Spatial disease anomaly detection under sparse data with stability metrics

Raphaella Diniz
Simon Fraser University
Canada
raphaella_diniz@sfu.ca

Renato Assunção
ESRI Inc.
U.S.A.
Universidade Federal de Minas Gerais
Brazil
rassuncao@esri.com

Pedro O.S. Vaz-de-Melo
Universidade Federal de Minas Gerais
Brazil
olmo@dcc.ufmg.br

ABSTRACT

The stability concept is a widely employed criterion for determining the optimal number of clusters in non-spatial datasets. This principle posits that an algorithm should successfully identify similar clusters across various perturbed instances of the data. In this study, we extend the application of stability concepts to spatial anomaly detection algorithms in the context of disease rate data. Our approach involves introducing novel methods for introducing noise to a singular instance of a disease map while preserving its spatial pattern, even in the absence of knowledge about the underlying distribution. Additionally, we present innovative metrics to assess the stability of spatial anomaly detection algorithms. Furthermore, we demonstrate the utility of the stability concept in analyzing real-world datasets. This involves aiding in the selection of the most suitable spatial anomaly detection algorithm for the given data, pinpointing the epicenter of the spatial anomaly, and quantifying the confidence level of the method for each area on the map. Importantly, our proposal is agnostic to the spatial configuration of the map, the underlying distribution of the data, the chosen spatial anomaly detection algorithm, or any prior information.

CCS CONCEPTS

• **Computing methodologies → Anomaly detection**; **Cluster analysis**.

KEYWORDS

Spatial statistics, Spatial disease cluster, Scan statistics, Cluster detection, Cluster stability.

ACM Reference Format:
Raphaella Diniz, Renato Assunção, and Pedro O.S. Vaz-de-Melo. 2024. Spatial disease anomaly detection under sparse data with stability metrics. In *1st ACM SIGSPATIAL International Workshop on Geospatial Anomaly Detection (GeoAnomalies'24), October 29-November 1 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3681765.3698466

1 INTRODUCTION

The significance of rapid epidemiological surveillance methods and improved predictive capacities has become more apparent in recent years, influencing the strategies employed by public health authorities in implementing epidemiological surveillance systems to track disease incidence. In conjunction with feedback from the general populace, these systems produce multiple localized spatial zones indicating heightened disease risk compared to the surrounding areas. For cancer surveillance, for instance, the current definition adopted by the US Center for Disease Control (CDC) is: "a cancer cluster is a greater than expected number of cancer cases that occurs within a group of people in a geographic area over a defined period of time" [1]. Suspect clusters require some response to community concerns and potentially a more thorough and costly investigation. Trumbo [44] reports that approximately 1,100 cluster investigations were requested in 1997 in the US while state health departments received from 1,300 to 1,650 requests to investigate suspected cancer clusters in the same period [15]. According to Thun and Sinks [43], the latest available public information, more than 1,000 inquiries about suspected cancer clusters must be responded to by USA state and local health departments annually.

The *spatial cluster detection* is an epidemiological and public health task that aims at detecting **geographically contiguous areas** that are *anomalous* with respect to certain reference framework, usually a generative probability distribution model [1, 21]. However, from the data mining (DM) and machine learning (ML) point of view, this is an unfortunate misnomer. Different from the usual clustering task in DM and ML, the objective is *not* to aggregate the small areas (such as Census counties) with similar risks into larger regions by partitioning the geographical map, as the map under analysis is known to have relatively homogeneous risk all over. However, there may be a few areas covering a small region with an increased risk. What is called spatial cluster detection is the detection of that *spatial anomaly*, the small region where the risk is larger than in the rest of the map. Therefore, a more appropriate name for this task should be *spatial disease anomaly detection* (**SDAD**). To avoid mixing the usual ML/DM clustering task with such spatial anomalies detection, we will avoid using "spatial cluster", the CDC chosen name.

Health authorities encounter a significant challenge due to the sparse distribution of cancer cases, often resulting in their concentration in specific locations and periods purely by chance, unrelated to underlying population risks. Boyle et al. [3] cite numerous real-life instances where extreme cancer incidence occurred in small areas. One such example is the occurrence of eight leukemia

ACM SIGSPATIAL '24, October 29 – November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1144-2/24/10
https://doi.org/10.1145/3681765.3698466

cases over 12 years in Fowey, an English village with only 2,320 inhabitants, yielding an incidence rate of 28.8 per 100,000, which is considerably higher than 5.0, the baseline rate [16]. Similarly, in Newquay, a small town on the north coast of Cornwall, England, with a population of 19 thousand, 17 cases were registered within a half-mile radius between 1948 and 1959 [49]. Despite extensive investigations, no underlying risk factors were identified in most of these instances, with chance remaining the most plausible explanation for the clustering of cases in spatial anomalies. Yet, inquiries unveiled that certain clusters stemmed from particular cancer-causing substances present in the environment. According to one review of 576 cancer cluster investigations, only 72 of the apparent clusters (less than 12%) confirmed a real increase in cancer rates [14]. However, public health authorities must address a large number of inquiries and respond to each one, even though most turn out to be false alarms.

The evaluation of **SDAD** algorithms typically involves assessing statistical power, which gauges the significance of the spatial anomaly identified by the algorithm, along with traditional machine learning (ML) metrics like precision and recall, which measure the algorithm's ability to detect the true anomaly. However, this framework proves inadequate for real-world scenarios due to three primary reasons.

The first reason pertains to the *spatial nature* of the dataset. Unlike the typical ML scenario with many independent instances, the data in spatial analysis consist of a single map divided into small areas, each containing counts of disease cases along with underlying population sizes. This setup resembles having a single time series from which to learn temporal patterns, rather than multiple independent realizations of the same stochastic process. In the spatial context, the conventional cross-validation approach based on training/test splits of the disease map dataset is not feasible. Attempts to create such splits by removing areas or cases from the map result in highly biased or distorted datasets. Incomplete maps used for anomaly detection introduce inefficiencies while case removal underestimates high-risk spatial anomalies. Consequently, there is no viable option for conducting a training/testing split, leading to overfitting and diminished generalization capacity in the learning process.

The second reason behind the misbehavior of metrics stems from the *nature of the data type* under consideration. Decisions are based on *disease rates*, which represent the ratio between the number of cases and the underlying population size in each area. This aspect also elucidates why traditional clustering ML methods, such as DBSCAN, are unsuitable for this problem. Each data point (i.e., each rate measured in a small area) exhibits different variance and substantial variability. Disease rates are highly sensitive to minor variations in the input disease count data, particularly in regions with small populations like rural areas. Even minor fluctuations of one or two disease cases in such areas can significantly impact disease rates, highlighting the widely varying variances among different areas. While areas with large populations typically exhibit stable and low-variance rates, sparsely populated regions experience substantial rate variations even with minor perturbations in disease counts. It is concerning that areas with the most extreme rates—whether high or low—attract the most attention on a disease map, yet they provide the least reliable information. In essence, data from different areas carry varying weights of evidence regarding the presence of disease clusters. We advocate for incorporating these distinct data evidence weights into the evaluation metrics.

The third reason for the inadequacy of metrics lies in the discrepancy between what is considered a "true spatial disease anomaly" and the evidence provided by the observed data. As outlined in Section 3, this fundamental challenge prompts us to adopt the clustering stability approach [46]. Consider an area known to belong to the disease cluster; however, when generating data, it may have zero disease cases in approximately 70% of instances. In other words, most of the time, there is no data evidence indicating that the area has a higher risk than the rest of the map. This scenario exemplifies the no-free lunch theorem [48] in action, signifying that no single spatial anomaly detection algorithm can be optimal for every possible data instantiation. The concept of stability proves useful by characterizing an area as part of a spatial anomaly if it consistently exhibits such characteristics across multiple spatial maps generated from the same data-generating process. In essence, a **SDAD** algorithm should yield similar results when applied to several spatial maps originating from the same underlying data.

