# A comparison of the frequentist and Bayesian spatial scan statistic in disease outbreak monitoring

Dongdong Lu

## Abstract

Early disease outbreak detection enables earlier action for public health measures. Early detection of disease outbreak can prompt early investigation of the causes such as naturally occurring epidemics, bioterrorist attacks, or environmental hazards and early implementation of public health measures. Daniel Neill put forward the Bayesian spatial scan statistic (BSSS) in 2006, claiming that it was a significant improvement over the traditional frequentist method. In this paper, we implement the BSSS method on different outbreak scenarios and evaluate the performances with measures such as detection time and detection rate. We have found that the frequentist scan statistic performs better than the BSSS in common metrics. Possible explanations are discussed. We also propose a hypothesis suggesting that under fair comparison requirements, the frequentist scan statistic cannot be beaten by other methods.

# Contents

# Introduction

Early disease outbreak detection enables earlier action for public health measures. Our goal is to detect clusters of disease cases, which can prompt earlier investigation of the causes such as naturally occurring epidemics, bioterrorist attacks, or environmental hazards. With timely and informed response, substantial reduction of life, economic and ecology cost can be achieved, while delayed or inaccurate responses can cost more.

Automatic surveillance systems potentially improve the efficiency and accuracy of responses by quickly detecting an emerging anomaly signal among the massive volume of public health data. To accurately detect and identify emerging patterns in a timely manner, an ideal disease outbreak detection system should monitor syndrome type (e.g., respiratory) in each spatial location (e.g., zip code) counts data daily. However, due to irrelevant events or other patterns in the data, the system can falsely sound the alarm. Therefore, we have two conflicting objectives to balance: making detection time short and keeping the number of false positives low.

Most scan methods have two hypotheses:

- $H_0$: there is no disease cluster in all locations
- $H_1$: there is a disease cluster in certain locations.

In 1997, Martin Kulldorff proposed the frequentist scan method [5], which soon became the most popular statistical tool for disease surveillance. This method assumes the disease cases are generated from a Poisson model and computes the likelihood ratio test statistic for a fixed set of

candidate clusters. The candidate cluster with the highest test statistic is known as the most likely cluster (MLC). Monte Carlo simulation is used to assess the significance of the MLC and secondary clusters.

In 2006, Daniel Neill extended Kulldorff's frequentist scan method to the Bayesian spatial scan statistic (BSSS) using a Gamma-Poisson model [9]. The Gamma-Poisson model has been used in disease mapping by Mollie [3], Clayton and Kaldor [8], to produce a smoothed map of disease rates. Similar to Kulldorff's frequentist scan method, BSSS assumes that two hypotheses $H_0$ and $H_1$ and disease counts are assumed to be generated by the Poisson distribution. However, the BSSS goes further in assuming the underlying risk is generated by a Gamma distribution. Instead of calculating the likelihood ratio statistic for a set of candidate clusters, it calculates the posterior probabilities of null and alternative hypothesis for each cluster and picks the one with the highest alternative posterior probability as the most likely cluster (MLC).

Neill et al. (2006) stated that the BSSS had the following advantages when compared with the traditional spatial scan statistic: 1. it uses prior information about the likelihood, size and impact of an outbreak, which can improve the detection power. 2. it uses a marginal likelihood approach rather than relying on the maximum likelihood estimates of these parameters, making the model more flexible and less prone to overfitting. 3. it does not need Monte Carlo randomization testing, and therefore can be performed at a significantly quicker speed.

In this project, we have implemented the code for the BSSS algorithm in R and compared its performance with the classic Kulldorff's spatial scan method using simulated Northeastern USA disease outbreak datasets. We evaluate the algorithms using the following criteria: days to

detection, percentage of outbreak detected, sensitivity, positive predictive value and accuracy (under 1 false positive/month). Since there is no real outbreak dataset available, we used a simulated dataset based on the Poisson model and manually injected outbreaks.

Overall, we have found that the frequentist method performs slightly better than the BSSS. Both methods share almost the same detection time, but the Bayesian method tends to produce more false positives and therefore drag down its performance in sensitivity, positive predictive value, and accuracy. In the Discussion section, we also propose a bold hypothesis: under fair comparison settings and the same false positive rate requirement, no other methods can perform better than the frequentist method.

# Method

We first introduce some notation, which we summarize in Table 1.

| Notation | |
|---|---|
| $\mathbf{c_i}$ | denote the disease counts in the i-th location |
| **D** | denote the collection of disease counts $c_i$ on the scanned day |
| $\mathbf{b_i}$ | denote the population of the i-th location |
| **q** | risk of an individual catching the disease |
| $\boldsymbol{G}$ | the set of all locations in the simulated data set (affected or not) |

*Table 1 Notation*

We assume that $c_i \overset{indep.}{\sim} \mathrm{Po}(qb_i)$, $i = 1,2,\dots,n$. Here $n = 1087$ in our simulation, since there are 1087 candidate clusters in the Northeastern USA.

## The frequentist scan statistic (frequentist method)

Kulldorff's spatial scan statistic is perhaps the most well-known statistical method for

disease outbreak detection. The method tries to identify clusters of cases using the likelihood ratio statistic between two hypotheses: $H_0$: all the regions have the same risk of disease and $H_1$: some regions have a higher risk than other regions. It assumes disease counts are Poisson-distributed with mean equal to the risk parameter times the baseline population $c_i \sim Po(qb_i)$. The goal is to find a collection of contiguous regions, denoted S, that have the most unusual collection of cases relative to those regions outside the cluster.

