Statistically-Robust Clustering Techniques for Mapping Spatial Hotspots: A Survey

YIQUN XIE, University of Maryland
SHASHI SHEKHAR and YAN LI, University of Minnesota

Mapping of spatial hotspots, i.e., regions with significantly higher rates of generating cases of certain events (e.g., disease or crime cases), is an important task in diverse societal domains, including public health, public safety, transportation, agriculture, environmental science, and so on. Clustering techniques required by these domains differ from traditional clustering methods due to the high economic and social costs of spurious results (e.g., false alarms of crime clusters). As a result, statistical rigor is needed explicitly to control the rate of spurious detections. To address this challenge, techniques for statistically-robust clustering (e.g., scan statistics) have been extensively studied by the data mining and statistics communities. In this survey, we present an up-to-date and detailed review of the models and algorithms developed by this field. We first present a general taxonomy for statistically-robust clustering, covering key steps of data and statistical modeling, region enumeration and maximization, and significance testing. We further discuss different paradigms and methods within each of the key steps. Finally, we highlight research gaps and potential future directions, which may serve as a stepping stone in generating new ideas and thoughts in this growing field and beyond.

CCS Concepts: • **Information systems → Spatial-temporal systems**; **Clustering**; • **Computing methodologies → Cluster analysis;**

Additional Key Words and Phrases: Hotspot, mapping, clustering, statistical rigor, scan statistics

ACM Reference format:
Yiqun Xie, Shashi Shekhar, and Yan Li. 2022. Statistically-Robust Clustering Techniques for Mapping Spatial Hotspots: A Survey. *ACM Comput. Surv.* 55, 2, Article 36 (January 2022), 38 pages.
https://doi.org/10.1145/3487893

1 INTRODUCTION

Identification of spatial hotspots, i.e., geographic regions with significantly higher rates of generating cases for certain types of events (e.g., disease, crime, and pollution), has become an important task and standard research methodology in a diverse range of applications (e.g., public health, public safety, transportation, and agriculture). In public health, for example, epidemiologists use

This work is supported in part by the NSF under Grants No. 2105133, 2126474, 2126449, 1901099, 1737633, and 2040459, the USDOD under Grants HM0210-13-1-0005, ARPA-E under Grant No. DE-AR0000795, USDA under Grant No. 2017-51181-27222, NIH under Grant No. UL1 TR002494, KL2 TR002492 and TL1 TR002493, Google AI for Social Good, and the University of Maryland.
Authors' addresses: Y. Xie, University of Maryland, Center for Geospatial Information Science, 7251 Preinkert Dr., College Park, MD 20742; email: xie@umd.edu; S. Shekhar and Y. Li, University of Minnesota, Department of Computer Science, 200 Union Street SE, Minneapolis, MN 55455; emails: {shekhar, lixx4266}@umn.edu.

0360-0300/2022/01-ART36 $15.00
https://doi.org/10.1145/3487893

spatial hotspots to find outbreaks of diseases and make timely interventions to reduce health risks to the public [78]. Similarly, in public safety, hotspots reveal areas with abnormally high-rates of crimes (e.g., theft, assault), which can help narrow the search space for serial criminals (e.g., arsonists) or other underlying factors [130]. Broad applications of hotspot mapping will be detailed in Section 2.

While spatial hotspots are naturally clusters of events, mapping of hotspots poses many unique challenges. Here, we summarize these challenges in four different aspects. *First,* spurious results often have very high social and economic costs in the application domains of hotspot mapping [16, 182]. For example, if a city falsely claims a neighborhood as a crime cluster, there can be many unnecessary negative impacts such as reducing property values in that region, hurting local small businesses, increasing mental pressure on residents, and potentially pushing them to move. Unfortunately, spurious clusters are very common in real-world datasets and can easily happen as a consequence of natural randomness. Thus, decision makers in important application domains require statistical rigor as a necessary component to robustly control of the rate of spurious patterns. *Second,* needs explicit consideration of underlying risk or known contributing factors, which are typically not considered in traditional clustering approaches [102, 133]. For example, having a hundred crime cases in a densely-populated urban area is different from having the same in a rural area of the same size. *Third,* hotspots are geographically contiguous regions and should not be fragmented pieces in space [191, 194]. As a result, spatial attributes cannot be simply mixed with other non-spatial attributes during the clustering process and need specific modeling. *Finally,* while a hotspot is often a geographically contiguous region, the observed events inside it may not necessarily form a continuous or smooth distribution of density, especially when the total number of observations is small (i.e., higher variance across locations). This undermines basic assumptions on contiguous density used in many traditional clustering approaches [29, 66].

Clustering has long been a core topic in data mining and machine learning, providing a vast set of techniques in various families. Partitioning-based clustering methods (e.g., k-means and Gaussian mixture models, CLARANS [153]) split input data into k groups to minimize dissimilarities within clusters. Density-based approaches (e.g., DBSCAN [66], OPTICS [14], and HDBSCAN [29]) apply local or global density criteria to separate out clusters from noises. These approaches do not rely on pre-defined number of clusters and can be made flexible for different shapes and varying densities. Many techniques also utilize the hierarchical structure of clusters, e.g., by first splitting data into small units and then sequentially merging them into final clusters based on similarity and adjacency, or in the opposite direction (e.g., Chameleon [99], BIRCH [224], CURE [80], and HDBSCAN [29]). These families of approaches are not necessarily mutually exclusive (e.g., HDBSCAN leverages both density and hierarchy). Moving further, there have also been many other types such as grid-based (e.g., STING [205], CLIQUE [9]), graph-spectrum-based (e.g., spectral clustering [202]), kernel-based (e.g., kernel k-means [177], SVC [20]), and so on. Given that these techniques have been well summarized in other surveys on general clustering (e.g., [217, 218]), here, we skip the details and concentrate on the topic of statistically-robust clustering for hotspot mapping.

Traditional clustering techniques are not sufficient in responding to the unique challenges posed by hotspot mapping. These techniques, for example, often do not consider the high costs of spurious results and are prone to return many clusters formed by natural randomness (e.g., k-means always returns k partitions, DBSCAN and its variations have difficulty in avoiding spurious high-density regions in random point distributions [29, 133, 211, 213, 214]). In addition, these methods normally do not consider underlying risk factors (e.g., population at risk as control), geographic-contiguity of outputs, and non-contiguous within-cluster density.