In this paper, we tackle these challenges by proposing solutions grounded in the principle of cluster stability. Our contributions can be summarized as follows:

(1) **We introduce methods to generate several spatial maps from the same data-generating process as the single observed disease map.** Unlike conventional methods in **SDAD** literature that rely on artificial and probabilistic-based synthetic maps, our proposed methods—the spatial bootstrap and a rewiring technique—offer alternatives free of stochastic model assumptions. These methods enable cross-validation algorithms for **SDAD**. Importantly, we demonstrate that these random maps preserve the same spatial patterns as the observed data, ensuring that any unknown spatial signal present in the data, such as the potential existence of a true spatial disease cluster, remains unaffected. The differences between the observed and random maps arise solely from random disturbances unrelated to any true underlying risk reflected in the original data. This approach yields a substantial number of datasets, enhancing generalization capabilities.

(2) **We propose two metrics to evaluate the stability of SDAD algorithms.** Effective algorithms should identify a spatial anomaly that exhibits stability across numerous spatial maps generated from the same *unknown* spatial data-generating process as the single observed disease map. These metrics aim to gauge the robustness of methods against small data perturbations that do not affect the underlying generating data distribution, ensuring that the identified spatial anomalies remain consistent and reliable.

(3) **We propose a framework to quantify the weight of the evidence that an area actually belongs to the real anomaly.** In contrast to our second contribution, which focuses on evaluating the anomaly detection algorithm's stability, this framework assesses the confidence level regarding the inclusion of a detected area in the true underlying anomaly.

It offers a means to address the challenges posed by the no-free lunch theorem, wherein a seemingly inferior algorithm may actually be the optimal choice for a specific application. Our framework guides the selection of the most appropriate algorithm for a given scenario, providing clarity amidst the algorithm selection process.

2 RELATED WORK

It is important to note that our paper focuses on modeling non-infectious diseases, such as cancer surveillance. When modeling the spatial patterns of infectious diseases like COVID-19, probability models typically account for the contagious nature of the disease, where the presence of cases in one location increases the likelihood of cases in nearby locations. This requires models that capture spatial dependence or clustering, such as spatial autoregressive models or point process models that include interaction terms between individuals and explicitly account for temporal dynamics. In contrast, for non-infectious diseases like cancer, the occurrence of cases is generally independent of the presence of other cases in nearby areas. This leads to the use of probability models that assume spatial independence or incorporate explanatory covariates to capture environmental or demographic risk factors. Thus, while infectious disease models emphasize transmission dynamics and spatial diffusion, non-infectious disease models focus more on identifying underlying spatial risk factors without considering direct spatial interaction between cases.

The circular spatial scan statistic is the seminal statistical test and still the most popular algorithm for the spatial anomaly detection task [25]. It scans all circular regions over the map with different radius and outputs that one maximizing a likelihood function. The strength of this method is the delivery of a valid p-value for the null hypothesis of no spatial anomaly overcoming the multiple comparison problem. This p-value is computed through Monte Carlo hypothesis testing. Over the years, many different algorithms were presented, some having different shape restrictions [24, 34] and others searching for irregular shapes [9, 40].

Spatial scan statistics analysis has been applied in the past years in many fields such as public security, transportation, agriculture, and especially in public health. Examples of recent applications are the analysis of anemia in Ethiopia [10], malaria in Cambodia [37] and fluoride concentrations in Tanzania [17]. During the COVID-19 pandemic, spatial scan statistics analysis has been applied in many scenarios. Cordes and Castro [6] analyzed the relationship between COVID-19 testing rates and positive cases in New York City while Escobar et al. [11] investigated the impact of race/ethnicity on SARS-CoV-2 testing, infections, and outcomes in Northern California. Many other studies were published evaluating hotspots of COVID-19 cases on several countries [18, 32]. Beyond the detection of spatial disease clusters, clustered anomaly detection using scan statistics is also applied in different computer science-related areas. del Gobbo et al. [7] and Camacho et al. [5] applied scan statistics in geolocated Twitter data to discover spatial anomalies of users' opinions while Nguyen [35] aimed at detecting rumors spatially using social media data. Scan statistics algorithms are also widely employed in pattern and anomaly detection in graphs [4, 13, 47**?**].

As a reviewer pointed out, there is a potential connection between the traditional machine learning literature on clustering methods specialized for spatial objects and our problem of spatial anomaly detection. Some studies have explored spatial clustering with uncertain data. For instance, [50] proposed a model in which statistical significance can be assigned to clusters detected by DBSCAN. Another approach is representative clustering [51], which samples possible worlds, performs clustering on each, and then identifies clusters that consistently appear across multiple worlds.

A further approach involves Bayesian modeling of the spatial partitioning of maps based on rates [41, 42]. This method generates a posterior probability distribution of spatial partitions, allowing possible partitions to be ranked according to their probability.

In a recent position paper, Sculley *et al.* made an impacting statement about how empirical rigor is not keeping pace with advances in ML [38]: "The goal of science is not wins, but knowledge." An important aspect of rigorous empirical evaluation is to assess the limits and shortcomings of evaluation metrics as well as the proposition of more adequate metrics [12, 28]. Like us, many others are taking a step back and reviewing the current state of the art of different areas in ML through rigorous and thorough empirical evaluations [2, 19, 29, 31, 36, 39, 45]. The **SDAD** task, in particular, presents unique challenges and problems that necessitate new evaluation procedures [8]. In this paper, we discuss the peculiarities and the evaluation metrics used in the **SDAD** task. While [8] revisit traditionally used metrics for this task, we propose new ones.

Stability is a concept applied to find an appropriate number of nonspatial anomalies in a dataset [27, 46]. It was also considered to evaluate clustering solutions in particular data sets, such as time series [20], images [33], and cell data [30]. An algorithm should be able to detect nearly the same cluster under a slight perturbation of the data. To the best of our knowledge, this concept has never been explored before for the spatial anomaly detection task. However, it can be very useful, mainly when dealing with sparse data. Beyond the traditional metrics for the **SDAD** task, a spatial anomaly detection algorithm should also be evaluated for its ability to find the same spatial anomaly under slight perturbations. This work proposes spatial anomaly stability as a necessary checkpoint for anomaly interpretation and introduces a new systematic and standardized analytical framework for the assessment of spatial anomaly detection results.

3 DEFINITIONS AND PROBLEM STATEMENT

Consider a map partitioned into N distinct areas, denoted by $i = 1, \ldots, N$. Each area i is characterized by its underlying population size n_i and the number of observed disease cases y_i within a specific time frame. We model the occurrence of disease cases as a probabilistic distribution, such as $y_i \sim \text{Poisson}(\theta_i n_i)$, where θ_i represents the unknown incidence rate. The map is split into two disjoint sets: a contiguous subset of areas, denoted as a spatial anomaly C, where $\theta_i = \theta_1$, and the remaining areas where $\theta_i = \theta_0$, with $\theta_1 > \theta_0$. The objective of a **SDAD** algorithm is to identify a subset $\hat{C}$ of areas solely based on the underlying population n_i and observed case counts y_i, aiming to accurately identify $\hat{C} = C$. In this study, we focus on detecting the presence of a single spatial anomaly. Consequently, both C and $\hat{C}$ consist only of connected areas, ensuring that it is possible to travel between any pair of areas within each set while passing exclusively through areas within the set.

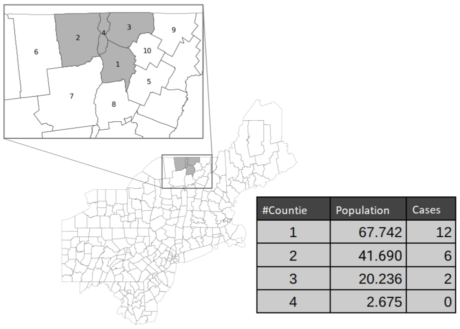
All **SDAD** methods detect a geographical zone $\hat{C}$ if the associated statistical test yields significance, typically indicated by a p-value below a specified threshold, often set at 0.05 or 0.01. If the test fails to reach statistical significance, no spatial anomaly is detected, and the algorithm returns $\hat{C} = \emptyset$.