The spatial scan statistic [7] can be computed using the formula:

$$F(S) = \frac{P(Data|H_1(S))}{P(Data|H_0(S))} = \left(\frac{C_{in}}{B_{in}}\right)^{C_{in}} \left(\frac{C_{out}}{B_{out}}\right)^{C_{out}} \left(\frac{C_{all}}{B_{all}}\right)^{-C_{all}}$$

where $C_{in}$ is the total disease counts inside cluster S, $C_{out}$ is the total disease counts outside cluster S and $C_{all}$ is the total disease counts of all regions. $B_{in}$ is the total population inside S and $B_{out}$ is the total population outside S, and $B_{all}$ is the total population of all regions.

The space-time scan [6] extends the above spatial scan into three dimensional scenarios by using cylindrical window in three dimensions. The base of the cylinder is space and the height is time. The geographical base and starting dates are both flexible and we only consider those cylinder with end date of as the end of availability in the dataset. In this way, all the detected clusters are still alive and relevant for public health concerns. The likelihood statistic is calculated using the same method as for the purely spatial scan statistic. The time periodic surveillance can be done by repeatedly scanning the dataset at each time the dataset adds new counts.

We use the term grid to describe a dataset of disease counts in the designated spatial regions which can be real or simulated. The general detection process follows the following steps: we first

find the highest scoring candidate cluster $S^* = \text{argmax}_S F(S)$ of grid G, and its score $F^* = F(S^*)$. To determine the statistical significance of this candidate cluster, we use the Monte Carlo replication method: first we randomly create a large number R (i.e. 999) of replica grids by sampling under the null hypothesis $c_i \sim Po(q_{all}b_i)$; then we find the highest scoring candidate cluster and its score for each replica grid. Next, we find the p-value of $S^*$ which is $\frac{R_{beat}+1}{R+1}$ , where $R_{beat}$ is the number of replicas $G'$ with $F^*$ higher than the observed grid. If this p-value is less than a pre-set threshold (e.g. 0.05), we can report the candidate cluster as a significant spatial cluster; otherwise, no significant cluster is detected.

## The Bayesian spatial scan statistic (BSSS)

Daniel Neill proposed the BSSS as a natural extension of Kulldorff's spatial scan, by introducing the Gamma-Poisson model. The Gamma-Poisson model has been widely used in the disease mapping, and we primarily consider its utility in computing posterior probabilities. We now compute the null and alternative posterior probabilities of the dataset D. The risk was assumed to be a higher constant in the affected region while the risk was assumed to be a lower constant in outside regions.

Similar to the frequentist approach, we assume the same data distribution $c_i \sim Po(qb_i)$. But under $H_0$, we assume $q = q_{all}$ for all $s_i \in G$, where $q_{all} \sim Ga(\alpha_{all}, \beta_{all})$ and under $H_1$, we assume $q = q_{in}$ for all $s_i \in S$ and $q = q_{out}$ for all $s_i \in G - S$. In particular, the latter risks are drawn from two different gamma distribution respectively: $q_{in} \sim Ga(\alpha_{in}, \beta_{in})$ and $q_{out} \sim Ga(\alpha_{out}, \beta_{out})$. Computing the posteriors are straightforward: $P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D)}$ and $P(H_1|D) =$

$\frac{P(D|H_1)P(H_1)}{P(D)}$, where $P(D) = P(D|H_0)P(H_0) + \sum_S P(D|H_1(S))P(H_1(S))$. Here $P_1$ is prior probability of existing an ongoing outbreak which is to be chosen by the experimenter and $P(H_0) = 1 - P_1$ and $P(H_1(S)) = \frac{P_1}{N_{reg}}$. Let $E_{sample}\left[\frac{C_{all}}{B_{all}}\right]$ and $Var_{sample}\left[\frac{C_{all}}{B_{all}}\right]$ be the sample mean and variance of estimated risk across days for the baseline data. Then $\alpha_{all}$, $\beta_{all}$ can be calculated as $\alpha_{all} = \frac{\left(E_{sample}\left[\frac{C_{all}}{B_{all}}\right]\right)^2}{\left(Var_{sample}\left[\frac{C_{all}}{B_{all}}\right]\right)}$, $\beta_{all} = \frac{E_{sample}\left[\frac{C_{all}}{B_{all}}\right]}{Var_{sample}\left[\frac{C_{all}}{B_{all}}\right]}$. $\alpha_{in}$, $\beta_{in}$, $\alpha_{out}$, $\beta_{out}$ are calculated in the same ways. Finally, we can return all candidate zones with non-negligible posterior probabilities and pick the highest as our $F^*$ for the day and record the posterior probabilities of an outbreak.