To address these challenges, in the last decades there have been a blossom of research focusing on statistically-robust clustering techniques with explicit consideration and modeling of these various needs of hotspot mapping. As many of the methods are based on formulations from scan statistics, which was first developed by the statistics community, statistically-robust clustering is also known as scan statistics. Since both clustering and scan statistics have been widely used in literature, we use "clustering" in the rest of the survey to better connect to intended audience from data mining. *First*, statistical significance of cluster candidates was modeled and incorporated into the detection process to provide robust control of the rate of spurious results. New algorithms were also developed to reduce the computational cost of significance testing, which often requires Monte-Carlo-based estimation due to the complexity of test statistics. *Second*, underlying risk factors were explicitly modeled into the test statistics used for candidate evaluation, allowing the scores to reflect differences in statistical processes rather than directly observed densities of events. *Third*, statistically-robust clustering techniques can also guarantee that the output clusters are contiguous in the geographic space by modeling the problem as a region-maximization problem. In this paradigm, the enumeration space of region-candidates are clearly-defined (e.g., circles, rings, or linear paths formed by subsets of data) and a wide variety of efficient region-maximization algorithms have been developed. This enumeration scheme also addresses the fourth challenge that within-cluster density may not be smooth or continuous due to the effect of natural randomness. By directly enumerating regions instead of forming them by local or hierarchical density criteria, regions can be evaluated regardless of their inner-density distribution.

There are several related surveys on traditional and statistically-robust clustering. For traditional clustering, multiple taxonomies are provided in [22, 124, 160, 217, 218] to summarize a broad spectrum of techniques, including partition-based, density-based, hierarchy-based, graph spectrum based, kernel based, spatial and non-spatial, and so on. These surveys do not cover the models and techniques developed for statistically-robust clustering. A few surveys also exist on statistically robust clustering from different perspectives. Kulldorff [1999] [103] provides a summary of early methods and applications of spatial scan statistics, as well as basic enumeration algorithms; an overview of earlier domain applications was also discussed later in [46]. Neill and Moore [2006] [147] describe a few extensions of spatial scan statistic, including efficient computational strategies for rectangular-shaped clusters and an expectation-based approach to improve space-time cluster detection. Later, two sequential handbooks on scan statistics were developed mainly by the statistics community [73, 75], covering both new and existing theoretical results on distribution estimation (e.g., closed-form solutions, approximations), extreme values, sequences, and so on. The handbooks also included a few chapters that are related to data mining, including summaries on Bayesian scan statistics [139] and irregular-shaped cluster detection [59]. The topic of statistically-robust clustering (a.k.a. hotspot detection) has also been briefly discussed (e.g., as a paragraph or section) in broad spatio-temporal data mining and data science surveys [16, 182, 209]. It is worth-mentioning that hotspot is also related to another common pattern—spatial anomaly/outlier—in literature [10, 16, 35, 182, 209]. One key difference between the two patterns is that anomaly or outlier is often used to describe rare observations, e.g., as defined in existing surveys such as "interesting but rare phenomena" [16], "rare occurrences" [10], "rare patterns" [182], and so on. In contrast, significant clusters or hotspots often represent a substantial/large portion or sometimes the majority of observations.

Given the richness of application contexts and the large variety of techniques developed along different dimensions of statistically-robust clustering, there is a need for developing a taxonomy from the computer science or data mining perspective to decompose the complex modeling and detection process into a set of key steps, highlight their roles and mutual relationships, and discuss the advances in each key step by extracting corresponding contributions from the vast (and often versatile) literature. This can promote cross-pollination and synergistic integration of ideas across various research communities, especially considering that the developments of general clustering and statistically-robust clustering have largely run on siloed tracks. This can also help engage the broader data mining community in addressing the challenges and fostering future directions in statistically-robust clustering.

This survey aims at fulfilling this need as follows. First, we develop a set of taxonomies to highlight the key components of statistically-robust clustering and its sub-areas (e.g., data models, spatial point processes (SPP), test statistics, enumeration space, computational algorithms, and test paradigms). As research papers in the field often touch on multiple aspects of the overall taxonomy, we further decompose a broad set of the papers into contributions to each component of the taxonomy, which are otherwise difficult to distill and hinder cross-discipline dissemination of the research advances. This process also helps extract ideas that are applicable, and potentially valuable, to general studies in the field beyond their original scopes. Finally, based on the identified research landscape, we expose several open questions and future research directions to harness emerging sources of spatial big data and promote synergistic integration of ideas across multiple domains.

Scope and outline: This review article surveys the research progresses from a data mining perspective for computer science audience, and does not intend to provide a comprehensive survey on related statistics literature, which is another important line of research that often focuses more on statistical theories (e.g., distribution approximation) and has been well summarized in [73, 75]. The rest of the survey is outlined as follows: Section 2 introduces the application domains of hotspot mapping; Section 3 summarizes techniques for statistically-robust clustering, with Section 3.1 providing general key concepts, Section 3.2 providing an overview of building blocks, Section 3.3 providing a taxonomy of data models, Section 3.4 providing a taxonomy on various statistical processes and test statistics, Section 3.5 providing a taxonomy on different region definitions as well as their enumeration and maximization algorithms, Section 3.6 summarizing different paradigms and methods for significance testing, and Section 3.7 discussing evaluation methods and metrics. Finally, Sections 4 and 5 discuss challenges and opportunities for future research, and conclude the article with highlights.

2 APPLICATIONS OF SPATIAL HOTSPOT MAPPING

Statistically significant clustering for hotspot mapping has been widely used with an ever-growing variety of spatiotemporal data sources from different domains. In this section, we highlight example applications from several domains where statistically interpretable outputs and robustness against spurious results are especially important for both research investigation and policy making.