SDAD methods fall within the category of unsupervised classification. Unlike supervised classification, there exists no ground truth against which to validate clustering outcomes. To evaluate the performance of **SDAD** algorithms, synthetic datasets are commonly employed due to the absence of labeled maps with true disease clusters. These datasets introduce an artificial spatial anomaly into the map, where a probabilistic distribution for disease counts y_i in each area is defined to yield a higher incidence rate θ_i within the spatial anomaly. A substantial number K of independent dataset instances, each containing the same spatial anomaly, is generated, distributing disease counts across the map according to the defined distributions. The algorithm is then executed on each instance, resulting in the detection of an anomaly $\hat{C}_k$ in instance k. Performance metrics such as precision, recall, and statistical power are computed by averaging the results across instances. In Section 5, we detail the benchmark dataset widely employed for evaluating **SDAD** algorithms. This benchmark comprises numerous map datasets featuring spatial anomalies of various sizes located within regions characterized by rural (and small), mixed, or urban (and large) population.

3.1 Evaluation difficulties

Recognizing the necessity for rethinking evaluation measures, consider a scenario where the true anomaly comprises four areas situated in a sparsely populated rural region (see Figure 1). We simulated disease counts with a much larger relative risk within the spatial anomaly. Even so, area 4 had zero observed cases in 70% of the simulated instances (one such instance is shown in the figure). That is, in the majority of instances, there exists no evidence indicating that this anomaly area belongs to a genuine spatial anomaly, as the observed count is zero, the smallest possible. Despite belonging to the high-risk spatial anomaly, this area lacks disease cases. Such occurrences are natural consequences of small populations and are not uncommon.

The issue arises from the observation of zero disease cases in an area, providing no evidence that its population faces a high-risk situation. Strictly based on observed data evidence, we should refrain from including area 4 in a spatial anomaly unless we possess prior reasons to incorporate it into $\hat{C}$. For instance, strong prior beliefs dictating that the cluster must exhibit a circular shape may compel the inclusion of areas with zero case counts into a cluster estimate. A sparsely populated area surrounded by regions with a high number of cases may prompt the aggregation of the zero count area into the cluster, considering its small population. However, the inclusion of a single additional case can lead to varying detected clusters depending on the **SDAD** algorithm utilized. Real-world scenarios often involve slight data variations, such as case registration errors or population count errors stemming from outdated records. Therefore, **SDAD** algorithms should demonstrate robustness to such minor variations, consistently identifying the same spatial anomaly.



#Countie	Population	Cases
1	67.742	12
2	41.690	6
3	20.236	2
4	2.675	0

Figure 1: Map with the NE United States counties with their female population and the corresponding disease cases.

The situation becomes more intricate when assessing the **likelihood** of area 4 belonging to the spatial anomaly. Firstly, even if it belongs to the cluster and exhibits a higher risk than the rest of the map, observing zero disease cases is the most probable outcome in this area. Its population is so minuscule that it would require an immense relative risk to potentially produce even a single case. Secondly, complicating the analysis further, the probability of observing zero cases is greater under the null hypothesis than under the alternative hypothesis. In other words, even if area 4 belongs to the spatial anomaly and possesses an underlying higher risk than the rest of the map, observing zero cases is more indicative of the *not in the spatial anomaly* scenario than of the *in the spatial anomaly* scenario: $\mathbb{P}(y_i = 0 \mid i \notin C) > \mathbb{P}(y_i = 0 \mid i \in C)$.

In the first row of plots in Figure 2, the red line represents the probabilities $\mathbb{P}(y_i = k | i \in C)$ for the four cluster areas depicted in Figure 1, with the horizontal axis indicating the possible number of disease cases k. The blue line connects the probabilities $\mathbb{P}(y_i = k | i \notin C)$, representing the probabilities of each possible count k if the area did not belong to the spatial cluster. A vertical dashed line indicates the number of cases observed in these areas in the simulation illustrated in Figure 1. In the first three areas, the likelihood of the observed count y_i is higher under the spatial anomaly hypothesis than under the null hypothesis, aligning with our expectations. However, in area 4, a peculiar behavior is observed: the value $y_4 = 0$ is more probable under the null hypothesis than under the spatial anomaly hypothesis, *despite our prior knowledge that area 4 belongs to the spatial anomaly.*

In summary, the crux of the issue lies in the fact that observing zero counts in this area is a very common occurrence. Indeed, out of the 10,000 simulations of this spatial anomaly, $6,947$ instances resulted in $y_4 = 0$. This leads to a puzzling scenario where an area is part of a high-risk zone, yet there is a 0.7 probability of observing zero cases there. This discrepancy arises because the definition of spatial anomaly is based on the generative probabilistic process that generates random cases in each area, rather than on the realized data instantiation.

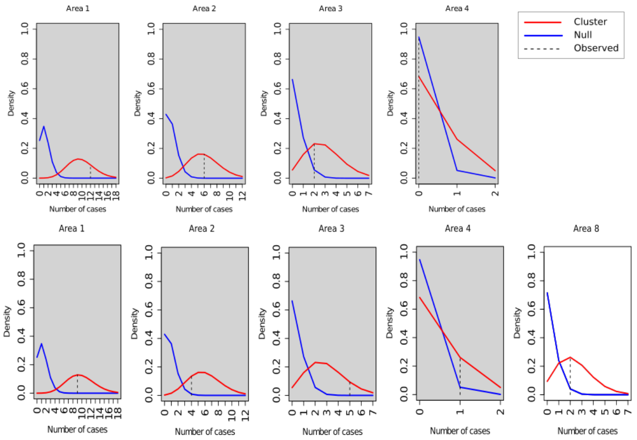


Figure 2: First Row of Plots: $\mathbb{P}(y_i = k|i \in C)$ (red line) and $\mathbb{P}(y_i = k|i \notin C)$ (blue line) are depicted for the four areas in Figure 1. The vertical dashed lines represent the observed cases. Second Row of Plots: Observed values in a different simulation, along with a fifth area neighboring the anomaly but not belonging to it.

3.2 Precision and recall issues

The **precision metric** is also influenced by these random fluctuations. The second row of plots in Figure 2 illustrates another simulation under the same hypothesis of a high-risk spatial anomaly in the four areas. Observe that the random number generation produced $y_4 = 1$ this time, a situation more aligned with the hypothesis that area 4 may belong to the spatial anomaly. However, we also evaluate the result $y_8 = 2$ in area 8, a neighboring area of this spatial anomaly. In this instance, although area 8 is not part of the cluster, its observed value is more compatible with the spatial anomaly hypothesis than with the null hypothesis, owing to the nature of the generative process and its small population.

Finally, consider the **recall metric** calculated for the cluster composed of the four areas shown in Figure 1. Recall is estimated as the average of $|C \cap \hat{C}|/4$ over thousands of independent simulations of this ratio. However, one of the areas in the denominator, area 4, has zero cases in approximately 70% of the simulations. The decision to count or not count this area in the numerator has a significant impact, as it can add or subtract 0.25 from a metric bounded between 0 and 1. Thus, we advocate that an area with zero disease cases ($y_i = 0$) should not heavily penalize any **SDAD** method if it returns $i \notin \hat{C}$ (if it leaves area i out of the detected cluster). The rationale is that a data-driven algorithm, one not based on prior knowledge about the true spatial anomaly shape, should not be excessively penalized for failing to detect such an area. Hence, there is a need to rethink the evaluation metrics for this **SDAD** task.

In the next section, we adopt a principled probabilistic approach to redefine the usual metrics for this task, giving less weight to regions where the disease incidence rate varies dramatically due to small data fluctuations.

4 STABILITY METRICS FOR SDAD

4.1 Sampling methods

In typical ML applications, we often partition the dataset into training and testing samples to assess the results *without imposing any distributional assumptions about the data generation process*. This evaluation approach is application-specific and entirely data-driven, bypassing the need for synthetic datasets. By doing so, we mitigate the risks of overfitting and counteract biased outcomes. However, given the unknown nature of the true anomaly and the underlying distribution, generating additional instances under the same model to evaluate various **SDAD** methods becomes non-trivial. This presents a unique challenge in assessing the efficacy of such methods.