To get the marginal likelihoods of the data under $H_0$ and $H_1$ respectively, we can integrate over all possible values of the parameters ($q_{in}$, $P(D)$, $q_{out}$, $q_{all}$) weighted by their corresponding probabilities. By the conjugate prior properties of Gaussian-Poisson model, we can compute a closed-form solution for these likelihoods:

$$P(D|H_0) = \int P\left(q_{all} \sim Ga(\alpha_{all}, \beta_{all})\right) \prod P\left(c_i \sim Po(q_{all} b_i)\right) dq_{all}$$

$$P(D|H_1(S)) = \int P\left(q_{in} \sim Ga(\alpha_{in}, \beta_{in})\right) \prod P\left(c_i \sim Po(q_{in} b_i)\right) dq_{in} \times \int P\left(q_{out} \sim Ga(\alpha_{out}, \beta_{out})\right) \prod P\left(c_i \sim Po(q_{out} b_i)\right) dq_{out}$$

Notice that we have the same type of integral to be computed in the above formulas, and let $C = \sum c_i$ and $B = \sum b_i$ for simplification:

$$\int \mathrm{P}\left(\mathrm{q} \sim \mathrm{Ga}(\alpha\,,\,\beta)\right)\prod\prod \mathrm{P}\left(\mathrm{c_i} \sim \mathrm{Po}(q\mathrm{b_i})\right)\mathrm{dq} = \int \frac{\beta^{\alpha}}{\Gamma(\alpha)} q^{\alpha-1}e^{-\beta q}\prod_{S_i}\frac{(qb_i)^{c_i}e^{-qb_i}}{(c_i)!}dq$$

$$\propto \frac{\beta^{\alpha}}{\Gamma(\alpha)}\int q^{\alpha-1}e^{-\beta q}q^{\sum c_i}e^{-q\sum b_i}dq \propto \frac{\beta^{\alpha}}{\Gamma(\alpha)}\int q^{\alpha+C-1}e^{-(\beta+B)q}dq$$

$$= \frac{\beta^{\alpha}\Gamma(\alpha+C)}{(\beta+B)^{\alpha+C}\Gamma(C)}$$

Apply the above results, we get:

$$\mathrm{P(D|H_0)} \propto \frac{(\beta_{\mathrm{all}})^{\alpha_{\mathrm{all}}}\Gamma(\alpha_{\mathrm{all}}+\mathrm{C_{all}})}{(\beta_{\mathrm{all}}+\mathrm{B_{all}})^{\alpha_{\mathrm{all}}+\mathrm{C_{all}}}\Gamma(\alpha_{\mathrm{all}})}$$

$$\mathrm{P(D|H_1)} \propto \frac{(\beta_{\mathrm{in}})^{\alpha_{\mathrm{in}}}\Gamma(\alpha_{\mathrm{in}}+\mathrm{C_{in}})}{(\beta_{\mathrm{in}}+\mathrm{B_{in}})^{\alpha_{\mathrm{in}}+\mathrm{C_{in}}}\Gamma(\alpha_{\mathrm{in}})} \times \frac{(\beta_{\mathrm{out}})^{\alpha_{\mathrm{out}}}\Gamma(\alpha_{\mathrm{out}}+\mathrm{C_{out}})}{(\beta_{\mathrm{out}}+\mathrm{B_{out}})^{\alpha_{\mathrm{out}}+\mathrm{C_{out}}}\Gamma(\alpha_{\mathrm{out}})}$$

Neill (2006) gives two methods for computing the score $\mathrm{F^*}$: 1. Bayesian maximum method: $\mathrm{F^*}$ is the maximum posterior probability $\mathrm{P(H_1(S)|D)}$ over all clusters 2.Bayesian total method $\mathrm{F^*}$ is the sum of posterior probabilities $\mathrm{P(H_1(S)|D)}$ over all clusters (the total posterior probability of an outbreak). The author points out the maximum posterior probability method generally performs best. So for our comparison, we also use this method for consistency.

# Benchmark Evaluations

## Background

Since real disease outbreak datasets are hard to find, we simulated fictional disease syndrome counts for 245 counties spread throughout Northeastern USA states: Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and the District of Columbia. Each county has a centroid coordinate. To generate a simulated cluster, we first start with a single county. We then simulate

an outbreak such that only that county has an outbreak of disease and the remaining counties do not have an outbreak. To create the next cluster, we then add the nearest neighbor county to the previous county. The next cluster has two counties. We then simulate an outbreak such that these two counties have an outbreak while the remaining counties do not. We continue to add counties to the previous cluster and simulate data in the same way until adding a new county to the previous cluster causes the cluster to have more than 1% of the total population across all regions. We repeat this process for each county in our study area. This results in 1087 different clusters. We simulate an outbreak data set for each cluster. Figure 1 is a map of the 245 counties with blue points at the centroids.