Public health: The wide adoption of hotspot mapping started from public health with the popularity of the spatial scan statistic [102]. The most typical use is to identify the outbreaks of diseases, including cancer (e.g., childhood leukemia [84], cervical cancer [41]), infectious diseases (e.g., malaria [129], dengue [78, 188], COVID-19 [44, 85, 109]), and many others [63, 78, 116]. With the awareness of both diseases cases and population at risk, detected hotspots of diseases reveal abnormal underlying factors that are otherwise hidden from public health researchers and policy makers. Many of these applications have been considered as standard practices in public health studies (e.g., National Cancer Institute [1]). In addition to case data, other sources of related disease information, including symptoms (e.g., diarrhoea and fever [106]) and increase in pharmacy visits [140], have been used to improve the timeliness of outbreak alarms. With the growing variety of spatiotemporal datasets, recent applications have also explored the incorporation of proxy data from social networks (e.g., geotagged tweets about dengue fever [187]) and trajectories.

Public Safety: Hotspot detection is popularly used in public safety applications to identify statistically alarming clusters of criminal activities, accidents, suicides, severe disaster impacts, and so on. In crime analysis, such significant clusters are used to reveal patterns of homicide and drug traffic [222], violence-related deaths [126], police deaths in the line of duty [98], and more, with respect to many underlying factors including natural disasters, weather, new construction, and so on. Geographic profiling of serial criminals, characterized as ring-shaped hotspots of serial crimes (e.g., arson), is also used to help narrow down potential residence locations of criminals [64]. Similarly, adoptions can be found in early warning of terrorism [71], minefield detection [201], and identification of significant concentration of accidents (e.g., falls, fire injuries, shark attacks) [13, 54, 206].

Urban Planning and Transportation: Statistically significant clustering helps urban policy makers and designers better monitor, manage and plan infrastructure changes. Typical uses include locating breakage of distribution pipes [51], mapping city areas with low accessibility to critical infrastructures or housing difficulties [163], identifying disparities in air quality [81], revealing regions with aging problems [219], and many others [33, 83]. Generalized hotspots can also reflect levels of diversity (e.g., trees in urban forests, types of business services) and have been used to detect non-resilient designs or distributions of infrastructures within cities [207]. The methods have also been extended to cover network-based applications in transportation. In transportation-safety related use cases, road segments—including both intersections and linear paths—have been used to identify subspaces in road networks that have significantly higher rates of traffic accidents (e.g., pedestrian fatalities) [156, 196]. With emerging data sources such as on-board-diagnostic data (i.e., vehicle trajectories with hundreds of engine measurements), hotspot detection is also used to find sub-paths along routes that cause divergent behaviors in emission or energy consumption [115].

Others: Forestry studies use hotspot mapping to analyze forest health at large scales [171] and detect clusters of forest fires [157], tree regeneration [68], and so on. It is also used in both plant and animal agriculture. In plant agriculture, for example, spatial concentrations of high and low ratios of production-to-environment-cost are used to suggest alternative plans for agricultural conservation planning [178]. Hotspots of crop diseases were also used to map disease spread across plants [48]. On the agricultural operation safety side, the approach is used to find spatial clusters of field accidents (e.g., tractor overturns [176]). In animal agriculture, hotspot mapping has been widely employed to identify outbreaks of diseases in stock herds [193]. Finally, a great variety of applications exist in many other fields, including astronomy [32], medical imaging [220], environment [101], climate change [216], fishery [192], entomology [19], and many more.

3 STATISTICALLY-ROBUST CLUSTERING: AN OVERVIEW AND TAXONOMIES

In this section, we first provide key concepts and a high-level taxonomy to illustrate the building blocks of statistically-robust clustering. Then, we will dive deeper into each building block and discuss the detailed taxonomy and summary of developments.

3.1 Key Concepts and a General Problem Statement

In the following, we summarize the key concepts and general problem definition for statistically-robust clustering. Here, we keep the definitions at a high-level given the variety of formulations when fine details are considered, which will be discussed in later sections.

Definition 3.1 (Case and Control). A **case** C_i is an observed incident that is related to a real-world event or phenomenon E of interest (e.g., the spread of a disease). The number of cases is the number of observed incidents (e.g., the number of people with positive test results for a disease). In contrast, **control** represents the base population of candidates for the same event, where each candidate B_j is subject to being a case of the event (e.g., all people being tested for a disease or at risk of being infected). Data on case and control can be presented at either individual (e.g., points) or aggregated levels (e.g., polygons, each of which contains the number of case or control points inside).

Definition 3.2 (Point Distribution and Point Process). A **point distribution** is a spatial distribution of cases related to an event E. A **point process** governs the generation of random point distributions given a control distribution. In hotspot detection, it—for example—can determine the probability of each control point B_j being a case C_i of E based on its location. If the spatial distribution of control is not discrete but continuous, the point process can then determine probability density across space. Here a "point" refers to a general spatial data object, which can be a real point (location), a trajectory, and so on.

Definition 3.3 (Hotspot). A hotspot is a sub-region of a given input space, where the rate of generating case points inside p_{in} (or the probability of each individual control point being a case point) is higher compared to its outside p_{out}. The existence of a hotspot means the point process is not homogeneous in the input space but clustered at the hotspot region.

Given case and control distributions as inputs, statistically-robust clustering aims at identifying clusters formed by true hotspots rather than by pure random chance. The output clusters can either be represented using sub-regions of the input space or subsets of case points. Different from traditional clustering, output clusters are not only regions with high-intensity of case points. In contrast, the intensity is statistically adjusted by an underlying control distribution and output clusters must pass significance testing defined by a point process.

3.2 A High-Level Taxonomy

Figure 1 shows a high-level taxonomy of the key steps in statistically-robust clustering for hotspot mapping. Since hotspots can be considered as one type of clusters, in the rest of the article the two names are used in an interchangeable manner.

At the beginning of the process is the input spatial data model (Figure 1). One of the key characteristics of data in hotspot mapping—case vs. control—is highlighted under the step. A variety of commonly used data types will be discussed and categorized in detail in Section 3.3.