Here, we introduce two novel methods for generating additional instances of a single disease map: the bootstrap and the rewiring samples. The underlying concept in both methods is to generate maps for multiple rounds of cross-validation. The maps originate from the same statistical population and preserve the same spatial pattern observed in the original dataset, *regardless of our knowledge about the true probability distribution generating the observed data.*

4.1.1 Bootstrap samples. Let $Z_{ij} = 1$, if the j-th individual in area i is a disease case, and $Z_{ij} = 0$, otherwise. With a disease rate (per capita) equal to θ, we have a binomial distribution for the number of cases:

$$(y_i|n_i) = \sum_{i=1}^{n_i} Z_{ij} \sim \text{Bin}(n_i, \theta) \approx \text{Poisson}(n_i\theta) \ .$$

The Poisson approximation is valid when θ is close to zero, which is usually the case for typical diseases such as site-specific cancers. Independently for each j-th individual in area i, generate the binary indicator $W_{ij} \sim \text{Bernoulli}(\pi)$ using a spatially constant probability hyper-parameter $\pi \in (0, 1)$. Then, conditionally on n_i, we have:

$$(n_i^*|n_i) = \sum_{j=1}^{n_i} W_{ij} \sim \text{Bin}(n_i, \pi)$$

and

$$(y_i^*|n_i) = \sum_{j=1}^{n_i} W_{ij} \cdot Z_{ij} \sim \text{Bin}(n_i, \pi\theta)$$

The collection $\{(n_i^*, y_i^*), i = 1, \ldots, N\}$ is a randomly thinned version of the original map with $\{(n_i, y_i), i = 1, \ldots, N\}$. The map with this π-thinned random sample of cases and population is called a *bootstrap sample* or *bootstrap map*).

The important point is that, if there is any spatial anomaly in the original data, it is randomly reflected with the same spatial pattern in the bootstrapped map. The differences are due to random fluctuations not associated with the intrinsic disease rates θ_i. They are randomly thinned versions of the original map. Indeed, suppose that some areas have high incidence disease rates producing a map with spatially varying rates. Let $y_i \sim \text{Bin}(n_i, \theta_i) \approx \text{Poisson}(n_i\theta_i)$. As the bootstrap probability π does not vary spatially, the observed map spatial pattern is reflected on the bootstrap map: $y_i^* \sim \text{Poisson}(n_i^*\pi\theta_i)$. Hence, the odds between disease rates from any pair of areas remain the same as in the original map. For example, while the odds between rates from areas i and j is θ_i/θ_j in the original map, it is equal to $(\pi\theta_i)/(\pi\theta_j)$ in the new bootstrap samples. The spatial pattern of the disease rates θ_i is retained in the bootstrapped map, whatever this spatial pattern is.

By randomly sampling W_{ij} independently, we generate B bootstrap maps with different disease and population counts from our original data map. Each bootstrap disease map in this stack of B maps retains the same spatial pattern as the original map but differs in two respects. First, it has a smaller number of cases and population. Each area has approximately a proportion π of the original cases and population counts. Second, as the selection of retained cases and population is random, the bootstrapped maps differ. In short, this technique generates different maps from the real data but retaining the same spatial pattern, so one can test the resilience of the **SDAD** algorithms with them. We show how to explore this in Section 4.2.

4.1.2 Rewiring. Rewiring is a different way to generate pseudo maps that retain the spatial pattern from the original map but provide enough diversity to allow generalization capabilities for the algorithms. The population size is not altered, only the number of cases of each area can change. In contrast with the bootstrap approach, the total number of cases in the map, $\sum_i y_i$, is kept constant. The only change is that some of the observed cases may be randomly assigned to neighboring areas. The main idea is that each area may rewire a small proportion of its cases to neighboring areas. As all areas are rewiring, the expected number of cases in each area is kept approximately constant.

Let $\pi \in (0, 1)$ be a hyper-parameter. Rather than the true observed number y_i of cases in each area, we generate a random number $y_i^* = K_i + R_i$ by keeping approximately a proportion π of its original count y_i (the K_i count) plus additional cases coming from the neighbors (the R_i count). More specifically, $K_i \sim \text{Bin}(y_i, \pi)$. The neighbors of area i receive the residual $y_i - K_i$ cases. This distribution is made according to a multinomial distribution, with the probability of selecting a given neighboring area proportional to its population. To be precise, let V_{ij} be a binary indicator with $V_{ij} = 1$ if areas i and j share boundaries, and $V_{ij} = 0$ otherwise. We set $V_{ii} = 0$. Then,

$$R_i = \sum_{j=1}^{N} V_{ji} \cdot \text{Bin}\left(y_j - K_j, \frac{n_i}{\sum_k V_{jk} n_k}\right) .$$

As a consequence, the rewired expected disease rate in area i is given by

$$\begin{aligned} \mathbb{E}\left(\frac{y_i^*}{n_i}\right) &= \pi\theta_i + \frac{1}{n_i} \sum_{j=1}^{N} V_{ji} \mathbb{E}\left((y_j - K_j) \frac{n_j}{\sum_k V_{ik} n_k}\right) \\ &\approx \pi\theta_i + \frac{1-\pi}{\nu} \nu \bar{\theta}_i \approx \theta_i \end{aligned}$$

where $\bar{\theta}_i$ is the average of θ_j values over the neighbors of area i and ν is the average number of spatial neighbors of a given area in the map. For example, $\nu = 5.6$ for the USA continental counties.

Hence, the expected rewired rate in each area is the same as in the original unknown mechanism that generates the observed data. As in the bootstrapped maps, the rewired maps will retain approximately whatever spatial pattern is present in the original map. This technique simulates scenarios where an incorrect assignment of cases to regions is possible, so one can test the resilience of the algorithms under this circumstance.

The choice of the π parameter is important in both, the Rewiring and Bootstrap methods. With π equal to 1, all generated instances will be exactly the same as the original data. On the other hand, the smaller the value of π, the greater the randomness added and, therefore, the more distant from the spatial pattern of the original data the perturbed instances will be. Ideally, the value of π should be close to but not equal to 1. Section 5.1 shows how to visually find a good value for the parameter π for real datasets.

4.2 New Performance Measures: Stability Metrics

In this section, we introduce our new metrics to evaluate the stability of **SDAD** algorithms. In non-spatial settings, an increasingly popular method to select the number of clusters is based on the stability idea [46]. The main idea is that a clustering algorithm should obtain similar results if applied to several datasets from the same data-generating distribution. While cluster detection algorithms are used to identify the subset of areas that are most likely to have a high risk, stability measures are used to indicate whether the identification of subsets is robust. We adapt this general philosophy to the spatial disease cluster problem. None of the stability algorithms proposed for non-spatial settings can be applied due to the spatial nature and the type of data, so new approaches need to be developed.

By the stability principle, a **SDAD** algorithm is stable if it returns similar results over the K instances generated under the same probabilistic model. In Section 5, we show that our stability metrics can be applied in two distinct situations. One considers the stability of a single cluster detected in a specific application, with no knowledge about the data-generating distribution. In this case, we use the instances produced by the bootstrap or rewiring methods described previously. The other is concerned with the typical performance evaluation of several **SDAD** algorithms, which resort to synthetic data instances generated by a known and user-determined probability distribution such as that described in Section 3. We distinguish these two types of use of our stability metric in Section 5.