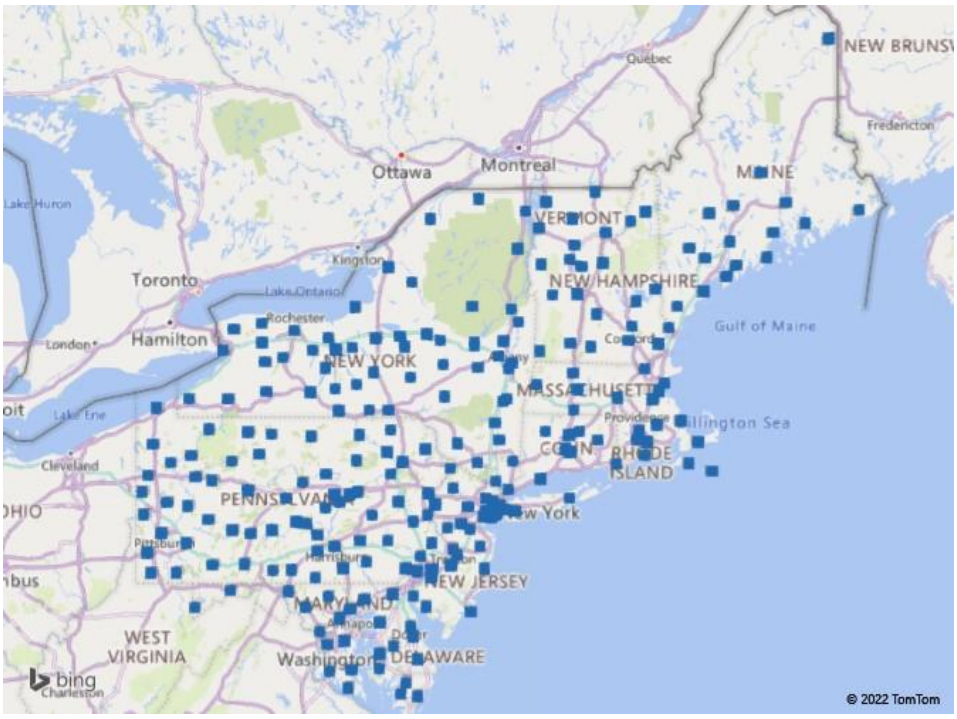


*Figure 1 NE USA Counties and their centroids (retrieved from [7])*

The Figure 2 is an example where we injected an outbreak in 4 nearby counties in Maryland: Harford (1), Cecil (2), Kent (3) and QueenAnne's (4). The color red means higher

intensity of disease counts, while the color green means lower intensity. Notice that since there is randomness in the data generation process, QueenAnne's (4) was injected with an outbreak (increased risk), it is not very red yet, while other counties with no ongoing outbreak may temporarily appear to be red.

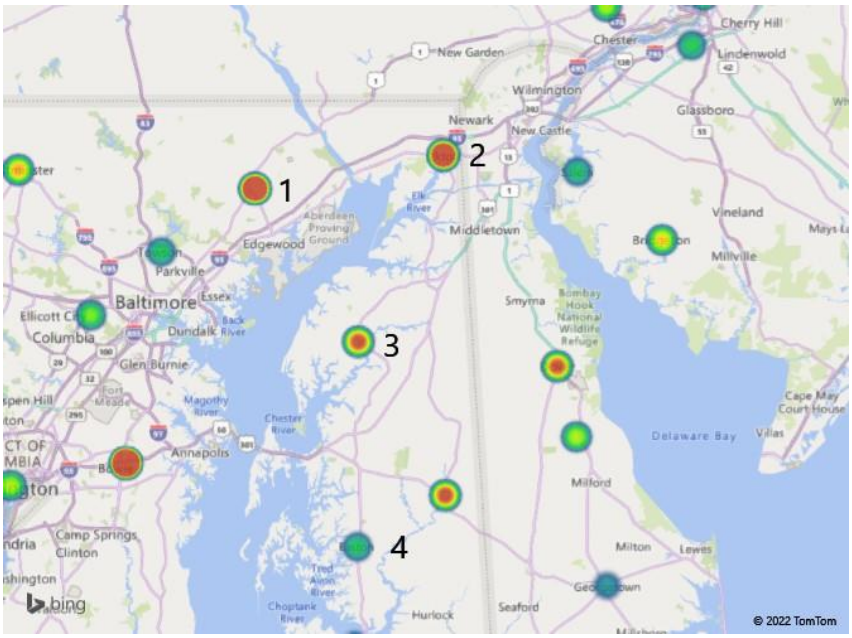


*Figure 2 Example of a Simulated Outbreak*

The original paper Neill (2006) used the "Fictional Linear Onset Outbreak" (or "FLOO") method to inject an outbreak. An $\mathrm{FLOO}(\Delta, \mathrm{T})$ outbreak is relatively simple: it generates $\mathrm{t}\Delta$ cases in each designated region on day $\mathrm{t}$ of the outbreak$(0 < \mathrm{t} \leq \mathrm{T}/2)$, then generates $\mathrm{T}\Delta/2$ cases per day for the remainder of the outbreak. This is the original outbreak simulation method stated in Neill (2006). However, we recognize the disease counts do not always increase linearly in the noisy real world. To compensate for this weakness, we propose our second outbreak simulation method: the "Revised Fictional Linear Onset Outbreak" (or "RFLOO"). The principle is the same

with FLOO, but it only assumes the risk is increasing gradually over time. Recall that disease cases are generated by Poisson distribution of risk times underline population. An increased risk only increases the possibility of getting a higher disease case count (which may not happen in certain locations on certain day). In this way, we can add more randomness to mimic the noise in real data, making it more similar with the real situation. The picture below shows how RFLOO generates more randomness in the added cases while compared with the original FLOO.