The second key step is the statistical modeling needed for rigorous control of spurious results. The first essential element within the step is a spatial (or spatiotemporal) point process needed to model the generation of a point distribution. This is often determined by assumptions in domain applications (e.g., Poisson point processes are common choices for monitoring disease outbreaks). Then, a hypothesis is needed to provide a clear mathematical definition of a "hotspot" (e.g., inside probability density of generating events is higher than outside). This definition is important as it spells out the criterion of a true hotspot, providing a mathematical ground to generate ground-truth synthetic data during evaluation (Section 3.7). Finally, a test statistic is needed to compute a score of each candidate cluster, which is necessary to allow comparison and ranking among a large number of candidates.

The third key step focuses on finding "where" a hotspot is based on the statistical modeling. This is achieved by clearly defining an enumeration space (e.g., a parameter space containing all candidates) and searching over regions to maximize test statistic values. The definitions of enumeration spaces can be used to guarantee that all candidate regions are spatially contiguous and meaningful to decision makers. Additional constraints can be applied to the enumeration space to avoid undesired outputs.

The last step before a hotspot can be added to the output is significance testing. This is the final guard in this process to enforce robust control of the rate of spurious clusters. The testing can be

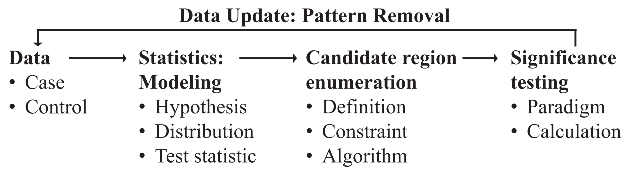


Fig. 1. Taxonomy based on the process of statistically-robust clustering.

performed using both the Frequentist and Bayesian views depending on the application need. The view should be consistent with the point process modeled in the second key step. Due to the typical need of Monte-Carlo-based estimation during the test, acceleration techniques are important to reduce the high cost leveraging unique characteristics of the simulation process.

Finally, in Figure 1 there is an arrow connecting back to the beginning of this process, which is data update. In this step, any detected significant hotspot is removed from the data before continuing to identify the other hotspots. This is often necessary due to the shadowing effects between hotspots in the data, which will be illustrated in Section 3.6.1.

3.3 Data Types and Modeling

Mapping of hotspots covers a wide variety of data types and there are many different criteria that can be used to classify these data types into different groups. Table 1 shows a detailed taxonomy of various ways of classification.

3.3.1 Types of Observation. A key characteristic of data used in statistically-robust clustering is the need of both case and control data, which is not the case in traditional clustering approaches (i.e., no control). In many critical societal applications (Section 2), the emergence of events are closely linked to other related distributions. In transportation, for example, a road segment with a higher traffic volume is also more likely to have a higher density of traffic accidents, but such "clusters" are not interesting to decision/policy makers as they are expected and often do not call for additional investigation and intervention [115, 196].

To have a more holistic view of this process, we can consider two levels of an observed dataset: (1) incidents of one or more events (e.g., crime cases); and (2) underlying contributing factors to the events (i.e., control). When control is ignored, the resulting clusters are representations of an aggregation of all the underlying contributing factors—which together make the cluster regions have higher rates of generating the observed data samples. Having control data as an input allows users to take out the contributions from known factors (e.g., traffic volume)—either interesting or not—and let resulting clusters reveal unknown factors causing higher rates of the events. More examples are provided in Table 1.

3.3.2 Types of Event Data Models. Statistically-robust clustering takes a wide variety of input data models, ranging from points, polygons, grids, trajectories to time-series data. Examples of data models are shown in Figure 2 as well as Table 1, and illustrated as follows:

— **Point:** Point is the most popular and widely used vector data model for mapping hotspots [63, 106, 126, 129, 222]. Each point has geographic coordinates, observation type (i.e., case, control), and event information (e.g., event type such as disease or crime, event subtype, event value such as air pollution indices if data is continuous).
— **Polygon and grid:** Polygons are typically used when observations are given at aggregated levels [30, 41, 78, 81, 84, 89, 104, 105, 219]. For example, population and demographic information are normally only available through census tracks, which may be shared at higher

Table 1. Taxonomy of Data Types

Classification criteria	Categories	Examples and use cases
Type of observation	Case	• Public health: Disease incidents (e.g., COVID), symptoms, tweets and pharmacy visits [85, 116, 149, 188] • Public safety: Crime cases, signals of terrorism and police deaths in the line of duty [64, 71, 98, 215, 222] • Transportation and urban infrastructure: Vehicle accidents, pedestrian fatalities, high emissions and underground pipe breakage [51, 115, 156, 184, 199]
	Control	• Public health: Population, population at risk, all tweets [102, 116, 129, 146, 187, 188] • Others: Traffic volume of Vehicles or pedestrians, forest volume and crop plant population [48, 68, 115, 157, 171]
Type of data models	Point	• Crime cases, injuries, fires, geo-tagged tweets, sampling sites [13, 64, 101, 157, 187, 188, 206, 215, 222] • Traffic accidents and vehicle tows [91, 156, 196, 199, 214]
	Polygon	• County or city district maps with statistics of events (e.g., disease), shorelines [13, 41, 85, 102, 116, 140] • Grid cells [90, 135, 145, 146, 150, 151, 216, 220]
	Trajectory	• Vehicle trajectories, smartphone or geo-tagged social media trajectories [15, 90, 115, 122, 187, 188]
	Time-series	• Hospital visits over the past days/weeks [94, 132, 135, 136, 140, 151]
Type of spaces	Euclidean	• A state, city or country [41, 85, 116] • A forest area [68, 157]
	Spatial network or graph	• Road or river network [91, 115, 156, 196, 199] • Utility network or graph representation of space [28, 51] • Social network with geo-tags [36, 187, 188]
Level of uncertainty	Individual	• Locations of crimes, traffic accidents or crop disease incidents (e.g., point-type examples above) • Perturbed or inaccurate locations and timestamps [90, 120, 155]
	Aggregated	• County maps with counts of cases and population, where uncertainty comes from both the aggregation and inaccuracies (e.g., [25, 52, 117])

aggregation levels (e.g., census blocks or census block groups) due to privacy concerns. Similarly, certain types of events (e.g., cancer data) may only be available at city- or county-levels. Such datasets are often given in the form of polygons representing boundaries of the corresponding aggregation units. Some techniques take inputs in the grid format mainly for computational purposes (e.g., easier to enumerate contiguous rectangular-shaped candidate clusters) [154, 171]. In each grid cell, a count is recorded for different types of observations and events. When the original dataset provided is not in the grid format, grid-based methods will preprocess it into a grid format which may require additional assumptions for data given in polygon formats (e.g., homogeneous distribution within each polygon).

