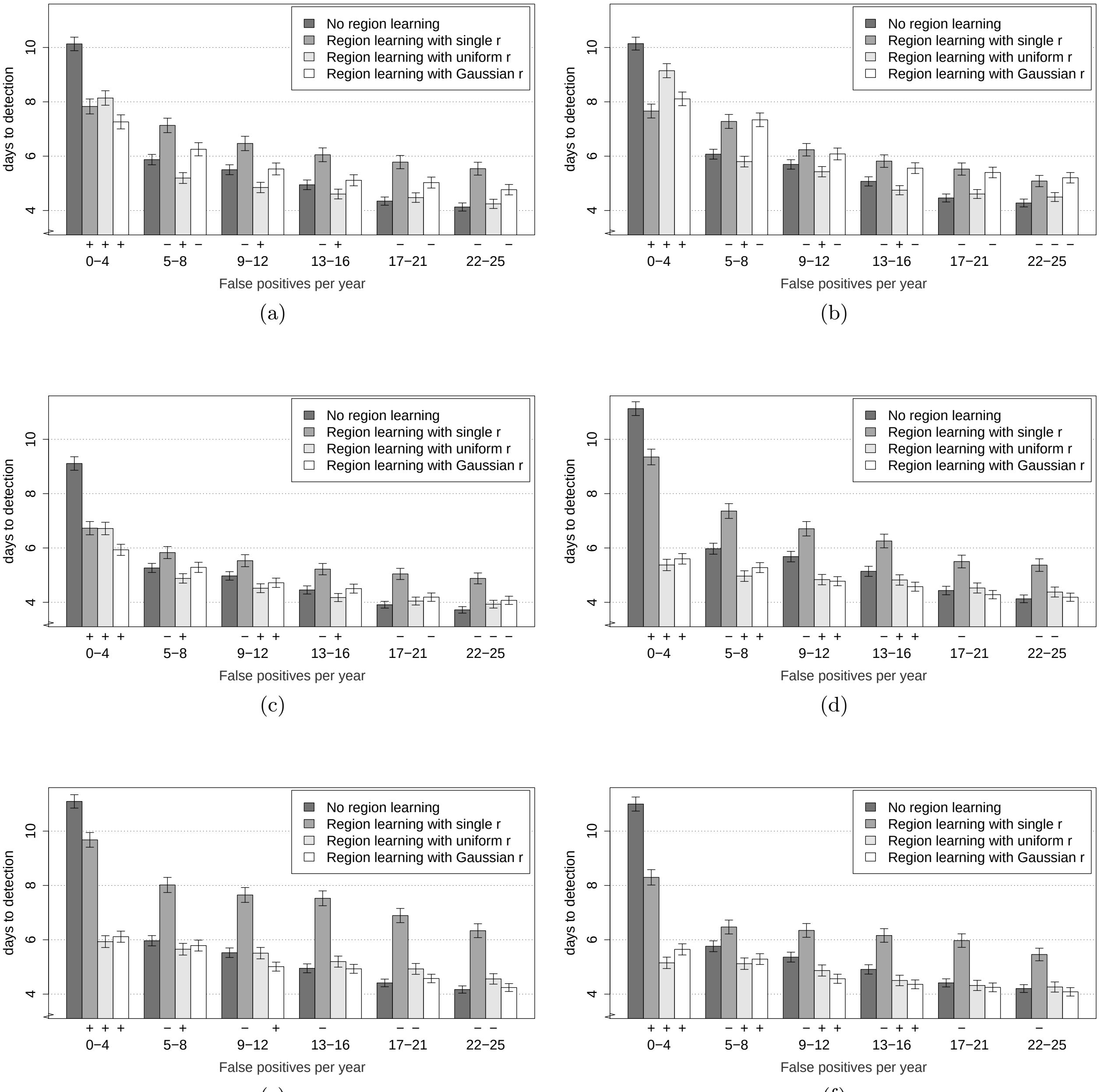
Figure 2. A latent center model of an outbreak region.



(a)

(b)

Figure 3. Normalized distributions of outbreak radii in the training ($M = 50$) and test ($M = 1000$) datasets for the outbreaks centered (a) in the dense zipcode (downtown) area and (b) away from the dense zipcode area (suburbs).