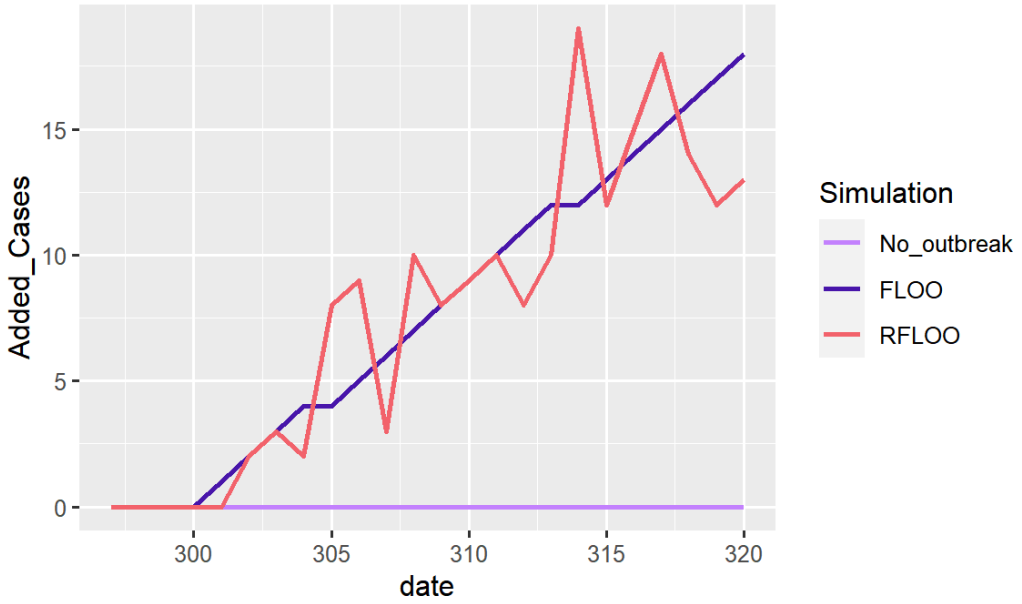


*Figure 3 Added Cases vs dates*

We injected 1 outbreak for every location cluster, which contains no more than 1% of the total population. That is 1087 outbreaks for testing each algorithm. This would examine their detection abilities in a full spectrum of geographic clusters.

## Implementation

The frequentist spatio-temporal scan statistic is based on [6] and the **scanstatistics** package

[1] implements in in the **`scan_pb_poisson`** function. To account for multiple comparisons, if the p-value $< 0.033$ (one false positive per month or $1/30$ chance in signaling a false outbreak on a given day), we return the detected cluster, and the day of detection. Since false negatives are rare (less than 0.5%), while summarizing the results, we marked the false negatives with detection on the last day of injection (day 30) with a p-value of 1 in the frequentist method (meaning they are not significant at all).

The Bayesian spatial scan outbreak detection algorithm is based on Neill (2006). We were also tempted to modify the **`scan_bayes_negbin`** function in **scanstatistics** to build the algorithm. However, it is clear from the first paragraph of Section 2.1 of Neill (2006) that each candidate cluster has a unique $\alpha_{in}$ and $\beta_{in}$. The **`scan_bayes_negbin`** uses a common $\alpha_{in}$ and $\beta_{in}$ for all candidate clusters. To see objectively evaluate the performance of the original paper, we decided to code the original algorithm with the assistance of three packages: **neastbenchmark**, **smerc**, **Brobdingnag**.

The **smerc** package was used to create the candidate clusters. Since we are working with early disease outbreak monitoring. I set the upper-bound of the population involved in the outbreak to be 1% of the overall population since a larger population is the opposite of the spirit of "early detection". If an outbreak is affecting areas with more than 1% total population, we do not need to use any algorithm – the news is everywhere already.

The **neastbenchmark** package [4] provided the northeast USA population and geographic data. The **Brobdingnag** package [10] is used since part of the calculation of posterior probabilities involves manipulating huge numbers, such as $\text{Gamma}(10000)$, which is beyond the management

ability of regular R session. This package has the power of remembering any big number in its log powers and performs all the calculations based on their powers.

We first run our spatial scan statistics for each day of the first 300 days to estimate baselines and priors, in addition we obtain the score $F^*$ for each day. Then we set prior probability of an existing outbreak to be 0.01 and the threshold of detection as the 96.7-th percentile of $F^*$ of null days, which is approximately 1 false-positive per month. The prior probability of an ongoing outbreak is selected to be 0.1. Next, we inject the outbreak into the last 30 days of the data and obtain the score $F^*(t)$ for each day $t$ of the outbreak. By recording the first date and location for which $F^*$ goes above the threshold, we can calculate the mean days-to-detection and false positive rate.

## Power-related measures

### Notation

We start by introducing some notation which is summarized in the below in Table 2.

| Notation | |
|---|---|
| $\widehat{\mathbf{Z}}_\mathbf{i}$ | detected cluster for the i-th simulated dataset |
| $\mathbf{C}$ | injected cluster |
| $\mathbf{n(C)}$ | population size of C |
| $\mathbf{p}^{(i)}$ | (Frequentist) Monte Carlo p-value for the largest test statistics related with the i-th benchmark dataset or (Bayesian) posterior probability |
| $\boldsymbol{G}$ | the set of all locations in the simulated (affected or not) |
| $\boldsymbol{\alpha_{freq}}$ | threshold below which to return the detection in frequentist setting |
| $\boldsymbol{\alpha_{bay}}$ | threshold above which to return the detection in Bayesian setting |

*Table 2 Notation*

### Days to detection

We have 4 power-related measures that can be used in a prospective disease surveillance

context. The most basic measure is the number of days to detection, i.e. the number of days until the outbreak is detected after it starts. The weakness of this measure is that it does not indicate whether the detected cluster overlaps with the true cluster or not. The detected cluster theoretically can have no overlap with injected cluster while system return a detection.

### Sensitivity

The sensitivity is defined as the proportion of the true cluster that overlaps with the detected cluster, averaged over all 1087 simulations.

$$\text{sensitivity} = \frac{1}{1087}\sum_{i=1}^{1087}\frac{n(\widehat{z_i} \cap C)}{n(C)} I(p^{(i)} < \alpha_{freq} \; or \; p^{(i)} > \alpha_{bay})$$

One potential weakness is that sensitivity is generally non-decreasing as the size of the detected cluster increases, without considering the number of "false positives" of the detected clusters. If there is no cluster detected by the end of the detection period (30 days), the detected cluster will simply be denoted as empty and the sensitivity will be 0.