— **Trajectory:** As trajectory data are becoming available in greater varieties and volumes, it is also an important source of inputs for hotspot mapping. Examples of trajectory data

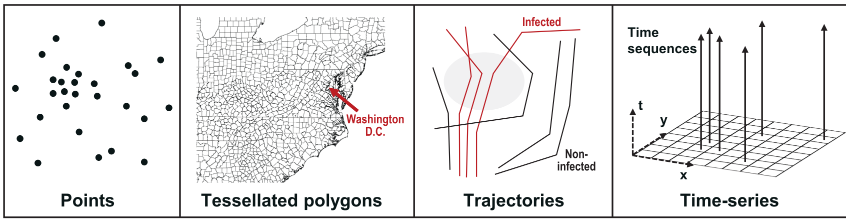


Fig. 2. Example data models commonly used for hotspot mapping.

that have been used in this field include social-media (e.g., a spatiotemporal sequence of geo-tagged tweets [187]) or smartphone trajectories of users; location traces of vehicles and properties (e.g., taxi trajectories [90, 122]; on-board-diagnostic data with hundreds of engine measurements [115]). Trajectory data can be used in two different ways. First, it can be used to identify hotspots that have a linear-path shape in a network space (e.g., a path with abnormally high energy consumption). Second, it can be used to identify hotspots as geographic regions [122], in which case events are no longer spatially static points but moving objects. This is especially valuable in epidemiology, in which a trace of locations provides much richer information than a static point (e.g., where an infected individual resides or is hospitalized) [187, 188]. One limitation of this type of data is its availability and quality (e.g., representativeness) for different types of applications.

— **Time-series:** In some application scenarios such as infectious disease monitoring, decision makers may be more interested in knowing about short-term dynamics reflected by the data rather than long-term phenomenons. Time-series data are very helpful for such purposes [106]. In time-series data, a point, polygon or grid cell becomes a sequence of observations at different timestamps (e.g., numbers of visits at a hospital on a daily basis [140]). This allows finding emerging hotspots, i.e., regions that were not but are becoming significant hotspots based on latest observations.

Sometimes the input datasets are a mixture of different data models based on their availability. For example, case data may be points of crime cases while control data may be polygons (e.g., census blocks) with population information.

3.3.3 Types of Spaces. Based on application contexts and characteristics of the phenomenon, different types of spaces may be used with input data. The two most common spaces used in hotspot mapping are Euclidean and network spaces.

Euclidean space is best suited when the goal of an application is to find geographic regions (e.g., a zone in a city) as hotspots or when the phenomenon may propagate and evolve with high-degrees of freedom (e.g., diseases that spread by air) [102, 150].

Network space should be used when the goal is to find a road segment, a path or a sub-network as a hotspot [51, 62, 91, 156, 199]. For example, traffic accidents mostly happen only on road networks, so a road segment or route often best represents the spatial footprint of the phenomenon. Similarly, in crime hotspot detection, if police officers are more interested in the travel distance of a serial criminal from a potential point (e.g., a residence location, which is unknown and needs to be enumerated) to a crime location [62], spatial networks should be used to model input data. In addition to road networks, other examples of networks include river networks, utility networks (e.g., powerlines, pipes [51]), communication networks, and social-spatial networks [36].

3.3.4 Levels of Uncertainty. Uncertainty may be introduced when data points are inaccurate either due to errors or privacy-related perturbation [90, 120, 155]. Inputs to statistically-robust clustering may also be provided at various aggregation levels, including no-aggregation (original point samples), aggregated with nearby cases within a certain distance, aggregated at census-block or county levels, and so on. Such aggregations also introduce additional uncertainty into the clustering process. The effects of these uncertainties on hotspot mapping have been examined and visualized by many [25, 52, 90, 117, 120, 155].

3.4 Statistical Modeling

Statistically-robust clustering uses a variety of spatial and spatiotemporal point processes to model different types of hotspots and their corresponding hypotheses, and has developed various test statistics to evaluate hotspot candidates. In the following, we first show a taxonomy of statistical models and then discuss formulations of test statistics as well as design strategies. It is worth-noting that while many of the formulations (e.g., test statistics) are originally proposed as a part of a technique (e.g., together with maximization algorithms), they can be generally used for different problems; an analogy is the modeling of loss functions in deep learning.

3.4.1 Spatial and Spatiotemporal Point Processes. Point processes have been developed based on a variety of statistical models to fit different types of input data models and application needs in hotspot detection. A point process mainly focuses on point distributions over space, and the following summary includes more generalized definitions that may also consider point attributes (e.g., related to marked point processes).

- **Bernoulli and Poisson:** In the Bernoulli model control data (e.g., population at risk for a disease) is a collection of discrete points in some space (e.g., Euclidean, network), and each point follows a Bernoulli distribution $X \sim Ber(p)$, where p is the probability of being an instance of an event (e.g., disease [102], crime [64], accident [196]). Statistical hypotheses are used to define hotspots (i.e., true clusters). The null hypothesis H_0 is that all points follow the same Bernoulli distribution (same p) in the study area, and the alternative hypothesis H_1 states that there exists a hotspot, where points inside follow $Ber(p_1)$ and outside follow $Ber(p_2)$, and $p_1 > p_2$. Instead of assuming p on individual points, Poisson point processes assumes expectations at regional levels [102].
- **Multivariate:** This model takes an integrated view of multiple types of events that are related to a common parent type. Examples include multiple symptoms that relate to a single disease [106, 144], multiple groups in a population [183], multiple actions (e.g., visiting a hospital, purchasing OTC medications) following an infection [140], and different types of crimes with similar motives. There are mainly two approaches to model the alternative hypotheses (H_0 still assumes a homogeneous process). The first approach considers a single H_1 that assumes there is a single parent event causing the increase of sub-event instances [106]. In contrast, the second considers a set of H_1, and each H_1 relates to a different subset of sub-events [140]. For example, both hospital and pharmacy visits are related to disease outbreak. However, pharmacy visits can also be caused by promotional events that are not related to an outbreak. Multiple alternative hypotheses enable such fine-grained characterization.
- **Multinomial:** This process also considers multiple types of events. However, these events are typically independent unlike the multivariate case [97, 207]. Examples of such types include races, species, unrelated diseases, and so on. This modeling is used when the goal is to find sub-regions where proportions of different types differ significantly from the rest. Thus, H_0 states that the proportion composition is the same across the study area, whereas H_1 states otherwise. Here, control data are typically not used.
- **Continuous:** This differs from the other processes by having continuous attributes on geo-located data points (e.g., air quality, house prices) rather than binary values or aggregated