(a)

(b)

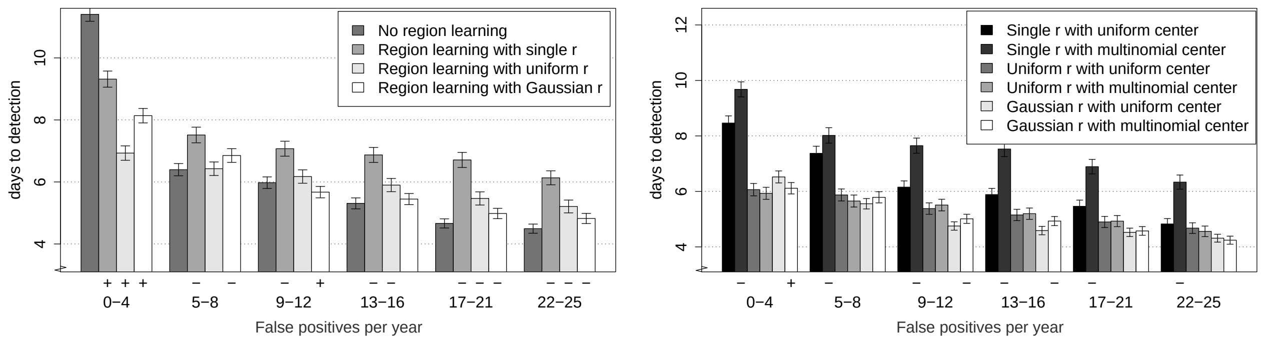
(c)

(d)

(e)

(f)

Figure 4. Days to detect versus false positives per year. Error bars indicate 95% confidence intervals. The + and − signs along the horizontal axis correspond to signs of significant differences between the respective bar and the first bar (BSS without region model) at $\alpha = 0.05$ according to two-sample t-test. The experiments cover different types of outbreaks: (a) Uniformly selected center of the outbreak and radius $r \sim U[0, 0.2]$; (b) Uniformly selected center of the outbreak and radius $r \sim N(0.15, 0.05)$; (c) Uniformly selected center of the outbreak, with $k \sim U\{1, \ldots, 25\}$ nearest neighbors affected; (d) Uniformly selected center of the outbreak, choosing $k \sim U\{1, \ldots, 25\}$ nearest neighbors to determine the radius r of the outbreak, and then tossing a biased coin for each location, with probability of a location being affected determined by the sigmoid function $\left(1 + e^{\frac{d-r}{h}}\right)^{-1}$, where $h = 0.1$ and d is the distance from the center of the outbreak (simulating noisy outbreak boundary); (e) Centers of the outbreaks are concentrated in the dense (downtown) location area, noisy outbreak boundary as in (d); (f) Centers of the outbreaks are concentrated away from the dense location area, noisy outbreak boundary as in (d). More details are in the text.



(a)

(b)

Figure 5. Days to detect versus false positives per year. Error bars indicate 95% confidence intervals. The experiments cover different types of outbreaks: (a) A union of two outbreaks each generated in the same way as the outbreak corresponding to Figure 4(d) (uniform distribution of centers, noisy outbreak boundary). The + and − signs along the horizontal axis correspond to signs of significant differences between the respective bar and the first bar (BSS without region model) at $\alpha = 0.05$ according to two-sample t-test; (b) The outbreak is identical to the one shown in Figure 4(e). For each of the three models of the radius distribution, we compare the version with uniform and with multinomial distribution of the outbreak centers. The + and − signs along the horizontal axis correspond to signs of significant differences between a method with the multinomial model of outbreak centers (even bars) and a respective method using a uniform model of outbreak centers (odd bars), at $\alpha = 0.05$ according to two-sample t-test. More details are in the text.

6. Conclusion

We address the problem of using a small amount of outbreak data to improve the performance of event detection methods by learning the distribution over affected regions. In particular, we augment the Bayesian scan statistic framework with a generative model of the outbreak region, attempting to capture the spatial connectivity, spatial size (spread) and spatial bias (urban vs. rural areas) of the outbreaks. Our results demonstrate that such region models can be estimated from small amounts of training data, significantly improve the time to detection as compared to the original Bayesian scan statistic method, and are robust to estimation errors even when the outbreaks significantly violate our modeling assumptions. The comparison between three models for distribution of the radius of the outbreaks demonstrates the improved performance of the uniform and Gaussian radius models as compared to learning only a single radius parameter, and also highlights the robustness of the uniform model to errors in parameter estimation. While the single radius model often underperformed the original BSS method, the uniform and Gaussian models tended to outperform the original BSS method by a large margin for low false positive rates (0–4 fp/year), with very similar performance for higher false positive rates.

Acknowledgments

The authors are grateful to Matthew Harrison and Larry Wasserman for their helpful feedback. This work was partially supported by NSF grant IIS-0325581 and CDC grant 8-R01-HK000020-02.

References

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, *26*, 1481–1496.

Lawson, A. B., & Denison, D. G. T. (Eds.). (2002). *Spatial Cluster Modelling*. Boca Raton, FL.

Mollié, A. (1999). Bayesian and empirical Bayes approaches to disease mapping. *Disease Mapping and Risk Assessment for Public Health*.

Neill, D. B., Moore, A. W., & Cooper, G. F. (2006). A Bayesian spatial scan statistic. *Advances in Neural Information Processing Systems 18* (pp. 1003–1010).

Neill, D. B., Moore, A. W., Sabhnani, M., & Daniel, K. (2005). Detection of emerging space-time clusters. *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* (pp. 218–227).

Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*, 103–134.