### Positive predictive value (ppv)

The positive predictive value (ppv) calculates the proportion of the detected cluster that overlaps with the injected cluster, averaged over all 1087 simulations.

$$\text{ppv} = \frac{1}{1087}\sum_{i=1}^{1087}\frac{n(\widehat{z_i} \cap C)}{n(\widehat{Z_i})} I(p^{(i)} < \alpha_{freq} \; or \; p^{(i)} > \alpha_{bay})$$

The weakness of the ppv measure is that the detected cluster can only be a subset of the injected cluster, while all other locations with true outbreaks are not counted.

### Accuracy

The accuracy of a method is the proportion of the total population correctly characterized by the detection method, averaged over all 1087 simulations.

$$\text{accuracy} = \frac{1}{1087}\sum_{i=1}^{1087}\frac{n(\widehat{z_i} \cap C) + n\left(\widehat{Z_i}^{C} \cap C^c\right)}{n_A} I\left(p^{(i)} < \alpha\right) + \frac{n(C^c)}{n_A} I(p^{(i)} \geq \alpha)$$

Note that the accuracy requires two inputs: the population experiencing an outbreak that is correctly detected and the population not experiencing an outbreak that is not part of a detected cluster. If a cluster is not detected, then all the population not experiencing an outbreak is categorized correctly.

# Results

Regarding the days to detection (DTD), we have the following key observations: both methods share a median 8 days, but the Frequentist method has a mean of 8.23 days with a standard deviation of 3.86 days, which is better than the Bayesian method's mean of 9.40 days and standard deviation of 5.06. For example, using the Bayesian method, injections at zones 191, 411, 487, 544, 658, 688, 742 have 22 DTD while they only have 14, 13, 12, 9, 14, 11, 18 DTD using the Frequentist method. The density plot and boxplot for both methods' DTD performance can be seen in Figure 4. To the right side the days to detection, we see a higher pink line (Bayesian method) over the blue line (frequentist method), indicating we have more late-days (day 20 – 30) detection

for the Bayesian method.

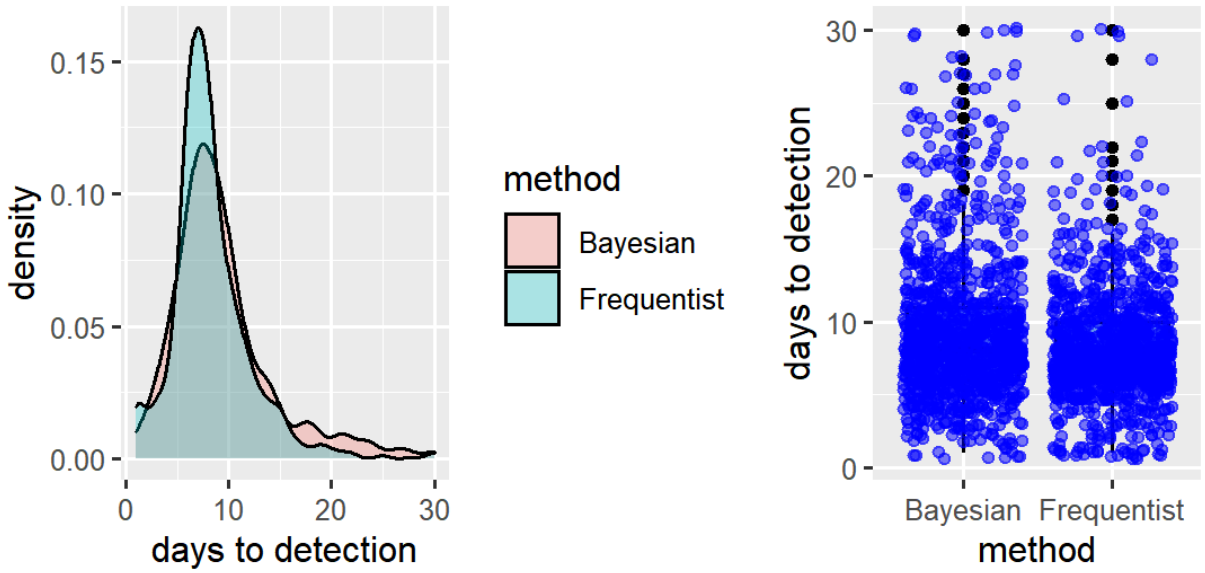


*Figure 4 Density plot and boxplot of days to detection*

The frequentist method fails to detect the outbreaks (false negatives) happened in the following clusters: 437, 720, 1002 by end of experimental period (day 30), while the Bayesian method produce detections for all the injections (not necessarily all correct). We describe an example of the Bayesian method's false positive in Table 3 below. While the injected cluster was a single region county 12, the detected cluster was counties 49, 50, 36, 12, 47, and 46. So even though the Bayesian method detected a cluster, it included only incorrect regions!

| **Injected Cluster** | **county: 12** |
|---|---|
| Detected Cluster | counties: 49 50 36 12 32 47 46 |
| Days to detection | 11 |
| Posterior probability | exp(-2.418417) |