counts. To model continuous values, methods typically use normal [89, 104], exponential [86], or other continuous distributions [49, 96]. Using normal distribution as an example, H_0 states that the mean is the same across the whole study area and H_1 states that points in hotspot regions follow distributions with higher means.

— **Expectation-based:** All the processes above are mainly designed for mapping spatial hotspots but can be extended to spatiotemporal use cases. In contrast, the expectation-based model is designed to focus on emerging (i.e., spatiotemporal) hotspots [135, 151]. It considers each location in a grid partition of space as a time-series of counts (e.g., number of visits to a hospital) following a Poisson process. Unlike the "inside vs. outside" view typically used in hypotheses of other processes, here it compares the most recent time period to the history of each spatial sub-region. H_0 states that the Poisson mean is the same for both the recent time window and past history, whereas H_1 says otherwise.

— **Bayesian:** This model also focuses on finding emerging hotspots on grid-aggregated data [94, 132, 136, 140, 141, 203]. As an example, [132] uses the Bayes' rule with the conjugate Gamma-Poisson distribution to model prior and posterior, where the mean of the Poisson is $p \cdot W$ where W is the control population at a location and p is the expected probability, and p further follows a Gamma prior. H_0 assumes that p follows the same Gamma distribution across the entire study area where H_1 states that there exists a sub-region where the prior is different from the rest.

3.4.2 Test Statistics and Measures. In the context of statistically-robust clustering, test statistics are used to assign a score to each candidate cluster for the purpose of comparison and ranking.

Early non-probabilistic test statistics are often used to find clusters of a fixed size (e.g., $s \times s$ squares) without underlying control distribution. With these simplifying assumptions, the pure cardinality of data points in a candidate is often sufficient as a test statistic. For more general cases with variable candidate sizes and underlying control distribution, simple extensions such as density or density ratio are used for normalization [91, 156, 196]. However, both density and density ratio are well-known to be strongly biased toward small candidates due to their monotonicity (e.g., any partitioning of a sub-region will yield a partition with a density greater than or equal to that of the sub-region). As a result, techniques based on these statistics lose detection power quickly as the size of a true hotspot increases [146, 213].

To mitigate the bias issues, most commonly-used test statistics in the frequentist framework adopted the likelihood ratio formulation (examples shown in Table 2) [63, 65, 102, 105, 146]. With data generation processes being defined by a parametric distribution, the general form of the likelihood ratio for a region S is $F(S) = \frac{Pr(Data|H_1(S))}{Pr(Data|H_0)}$, where $Pr(Data|H_1(S))$ and $Pr(Data|H_0)$ are the likelihoods of data being generated by the alternative hypothesis H_1 and null hypothesis H_0, respectively. A larger likelihood ratio indicates that an observation is more likely to exist under the alternative hypothesis. Different parametric distributions are used for different application scenarios (e.g., examples from Section 3.4.1), resulting in various forms of likelihood ratios. For example, Poisson based likelihood ratios can be used to identify outbreaks of diseases or hotspots of crime activities [63]; continuous versions can be used to find regions with high pollution; and expectation-based extensions are more sensitive to temporal changes [135]. Various studies have also explored integration of prior knowledge using regularizers in test statistics. For example, penalty terms are added to favor higher spatial resolution [70], discourage irregularly shaped and disconnected clusters [30], and promote dynamic clusters that change smoothly over time [191]. Methods were also developed to reduce the bias of test statistics across scales (cluster sizes) and improve detection power (e.g., average likelihood ratio [34, 69], spatial mixture index [207], non-deterministic normalization index [213]). For computational efficiency, simpler formulations of

Table 2. Examples of Test Statistics Using Point-Process Based Likelihood Ratios

Process types	Example test statistics	Interpretation
Poisson model	$$\log \left(\frac{n}{N \cdot \frac{a}{A}}\right)^n \cdot \left(\frac{N-n}{N \cdot \frac{A-a}{A}}\right)^{N-n} \cdot \mathbb{1}\left(n > N \cdot \frac{a}{A}\right)$$ where n and a are counts of case and control in a candidate cluster, respectively; N and A are the total counts of case and control of the study area; and the indicator function $\mathbb{1}(\cdot)$ guarantees the candidate is dense rather than sparse.	A log likelihood ratio (LR) using the Poisson point process. The denominators represent the expected number of cases inside and outside the candidate under H_0 (i.e., no hotspot) (e.g., [102, 146]).
Continuous value model	$$N \log \sigma + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - N/2 - N \log \sigma'$$ where n and N are the number of samples of a candidate cluster and the study area, respectively; x_i is the value of a sample in the candidate; μ and σ are the mean and standard deviation of all N samples; σ' is a standard deviation estimated assuming samples inside and outside a cluster follow different means (i.e., H_1).	A log LR using the continuous normal model, where H_0 states all samples follow the same mean-value and H_1 states that there exists a candidate region following a process with a higher mean (e.g., [89, 104]).
Multinomial model	$$\log \frac{\prod_{i=1}^{k} p_i^{n_i} q_i^{N_i - n_i}}{\prod_{i=1}^{k} (p_i')^{n_i} \cdot (q_i')^{N_i - n_i}}$$ where k is the number of types (e.g., disease types), n_i and N_i are number of type-i samples of a candidate cluster and the study area, respectively; $p_i = \frac{n_i}{n}$ and $q_i = \frac{N_i - n_i}{N-n}$ are the probability of a point being of type-i inside and outside the candidate; and $p_i' = q_i' = N_i/N$ is the probability of a point being of type-i in the whole data.	A log LR using the multinomial model, where H_0 states $p_i = q_i, \forall i = 1, \ldots, k$ across the whole space whereas H_1 states some regions have different complexions (e.g., [97, 207]).
Expectation based model	$$\log \max\left(\left(\frac{n}{b}\right)^n \cdot e^{b-n},\ 1\right)$$ where n is the number of cases observed in a candidate cluster at current timestamp t, b is the baseline learned from historical data, and the $\max(\cdot, \cdot)$ function is used to filter out scenarios where $n < b$.	Compared to the Poisson model shown above, this log LR compares the concentration of samples in a candidate cluster at t to its historical expectation b from $t' < t$ rather than the concentration outside the candidate (e.g., [135, 151]).