For the i-th area in the map, we calculate the proportion $\hat{p}_i$ of times among the K instances it was detected as part of the spatial anomaly by the considered **SDAD** algorithm: $\hat{p}_i = \sum_k \mathbb{I}(i \in \hat{C}_k)/K$ where $\mathbb{I}(A)$ is the binary indicator for the event A occurrence. This notation is adopted to clearly indicate that $\hat{p}_i$ is an unbiased empirical estimator of the true but unknown associated probability p_i. We define the set of areas D that are detected in more than half of the instances by the given **SDAD** algorithm: $D = \{i \in \{1, \ldots, N\} : \hat{p}_i \geq 0.5\}$. These are the areas that are systematically detected when slightly perturbed data are input. Let $\bar{D}$ represent all the other areas with $0 < \alpha \leq \hat{p}_i < 0.5$ where α is a hyper-parameter. These are areas that have been occasionally detected.

We introduce two sets of metrics. The first set measures the stability of the retrieved spatial anomaly $\hat{C}_k$ across the K instances. This metric focuses on whether the algorithm consistently identifies the same set of areas $\hat{C}_k$, regardless of whether these areas truly constitute an anomaly. It evaluates the stability of the identified areas across the K instances, prioritizing consistency over algorithm precision.

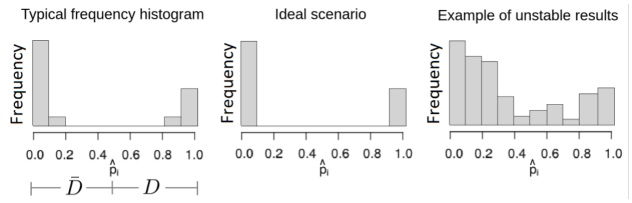


Figure 3: A typical satisfactory frequency histogram for $\hat{p}_i$ of all areas in the map, an ideal frequency histogram for $\hat{p}_i$, and an example of the values of an unstable algorithm.

For the first metric, we define a measure for the systematically detected areas in the set D

$$S_{\text{id}} = \frac{\sum_{i=1}^{N} [\mathbb{I}(i \in D)(\hat{p}_i - 0.5)/0.5]}{\sum_{i=1}^{N} \mathbb{I}(i \in D)} .$$

Similarly, we define a stability measure for the areas in $\bar{D}$, measuring the algorithm capacity of keeping the occasionally detected areas out of $\hat{C}_k$:

$$S_{\text{not-id}} = \frac{\sum_{i=1}^{N} [\mathbb{I}(i \in \bar{D})(1 - p_i/0.5)]}{\sum_{i=1}^{N} \mathbb{I}(i \in \bar{D})} .$$

Ideally, we want the **SDAD** algorithm to produce $S_{\text{id}} = 1$ and $S_{\text{not-id}} = 1$, showing that the method has detected one single set of areas in all map instances. To wrap the two measures into a single index, we take the F_1-score, calculating their harmonic mean: $S = 2(S_{\text{id}} * S_{\text{not-id}})/(S_{\text{id}} + S_{\text{not-id}})$.

Figure 3 gives an intuitive overview of the metrics. It shows an example of a typical frequency histogram for the $\hat{p}_i$ values obtained from all areas in the map. In the ideal situation, we wish that all $\hat{p}_i$ values for the areas in D are equal to or near 1, showing that the same set of areas was consistently detected in the map instances. For the set $\bar{D}$, we wish that all $\hat{p}_i$ values are near zero, meaning that the not systematically detected areas were returned almost in no instance. The worst scenarios are the ones where the $\hat{p}_i$ values fluctuate around 0.5. In these scenarios, the algorithm detected very distinct clusters along the simulations.

In contrast with the first set of metrics, the second set of metrics can be used only with synthetically generated datasets, when we know that there is a real spatial anomaly C in the map and which are the areas that comprise it. These metrics evaluate if the **SDAD** algorithm is able to consistently identify the ground truth spatial anomaly C and to avoid detecting the areas outside this true anomaly. That is, they measure the algorithm's ability to stably find the true anomaly even under data perturbation. Ideally, all the areas inside the true anomaly should be detected in all instances (resulting in p_i equals to 1) and the remaining areas should not be detected in any simulation, giving us p_i equals to zero. The stability for the positive areas is an average S_+ over the instances while, for the negative areas (those outside the true cluster), it is S_- and measured with the signed distance between p_i and 0. Also, we combine S_+ and S_- in an harmonic mean: $S_{\text{sign}} = 2(S_+ * S_-)/(S_+ + S_-)$. That is,

$$S_+ = \frac{\sum_{i=1}^{N} [\mathbb{I}(i \in C) p_i]}{\sum_{i=1}^{N} \mathbb{I}(i \in C)}$$

and

$$S_- = \frac{\sum_{i=1}^{N} [\mathbb{I}(i \notin C)\mathbb{I}(\hat{p}_i >= \alpha)(1 - p_i)]}{\sum_{i=1}^{N} \mathbb{I}(i \notin C)\mathbb{I}(\hat{p}_i >= \alpha)} .$$

5 RESULTS

5.1 Using the New Metrics in Real Datasets

We illustrate how our stability metrics S_{id}, $S_{\text{not-id}}$, and S based on the bootstrap or rewiring samples are useful in the analysis of real data. That is, a user selects a **SDAD** algorithm, applies it to a real disease map, and a set $\hat{C}$ of areas is detected. We do not know if there is a true high-risk spatial anomaly C present in the map or which areas it comprises. For each area $i \in \hat{C}$ eventually detected, we can obtain a stability measure associated with its membership in bootstrap or rewiring-generated instances. Also, we can compare different **SDAD** algorithms in that specific data analysis based on the stability metrics. The rationale is that, if there is no prior information about the anomaly shape, the stability measures help to identify the most appropriate **SDAD** algorithm to be used *in that specific application*. If an anomaly detected by an algorithm proves to be unstable when perturbing slightly the data, we have evidence that the algorithm is not able to identify the cluster confidently. This is so either because there is uncertainty about some areas having high rates or because the shape of the spatial anomaly is not compatible with the algorithm constraints.

We select two popular methods, the Circular Spatial Scan Statistics (**CSSS**) [22] and FleXScan [40], to illustrate how one can compare different algorithms using our measures. The first method constrains the spatial anomaly shape, scanning only over strictly circular-shaped candidates. A circular-shaped anomaly means that all areas' centroids are located within a certain circle. A circular-shaped anomaly is composed of all areas' centroids located within a certain circle. The second method detects spatial anomalies with any shape within a circular region up to a predefined maximum radius. We considered candidate anomalies with a population of up to 50% of the total population in the study region. For FleXScan, we considered all irregularly sized anomalies composed of up to 10 areas. We adopted the Poisson model to calculate the likelihood function. Although the results are presented considering these two methods, our approach is agnostic regarding the selected method. For example, we can use our stability-based methods with the approach proposed by [50], which was able to obtain statistical significance for spatial clusters detected by the DBSCAN algorithm.

The first step is to generate K perturbed instances of the original map using a sampling method (bootstrap or rewiring). One or more **SDAD** algorithms are run in each of these K pseudo-maps, returning a cluster $\hat{C}_k$, for $k = 1, \ldots, K$. For each algorithm and each i-th area in the map, we calculate the $\hat{p}_i$ and we compute the stability metrics S_{id}, $S_{\text{not-id}}$, and S for each method. The most stable **SDAD** algorithm for the given dataset is the one with higher S. Observing the $\hat{p}_i$ value for each area, we can also have an idea about how confident the method is for the result for that area.