*Table 3 An Example of False Positive for the Bayesian Method*

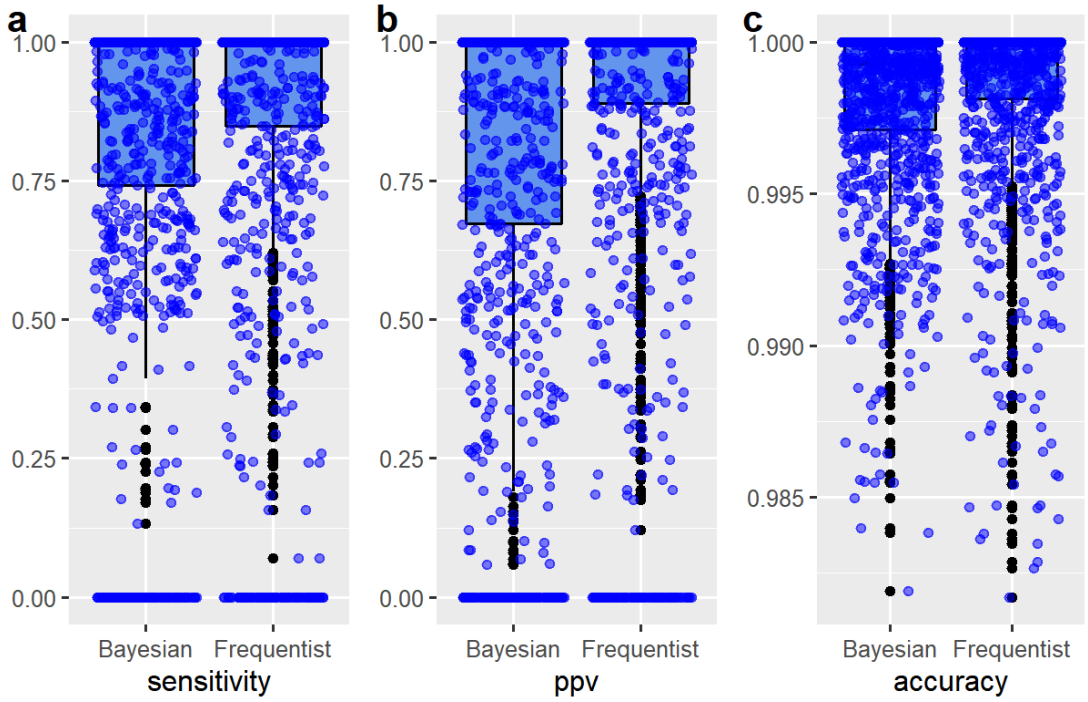


*Figure 5 Boxplots of sensitivity, ppv, accuracy for the two methods*

From the boxplots in Figure 5, we can see that the Bayesian method has significantly more data points falling below the lower quantile. This shows that false positives significantly drag down the Bayesian method's sensitivity, ppv and accuracy.

For sensitivity, ppv and accuracy, we can see from the below quintiles table in Table 4 that the Frequentist method is constantly performing better than the Bayesian method. For example, for sensitivity, the 30% quantile of frequentist method have reached 0.976 while it is 0.821 for Bayesian method; for ppv, the 30% quantile of frequentist method have reached 0.981 while the Bayesian method is only 0.781. Notice that since the major injections are correctly detected by both method which means all the above-mentioned three measures reach 100% (or near) around 50% quantiles. So we only need to make quantile tables from 0% to 50%.

| METRIC/METHOD | 0% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|
| sensitivity F | 0.0000000 | 0.2767776 | 0.7416916 | 0.9757980 | 1.0000000 | 1.0000000 |
| sensitivity B | 0.0000000 | 0.1738909 | 0.6512500 | 0.8205294 | 0.9757980 | 1.0000000 |
| ppv F | 0.0000000 | 0.3035883 | 0.7741308 | 0.9806508 | 1.0000000 | 1.0000000 |
| ppv B | 0.0000000 | 0.0648619 | 0.5391797 | 0.7810269 | 1.0000000 | 1.0000000 |
| accuracy F | 0.9816978 | 0.9950561 | 0.9974327 | 0.9987678 | 0.9994989 | 1.0000000 |
| accuracy B | 0.9819073 | 0.9938933 | 0.9963344 | 0.9977217 | 0.9987005 | 1.0000000 |

*Table 4 Quantile of sensitivity, PPV and accuracy for both methods (F: Frequentist, B: Bayesian)*

We can also see from the above three plots that the density of the sensitivity, ppv and accuracy are mostly distributed around the perfect "1" meaning the majority of injections are detected correctly. The Bayesian method is more left-skewed due to more false negatives it produced. Meanwhile the height of the blue line near "1" shows that the frequentist method successfully detects most injections. Regarding the running time, we have found it roughly takes 22 hours for the Bayesian method and 15 minutes for the frequentist method on an 8-core AMD Ryzen 5-2700X CPU with

parallelization.