test statistics are used, such as the elevated mean scan statistic [167]. All the above-mentioned test statistics assume the underlying probability distribution obeys a parametric model. In scenarios where prior knowledge on underlying distributions does not exist, nonparametric model based test statistics are introduced, such as Berk Jones [21], nonparametric scan statistics [36, 142], and Higher Criticism [56]. Without assumptions on statistical distributions, nonparametric test statistics often rely on empirical p-values for significance testing. There are also works that use a proxy to represent the underlying distributions. For example, for the continuous-value model, the Mann–Whitney statistic is used to measure the distribution of the rankings of point values instead of the actual values [49]. Finally, these test statistics can also be applied to identify significant sub-graphs (e.g., based on Steiner-connectivity) in network or graph data as summarized in [25].

Many Bayesian approaches have also been developed to complement the frequentist view for significant clustering [139]. Advantages of Bayesian measures include the capacity of incorporating informative prior information, when available; the flexibility of considering multiple hypotheses; and direct calculation of posterior probabilities. In general, Bayesian based methods do not rely on test statistics; rather, they directly calculate posterior probabilities using the Bayes' rule (e.g., univariate [132, 148], multivariate [140, 149], and PANDA-CDCA Bayesian network [43, 95]). With that being said, the Bayesian posterior probability can also be considered as a test statistic, and empirical thresholds can be applied as an approximation of significance levels in the frequentist setup. There are also frequentist methods that use the Bayesian modeling to design test statistics, such as the Bayes factor [226] and anomalous group detection [50].

3.4.3 Design Strategies of Test Statistics and Measures. A well-designed test statistic should be unbiased and ideally have the optimal asymptotic detection power under a significance level. Many theoretical insights and results have been developed to reveal major issues to consider during the design and recommend strategies to address them both statistically and computationally. For example, most of the existing test statistics used in research and applications are based on ratios between maximum likelihoods of hypotheses H_1 and H_0. Many of these standard formulations, including Kulldorff's likelihood ratio, have also been shown to be scale-sensitive. Specifically, these test statistics can be strongly biased toward small candidates [34, 174, 204, 213, 215] (albeit much better than density [145, 213]).

In significance testing, statistically-robust clustering techniques commonly estimate the distribution of test statistics using the best candidate (i.e., the maximum score) from the H_0-point-distribution in each Monte Carlo trial, which is needed to avoid multiple testing (Section 3.6.1). Thus, an intuitive interpretation behind the bias is that the distribution of the maximum score—as a random variable—is scale-dependent [61, 172, 174, 204]. In other words, the mean of maximum score of smaller candidates tends to strongly dominate that of larger candidates. As a remedy, a penalty term was constructed to correct the bias and restore the optimal asymptotic detection power for hotspots with different sizes [61, 166, 174, 204]. A further improvement uses square-roots of log likelihood ratios to construct the penalty term, requiring less case-specific design [172]; the extension also provides additional computational efficiency (near linear-time). Another approach is to correct the bias during significance testing, where a block criterion is developed to adjust critical values (i.e., thresholds on test statistic values to remove insignificant results) on candidates belonging to different sizes or size-groups [174, 204], guaranteeing optimal asymptotic power of detection. Average likelihood ratio approaches have also been developed to reduce the bias, where each score is a weighted average of those of neighboring candidates [34, 69]. According to theoretical results in [34], the original likelihood ratio in the spatial scan statistic has optimal detection power only for hotspots with smallest footprints whereas the average likelihood ratio has optimal power at all scales. The analyses were carried out for the discrete one-dimensional scenario. Another view of the average likelihood ratio approach is that it uses marginal likelihoods instead of maximum likelihoods [139], which may better utilize secondary cluster information as in the Bayesian test statistics [132, 148, 149]. In addition to the pure-scale interpretation of the bias, a spatial interpretation has been developed [213, 215]. The analysis shows that the likelihoods or probabilities used to score a candidate in many approaches are based on a space bi-partitioning where one partition represents the candidate C and the other for the outside. Likelihoods calculated using the bi-partition implicitly assume that the best candidate C_{ran} in a random dataset still locates in C's partition, which is rarely the case. This leads to a big underestimation especially for the likelihood of H_0 since the location of the best candidate should be nondeterministic in a purely random point process. To correct the bias, spatial-nondeterminism-aware test statistics were proposed [207, 213], which can be computed using dual-level Monte Carlo simulation [207]. The use of another simulation introduces additional computation overhead but can be generally applied for different point processes (e.g., Bernoulli, Poisson, multinomial), test statistics (e.g., density, likelihood ratio, spatial mixture index), candidate shapes (or enumeration spaces), maximization algorithms, and so on.