To illustrate this use, we will impose a true spatial anomaly in one map so we can verify how the metrics perform. However, this ground truth knowledge will not be available in the practical application of spatial anomaly detection and it is not used to calculate

S_{id}, $S_{\text{not-id}}$ and S. Figure 4 shows an example of the cluster identified by the **CSSS** and by the FleXScan algorithms for an irregularly shaped cluster. The first row of plots corresponds to the circular scan and the bottom row, to the FlexScan. The first column shows what each of these methods detected as a spatial anomaly using the original data. The true anomaly is shown in red, while the returned areas are in blue. We used the bootstrap pseudo-maps with $\pi = 0.9$ to run this cross-validation. The areas in the second column of maps are colored according to their $\hat{p}_i$ values: the larger the $\hat{p}_i$ value, the darker the gray tone. The third column shows the histogram of the $\hat{p}_i$ values for all areas in the map. Calculating the stability of each method for this scenario with $\alpha = 0.05$, we have $S_{\text{id}} = 0.57$, $S_{\text{not-id}} = 0.59$ and $S = 0.58$ for the **CSSS** algorithm. For FleXScan, we have $S_{\text{id}} = 0.89$, $S_{\text{not-id}} = 0.84$ and $S = 0.84$. Thus, although more flexible and prone to overfitting, FleXScan is more stable and, therefore, preferable to the circular method in this particular dataset. We observe more $\hat{p}_i$ values near 0.5 for the **CSSS** method and, as a consequence, fewer areas with values with $\hat{p}_i$ near 0 (detected in no instance) or 1 (detected in all instances). We have evidence to believe that the anomaly is composed only of the areas detected by the FleXScan method. Indeed, the true anomaly has an irregular shape and can not be fitted into a perfectly circular shape. When observing the number of simulations for each area using both methods, we can see the center of mass at the actual location of the real anomaly, in the southeast of the map.

We are able to assess how stable the algorithm is in each area by looking at its $\hat{p}_i$ value. Even for the areas i belonging to the true anomaly C, the value of $\hat{p}_i$ can be very different due to the evidence amount present in the data. Remember that small population areas have highly unstable rates. Some areas, even if returned as part of the detected anomaly, have much less evidence, being returned only in nearly 50% of the perturbed instances. In the same way, areas that were not part of the detected anomaly can be returned in a few but significant number of instances. In a real scenario, this analysis shows which areas are the most critical ones in the anomaly, where the source of the disease problem may be located and rapid actions are needed.

In Figure 5 we can see $\hat{p}_i$ for each area varying with different values of π for our two proposed sampling methods. With the original data ($\pi = 1$), we have no variation between the instances, and the algorithms always detect the same anomaly. If we add a small amount of noise, we can note a dramatic drop in the $\hat{p}_i$ values for some areas for the CSSS algorithm, indicating that it has less confidence about such areas. Indeed, the majority of these areas are not part of the true anomaly, represented as black lines on the graph. On the other hand, if we set a very high value for π, we note that the confidence decreases for positive areas and increases for negative areas as a result of too much noise. This type of graph can help in choosing a value for the π parameter. For example, in the graphs in Figure 5, values around $\pi = 0.9$ provide a good balance, as they maintain high confidence in true anomaly areas while still accounting for some noise, ensuring the algorithms' robustness across varying conditions.

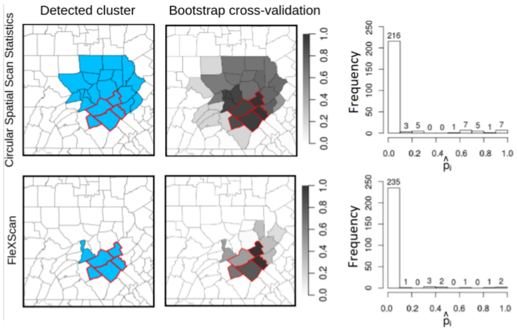


Figure 4: The first column shows the method-detected spatial anomaly using the original data. The 2nd and 3rd columns show the percentage of simulations for each area on the map and in a histogram based on bootstrap noise cross-validation with $\pi = 0.9$.

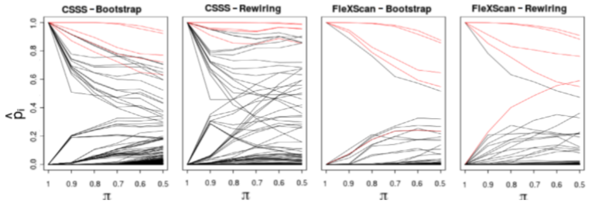


Figure 5: $\hat{p}_i$ for each area shown in figure 4 varying with π. Red lines correspond to the areas of the true cluster, while black lines correspond to the other areas on the map.

5.2 Selecting a SDAD algorithm

In this section, we show how our stability measures provide useful information to guide the **SDAD** algorithm general choice. In contrast with the previous section, we are not concerned with one specific application but rather with general algorithm properties. We specify a distribution probability to generate K synthetic datasets with a known spatial anomaly. In this way, besides S_{id}, $S_{\text{not-id}}$ and S, we can also calculate the metrics S_+, S_- and S_{sign}.

We know that areas with different underlying populations tend to exhibit varying levels of evidence in the data [8]. For this reason, we demonstrate our metrics on four anomalies, each composed of areas with different population profiles. The spatial anomalies were generated using the "Northeastern USA Benchmark Data, Purely Spatial" dataset collection, available at http://www.satscan.org/ [23]. It uses the geography and 1990 female population from 245 Northeastern U.S. counties. We created four spatial anomalies, each composed of 5 areas. The relative risks are defined so that the null hypothesis is rejected with probability 0.999 [26]. The format and areas that make up each anomaly can be seen by the areas with red borders in Figure 6. Each instance had a total of 600 disease cases.

The spatial anomaly or cluster A has a circular shape and its central area has a small population size. As shown in [8], areas with small population size have a high probability of receiving zero cases, even under a higher relative risk. By using **CSSS**, this central area will be forced into the detected circular cluster by the other high-risk and large population areas that form a ring around it. A method such as **CSSS** that restricts the potential anomalies to a circular shape tends to favor the inner-ring area. This is so even when $y_i = 0$ and therefore with no evidence of having a high underlying risk. The spatial anomaly B also has a circular shape but with a sparsely populated area on its edge, projecting out of the circular nucleus. A circular detection method tends to not identify this area, pointing to a spatial anomaly with a smaller radius unless the observed data presents strong evidence for high risk. This may also affect a flexible shape method such as FleXScan. Anomalies C and D have areas with diverse population sizes and do not have a circular shape. The smallest circle containing the spatial anomaly C is composed of 7 areas and hence incorporates 2 other areas with low disease incidence rates. The smallest circle containing the anomaly D is composed of 8 areas. Anomalies C and D should be more easily identified by irregularly shaped methods, but they can also be detected by the circular scan by means of a large circle.

For each spatial anomaly, we randomly generated 1000 instances by simulating disease cases in each area. We obtained $\hat{p}_i$ using $\alpha = 0.05$. Figure 6 shows on the first and second rows the $\hat{p}_i$ values for each area for CSSS and FleXScan, respectively, and the third row shows the stability metrics for each algorithm. There were no striking differences between the two methods. As expected, the CSSS algorithm performed better in all stability metrics for the cluster A, identifying exactly the true circular anomaly in almost all simulations. The FleXScan algorithm detected in fewer instances the areas in A with a small population and also identified more often some areas that do not belong to A. CSSS algorithm also performed better for the spatial anomaly B, assigning higher $\hat{p}_i$ to B areas, although FleXSCan was better at not detecting areas outside the anomaly. Note that both algorithms failed to detect the area with a small population in B. In anomaly C, FleXScan was more stable. CSSS detected many times a larger circular region containing the true spatial anomaly. The other areas were detected only in some simulations, giving to this algorithm a higher $S_{\text{not-id}}$ value. FleXScan found the same set of areas in almost all simulations with large values for S. However, it failed in detecting one area of C, leading to lower S_{+}. Due to the high occurrence of false positive areas detected by CSSS, FlexScan ended anyway with a better S_{sign} metric. For the anomaly D, CSSS and FleXScan had similar values for S. While CSSS detected consistently a larger spatial anomaly than the true one, FleXScan was better in giving lower $\hat{p}_i$ to areas not detected often. However, by analyzing the true spatial anomaly, we can note that FlexScan was better at finding systematically the true areas. In short, without prior information, CSSS would be recommended (because more stable) for spatial anomalies A and B, FleXScan for the anomaly C, and both algorithms for anomaly D.