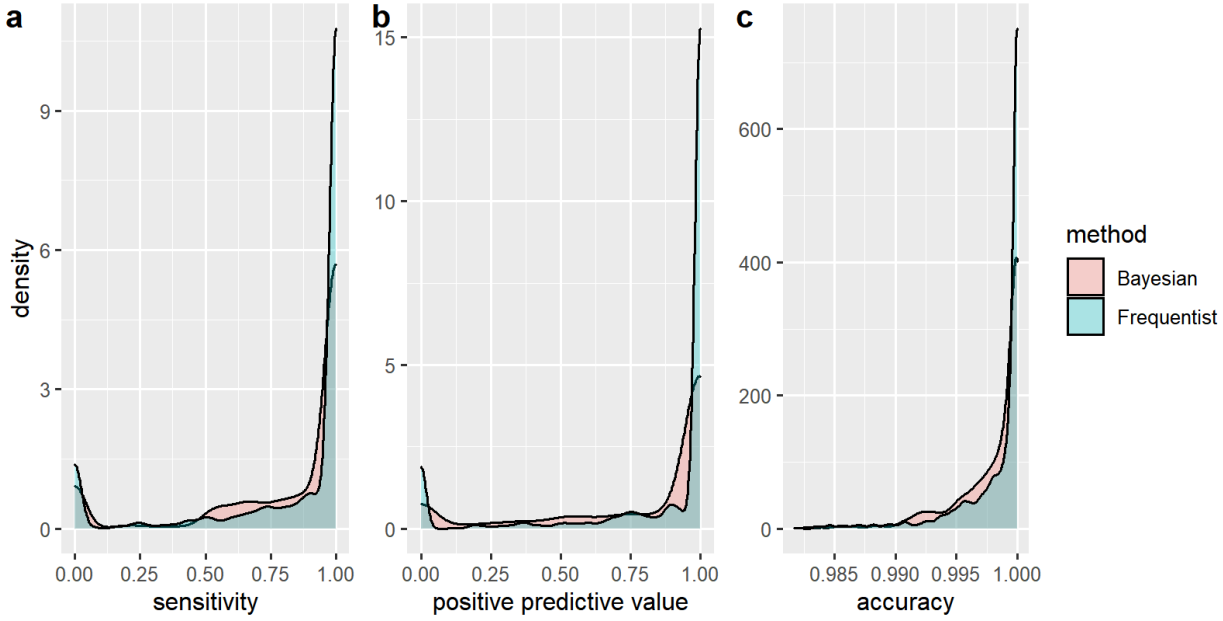


*Figure 6 Density plots for sensitivity, ppv, accuracy of two methods*

# Discussion

In this paper, we implement both of the frequentist spatial scan and the Bayesian spatial scan for the disease outbreak purpose on simulated data. We also came up with several slight improvements of outbreak simulation such as the revised fictional linear outbreak (RFLOO). Contrary to the claims in Neill (2006), we have found that the frequentist spatial scan performs better than the Bayesian spatial scan method in terms of common measures such as detection time, accuracy, sensitivity, ppv and computational time.

Why does Bayesian spatial scan method have more early false positives compared with the frequentist spatial scan method? Firstly, the frequentist spatial scan has the Monte Carlo replication

process to minimize the risk of multiple testing. However, the Bayesian spatial scan does not! Instead it runs the test individually on 1087 candidate clusters, no wonder there are so many early false positives. Secondly, we put forward a hypothesis that given the same false positive rate requirement, Kulldorff's frequentist method has already attained the best possible power. While we do not have a rigorous proof for it yet, the below is merely a rough sketch of the intuition behind our thinking. Kulldorff's frequentist scan method is not the original likelihood ratio test, but it is built by using the likelihood ratio, so it may still inherit certain essential property of the likelihood ratio. The Neyman-Pearson lemma [2] states that given the same false-positive (Type I error rate) rate, NO other test can do better than the likelihood ratio test in terms of power. Since the Bayesian method attempts to determine whether to reject the $H_0$ or not, this Bayesian algorithm is also a 'test'. Therefore, the Bayesian spatial scan statistic cannot beat the frequentist scan statistic in terms of power. Lower power means lower ability to reject $H_0$ when it is false, in our disease outbreak monitoring context, it translates into lower ability to declare the outbreak when it is really happening. So it would take a longer time for the Bayesian method to declare the outbreak, which is exactly what we saw in the simulation study, i.e., the average detection time is longer for the Bayesian method compared with the frequentist spatial scan.

The above hypothesis, if proven, could be significant result in this area. This means that the best method in disease surveillance has been found for a single information input of disease counts. Future development of this research field could only happen using multivariate surveillance techniques that is by combining multiple information inputs.

# Reference


[1] Allévius B (2018). “scanstatistics: space-time anomaly detection using scan statistics.” _Journal of Open Source Software_, *3*(25), 515. doi: 10.21105/joss.00515 (URL: https://doi.org/10.21105/joss.00515).

[2] Casella, G., & Berger, R. L. (2002). Statistical inference. Belmont, CA: Duxbury.

[3] Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized risks for use in disease mapping. Biometrics, 671-681.

[4] Joshua French (2021). neastbenchmark: Benchmark Data for Disease Clusters. R package version 0.2.

[5] Kulldorff, M. (1997). A spatial scan statistic. Communications in Statistics-Theory and methods, 26(6), 1481-1496.

[6] Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. Journal of the Royal Statistical Society: Series A (Statistics in Society), 164(1), 61-72.

[7] Microsoft Bing Map (2022, April 7). Northeastern USA Map. Retrieved from https://www.bing.com/maps

[8] Mollié, A. (1999). Bayesian and empirical Bayes approaches to disease mapping. Disease mapping and risk assessment for public health, 15-29.

[9] Neill, D., Moore, A., & Cooper, G. (2005). A Bayesian spatial scan statistic. Advances in neural information processing systems, 18.Rice, J. A. (2007). Mathematical statistics and data analysis. Belmont, CA: Thomson/Brooks/Cole.

[10] R. K. S. Hankin 2007. Very large numbers in R: Introducing package Brobdingnag R News 7(3)