5.3 Behavior of the metrics by varying π

We studied the stability for the detected and non-detected areas by the CSSS and FleXScan methods in the proposed clusters using the two noise addition procedures (bootstrap and rewiring). In all scenarios, the FleXScan method proved to be more stable for unidentified areas. That is, for areas not identified as part of the cluster, the number of simulations in which they are detected tends to be close to zero. As expected for cluster A, the CSSS method was more stable for the identified areas. This means that some areas were incorporated into the cluster in almost all noisy simulations. In fact, for cluster A, the circular method consistently identified the same cluster. For clusters B and C, the stability for the identified areas was similar for the two detection methods, with the circular method being slightly superior.

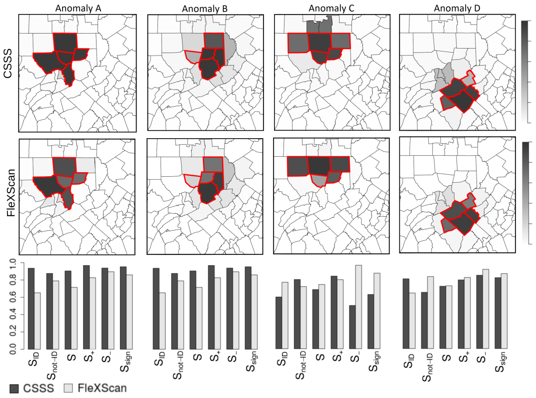


Figure 6: Percentage of simulations $\hat{p}_i$ for each area obtained by the CSSS (first row) and FlexScan (second row) algorithms for the spatial anomalies (areas with red borders). The third row shows the stability metrics of both algorithms.

We also studied the combination of stability for the identified areas and for the unidentified areas, which gives us the overall stability of the methods. Their stability is similar to each other, but there is a slight advantage for the circular method in cluster A and a slightly better result for the flexible method in cluster C. Additionally, we studied the stability for the truly positive and truly negative areas. Once again, we found that the FleXScan method is better when compared with CSSS. The reason is that it does not include truly negative areas in the detected cluster. The CSSS method, in turn, is more consistent for clusters A and B, correctly detecting the truly positive areas throughout the noisy simulations. For cluster C, the FleXScan method correctly identifies the cluster for low noise rates, but it is surpassed by the CSSS method as the noise increases. In summary, the stability of correctness for the circular method is superior for clusters A and B. For cluster C, the performance of the FleXScan method is better.

In our simulations, the results of the binomial and rewiring methods for adding noise were similar. Therefore, we do not favor either method. A more detailed analysis of the metrics can be made by examining the average proportion of times that each area was identified as part of the cluster detected during the noise simulations. By checking the distance from the average number of simulations in each area to zero or to the total number of simulations, we obtain evidence of the method's stability.

6 CONCLUSION

In this study, we extended the concept of stability, traditionally used to determine the optimal number of clusters in non-spatial datasets, to spatial disease anomaly detection (SDAD) algorithms.

We introduced innovative stability metrics and noise addition methods to assess and enhance the reliability of SDAD algorithms. Our approach, including bootstrap and rewiring sampling methods, ensures that spatial patterns in the data are preserved, providing a robust framework for evaluating SDAD performance.

Our results demonstrate that stability measures can significantly aid in selecting appropriate SDAD algorithms. For example, the CSSS algorithm showed better stability for spatial anomalies with circular shapes, while FleXScan performed better for irregularly shaped anomalies. This insight is crucial for practitioners to choose the most suitable algorithm based on the anticipated spatial anomaly shape.

However, our study has limitations. The proposed stability metrics are highly dependent on the hyper-parameter π, and selecting an inappropriate value can lead to misleading stability assessments. Future work could focus on developing adaptive methods for selecting π based on data characteristics.

Another limitation is our assumption that spatial patterns in disease data are consistent across different instances. In reality, external factors such as changes in population behavior, environmental conditions, or healthcare interventions can introduce variability in spatial patterns. Incorporating models that account for such influences could enhance the robustness of our stability measures.

Our evaluation relied on synthetic datasets, which, while allowing for controlled experimentation, may not fully capture the complexity of real-world disease data. Ground truth data for our task are unavailable. Although the COVID-19 pandemic provided extensive data, it does not offer suitable ground truth for our study due to uneven case recording, dynamic virus prevalence, and the complexities introduced by its infectious nature. COVID-19's rapidly changing patterns differ from the stable, slow-spreading diseases our work focuses on, and the need to model stochastic dependence due to interpersonal interactions complicates the use of standard disease mapping models.

In our study, we aimed to develop a robust framework for detecting spatial anomalies in disease risk, specifically under sparse data conditions. While synthetic datasets lack explicit ground truth, they allowed us to evaluate the feasibility and performance of our methods within the constraints of our defined scope. However, the importance of ground truth data cannot be overstated, and future research should aim to validate our findings using extensive real-world datasets to ensure the generalizability of our methods.

Another limitation is our focus on single spatial anomalies. Multiple overlapping or adjacent anomalies may occur in practice, complicating detection and stability assessment. Extending our stability metrics to handle multiple anomalies could provide a more comprehensive evaluation framework.

Lastly, while our stability metrics offer valuable insights into SDAD algorithms' reliability, they do not directly address the interpretability of the detected anomalies. For public health officials, understanding the underlying reasons for an anomaly is as important as detecting it. Integrating interpretability measures into our stability framework could make the results more actionable for decision-makers.

Our stability approach provides valuable insights into the given data and the spatial anomaly depicted in the map. It operates independently of the chosen SDAD method, the map topology, and the data-generating distribution. Moreover, it does not rely on any prior information or assumptions about the dataset. This work advocates for cluster stability as an essential checkpoint in interpreting spatial disease anomalies and introduces a systematic and standardized analytical framework for assessing spatial anomaly detection results.

REFERENCES

[1] Beth Abrams, Henry Anderson, Carina Blackmore, Frank J Bove, Suzanne K Condon, Christie R Eheman, Jerald Fagliano, Lorena Barck Haynes, Lauren S Lewis, Jennifer Major, et al. 2013. Investigating suspected cancer clusters and responding to community concerns: guidelines from CDC and the Council of State and Territorial Epidemiologists. *Morbidity and Mortality Weekly Report: Recommendations and Reports* 62, 8 (2013), 1–24.

[2] Guilherme Borges, Flavio Figueiredo, Renato M Assunçao, and Pedro OS Vaz-de Melo. 2020. Networked Point Process Models Under the Lens of Scrutiny. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD).*

[3] P Boyle, AM Walker, and FE Alexander. 1996. Methods for investigating localized clustering of disease. Historical aspects of leukaemia clusters. *IARC scientific publications* 135 (1996), 1–20.

[4] Jose Cadena, Arinjoy Basak, Anil Vullikanti, and Xinwei Deng. 2018. Graph scan statistics with uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[5] Ken Camacho, Raechel Portelli, Ashton Shortridge, and Bruno Takahashi. 2021. Sentiment mapping: point pattern analysis of sentiment classified Twitter data. *Cartography and Geographic Information Science* 48, 3 (2021), 241–257.

[6] Jack Cordes and Marcia C Castro. 2020. Spatial analysis of COVID-19 clusters and contextual factors in New York City. *Spatial and Spatio-temporal Epidemiology* 34 (2020), 100355.

[7] Emiliano del Gobbo, Lara Fontanella, Sara Fontanella, and Annalina Sarra. 2021. Geographies of Twitter debates. *Journal of Computational Social Science* (2021), 1–17.

[8] Raphaella Carvalho Diniz, Pedro OS Vaz-de Melo, and Renato Assunção. 2020. Evaluating the Evaluation Metrics for Spatial Disease Cluster Detection Algorithms. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems.* 401–404.

[9] Luiz Duczmal and Renato Assuncao. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* 45, 2 (2004), 269–286.

[10] Bilal Shikur Endris, Geert-Jan Dinant, Seifu H Gebreyesus, and Mark Spigt. 2021. Geospatial inequality of anaemia among children in Ethiopia. *Geospatial Health* 16, 2 (2021).

[11] Gabriel J Escobar, Alyce S Adams, Vincent X Liu, Lauren Soltesz, Yi-Fen Irene Chen, Stephen M Parodi, G Thomas Ray, Laura C Myers, Charulata M Ramaprasad, Richard Dlott, et al. 2021. Racial disparities in COVID-19 testing and outcomes: retrospective cohort study in an integrated health system. *Annals of internal medicine* 174, 6 (2021), 786–793.

[12] Jessica Zosa Forde and Michela Paganini. 2019. The Scientific Method in the Science of Machine Learning. In *ICLR Debugging Machine Learning Models Workshop.* 1–9. arXiv:1904.10922

[13] Yizhao Gao, Ting Li, Shaowen Wang, Myeong-Hun Jeong, and Kiumars Soltani. 2018. A multidimensional spatial scan statistics approach to movement pattern comparison. *International Journal of Geographical Information Science* 32, 7 (2018), 1304–1325.

[14] Michael Goodman, Joshua S Naiman, Dina Goodman, and Judy S LaKind. 2012. Cancer clusters in the U.S.A.: what do the last twenty years of state and federal investigations tell us? *Critical reviews in toxicology* 42, 6 (2012), 474–490.

[15] Michael Greenberg and Daniel Wartenberg. 1991. Communication to an alarmed community about cancer clusters: a fifty state survey. *Journal of Community Health* 16, 2 (1991), 71–82.

[16] E R Hargreaves. 1961. Epidemiological Studies in Cornwall. *Proceedings of the Royal Society of Medicine* 54, 3 (1961), 209–216.

[17] Julian Ijumulana, Fanuel Ligate, Prosun Bhattacharya, Felix Mtalo, and Chaosheng Zhang. 2020. Spatial analysis and GIS mapping of regional hotspots and potential health risk of fluoride concentrations in groundwater of northern Tanzania. *Science of The Total Environment* 735 (2020), 139584.

[18] Ariful Islam, Md Abu Sayeed, Md Kaisar Rahman, Jinnat Ferdous, Shariful Islam, and Mohammad Mahmudul Hassan. 2021. Geospatial dynamics of COVID-19 clusters and hotspots in Bangladesh. *Transboundary and Emerging Diseases* 68, 6 (2021), 3643–3657.

[19] MA Jie, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, Ben van Calster, et al. 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology* (2019).

[20] Gerhard Klassen, Martha Tatusch, Ludmila Himmelspach, and Stefan Conrad. 2020. Fuzzy Clustering Stability Evaluation of Time Series. 680–692.

[21] G. Knox. 1989. Detection of clusters. In *Methodology of Enquiries into Disease Clustering*, P. Elliott (Ed.). Small Area Health Statistics Unit, London, 17–22.

[22] Martin Kulldorff. 1997. A spatial scan statistic. *Communications in Statistics-Theory and methods* 26, 6 (1997), 1481–1496.

[23] Martin Kulldorff, Eric J Feuer, Barry A Miller, and Laurence S Freedma. 1997. Breast cancer clusters in the northeast United States: a geographic analysis. *American journal of epidemiology* 146, 2 (1997), 161–170.

[24] Martin Kulldorff, Lan Huang, Linda Pickle, and Luiz Duczmal. 2006. An elliptic spatial scan statistic. *Statistics in medicine* 25, 22 (2006), 3929–3943.

[25] Martin Kulldorff and Neville Nagarwalla. 1995. Spatial disease clusters: detection and inference. *Statistics in medicine* 14, 8 (1995), 799–810.

[26] Martin Kulldorff, Toshiro Tango, and Peter J Park. 2003. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis* 42, 4 (2003), 665–684.

[27] Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann. 2004. Stability-Based Validation of Clustering Solutions. *Neural Computation* 16, 6 (jun 2004), 1299–1323.

[28] Zachary C Lipton and Jacob Steinhardt. 2019. Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue* 17, 1 (2019), 45–77.

[29] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. 2018. Are GANs created equal? a large-scale study. In *Advances in neural information processing systems*. 700–709.

[30] Rossella Melchiotti, Filipe Gracio, Shahram Kordasti, Alan K. Todd, and Emanuele de Rinaldis. 2017. Cluster stability in the analysis of mass cytometry data. *Cytometry Part A* 91, 1 (jan 2017), 73–84.

[31] Gábor Melis, Chris Dyer, and Phil Blunsom. 2018. On the State of the Art of Evaluation in Neural Language Models. In *ICLR*.

[32] Rafael da Silveira Moreira. 2020. COVID-19: intensive care units, mechanical ventilators, and latent mortality profiles associated with case-fatality in Brazil. *Cadernos de saude publica* 36 (2020).

[33] Suvadip Mukherjee, Thibault Lagache, and Jean-Christophe Olivo-Marin. 2021. Evaluating the Stability of Spatial Keypoints via Cluster Core Correspondence Index. *IEEE Transactions on Image Processing* 30 (2021), 386–401.

[34] Daniel Neill and Andrew Moore. 2004. Rapid Detection of Significant Spatial Clusters. *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2004), 256–65.

[35] Thanh Tam Nguyen. 2019. *Graph-based rumour detection for social media*. Technical Report.

[36] Carlos Riquelme, George Tucker, and Jasper Snoek. 2018. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127* (2018).

[37] Mirco Sandfort, Amélie Vantaux, Saorin Kim, Thomas Obadia, Anaïs Pepey, Soazic Gardais, Nimol Khim, Dysoley Lek, Michael White, Leanne J Robinson, et al. 2020. Forest malaria in Cambodia: the occupational and spatial clustering of Plasmodium vivax and Plasmodium falciparum infection risk in a cross-sectional survey in Mondulkiri province, Cambodia. *Malaria journal* 19, 1 (2020), 1–12.

[38] David Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's curse? On pace, progress, and empirical rigor. In *ICLR 2018 Workshop, International Conference on Learning Representations*.

[39] Benjamin Strang, Peter van der Putten, Jan N van Rijn, and Frank Hutter. 2018. Don't Rule Out Simple Models Prematurely: A Large Scale Benchmark Comparing Linear and Non-linear Classifiers in OpenML. In *International Symposium on Intelligent Data Analysis*. 303–315.

[40] Toshiro Tango and Kunihiko Takahashi. 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics* 4, 1 (2005), 11.

[41] Leonardo Vilela Teixeira, Renato Martins Assuncao, and Rosangela Helena Loschi. 2015. A generative spatial clustering model for random data through spanning trees. In *2015 IEEE International Conference on Data Mining*. IEEE, 997–1002.

[42] Leonardo V Teixeira, Renato M Assunção, and Rosangela H Loschi. 2019. Bayesian space-time partitioning by sampling and pruning spanning trees. *Journal of Machine Learning Research* 20, 85 (2019), 1–35.

[43] Michael J Thun and Thomas Sinks. 2004. Understanding cancer clusters. *CA: A Cancer Journal for Clinicians* 54, 5 (2004), 273–280.

[44] Craig W Trumbo. 2000. Public requests for cancer cluster investigations: a survey of state health departments. *American Journal of Public Health* 90, 8 (2000), 1300.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[46] Ulrike Von Luxburg. 2010. *Clustering stability: an overview*. Now Publishers Inc.

[47] Bei Wang, Jeff M Phillips, Robert Schreiber, Dennis Wilkinson, Nina Mishra, and Robert Tarjan. 2008. Spatial scan statistics for graph clustering. In *Proceedings of the 2008 SIAM international conference on data mining*. 727–738.

[48] David H Wolpert. 1996. The lack of a priori distinctions between learning algorithms. *Neural computation* 8, 7 (1996), 1341–1390.

[49] Eileen E Wood. 1960. A survey of leukaemia in Cornwall, 1948-1959. *British Medical Journal* 1, 5188 (1960), 1760.

[50] Yiqun Xie and Shashi Shekhar. 2019. Significant DBSCAN towards statistically robust clustering. In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*. 31–40.

[51] Andreas Züfle, Tobias Emrich, Klaus Arthur Schmid, Nikos Mamoulis, Arthur Zimek, and Matthias Renz. 2014. Representative clustering of uncertain data. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 243–252.