Table 3. Region Enumeration and Maximization

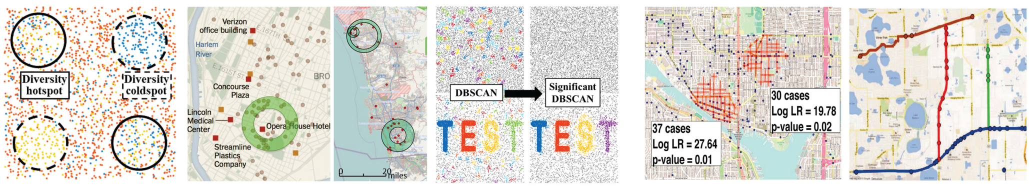
Classification criteria	Categories	Examples
Definition	Pre-defined shape	• Euclidean: Circles [65, 78, 102, 116, 122], ellipses [105], rectangles [145, 146], rings [63, 64] • Network: Linear paths [51, 91, 156, 196, 199], sub-networks with shape constraints [62, 198]
	Flexible shape	• Problem-specific heuristics: ULS [128, 161, 162], MST [45], FlexScan [194, 200], trajectory [53] • Meta-heuristics: Genetic algorithm [58, 175], simulated annealing [57], random walk [92] • General clustering based [211, 214]
Constraints	Computation-driven	• Euclidean: Finite enumeration space by data points (e.g., two-point circles, concentric four-point rings) [63, 64, 102, 145, 146, 204] • Network: Finite space plus shortest paths [51, 156, 196, 198], simple paths [199], or real-world trajectories [115]; relaxed sub-graph connectivity [11]
	Quality-driven	• Spatial contiguity/connectivity [11, 25, 173, 189–191] • Density contiguity [211, 214] • Shape regularity [30, 45, 58, 60] • Maximum population [63, 102]
Algorithms	Exact	• Bound based [63, 64, 145, 146, 150, 196, 197, 199, 214] • Incremental updates [133, 140, 207] • Linear-time subset scanning [25, 36, 137, 144, 190] • Shortest-path-tree [156, 196], FMGT [199]
	Approximation	• Approximation-ratio based [7, 8, 122, 173, 228] • Asymptotic-bound based [172, 174, 204]
	Heuristic	• Data sampling [121], space reduction [63, 207], meta-heuristics [57, 58, 92]

3.5 Region Enumeration and Maximization

Region enumeration is the core computational step in statistically robust clustering, which examines a mathematically well-defined candidate space of clusters to identify the best candidate that maximizes the test statistic for further significance testing. While this routine only returns one cluster due to maximization, techniques in the literature often repetitively apply it as a sub-routine to find arbitrary number of clusters, i.e., by removing the best cluster (if significant) from data after each round and then starting a new round for the next (e.g., [102, 145, 207]). If the best cluster at the current step is not significant, the algorithm terminates and returns all significant clusters. This enumeration-and-maximization (or one-cluster-at-a-time) scheme is widely used for statistically-robust clustering because of the fact that it reduces the mutual statistical influence (a.k.a. shadow effects) [112, 139] between multiple clusters in the same dataset (Section 3.6.1).

In the following, we will first summarize definitions of enumeration spaces and constraints that are popularly used in the literature, followed by maximization algorithms as well as acceleration techniques of various types. A summary taxonomy is provided in Table 3.

3.5.1 Enumeration Spaces. Definitions of enumeration spaces often take considerations of both application needs and computational feasibility/efficiency. These definitions can be well classified



(a) Circles, rings and irregular shapes (Euclidean) (b) Circles and linear shapes (network)

Fig. 3. Example shapes used to define enumeration spaces. (a) Circular-shaped hotspots and coldspots of diversity [207]; ring-shaped hotspots for finding serial crimes and the source of a disease [63, 64]; and arbitrarily-shaped hotspots by Significant DBSCAN, which also shows its robustness against spurious clusters (top: random data without clusters; bottom: clustered data) [211, 214]. (b) Iso-distance hotspots of crimes on road networks [65] and linear hotspots of traffic accidents (pedestrian fatalities) [156, 196].

using cluster shapes (i.e., predefined, flexible), with additional consideration of data types and data spaces (e.g., Euclidean, network).

Predefined geometric shapes: Majority of statistically robust clustering techniques were developed for a specific type of geometric shapes (e.g., circle, ellipse, rectangle, ring, linear path) that can represent real-world phenomenons in target application domains. The corresponding enumeration space then covers all valid sub-regions of the desired shape. Circle, for example, is the most widely used shape in epidemiology and public health research [1, 42, 65, 78, 84, 102, 116, 122] due to natural diffusion mechanisms, which tends to make infectious disease spread to an approximate circular shape; and due to computational convenience. Circles were extended to ellipses to better model effects of spatial anisotropy and provide more flexibility beyond circles [105]. Rectangles were also used as proxies for circles/ellipses because of the computational convenience they offer on gridded inputs [145, 146]. Ring shapes were developed for applications in public safety and criminology research. According to the routine activity theory, serial criminals (e.g., arsonists) tend to commit crimes neither too close to nor too far away from their home to lower both the risk of being recognized by neighbors and the cost of travel [64]. Thus, ring-shaped zones [62–64] were used to depict the footprints of serial criminals. While these shapes are more natural to point data (Section 3.3), they have also been used with polygon data (e.g., county polygons with total number of disease cases and population). In such cases, polygon data are first preprocessed to center points to perform region enumeration, and then in the output phase, the polygons containing points from each significant cluster are merged to generate the final results.

In transportation or city infrastructure related applications, vast majority of techniques use network space because data (e.g., locations of traffic accidents or pedestrian fatalities) or phenomenons are based on road, water, or utility networks. Correspondingly, linear paths or intersections along network edges are used to model shapes of candidate clusters [51, 91, 156, 184, 196, 199]. In addition, network-space has also been used to identify "areal" clusters (e.g., a circle becomes a sub-network where the network distances from all locations to a center are less than or equal to r), especially in urban areas where network-distance is a better measure of spatial connectivity than Euclidean distance [62, 198]. For example, two locations separated by a river can be very close in Euclidean space but far apart by network distance. Figure 3 shows several common examples of shapes used to define the enumeration space.

From a computational perspective, further simplifying constraints (e.g., two-point circles, shortest paths) are often needed to guarantee tractable search spaces; otherwise, a simple circular shape in a continuous space can lead to an infinite number of candidates. This will be detailed in Section 3.5.2.

Flexible shapes: A variety of techniques have been developed to detect arbitrarily shaped spatial hotspots to better capture the complex boundaries of real-world patterns caused by underlying physical and social contexts. Due to the non-parametric nature of arbitrary shapes, candidate enumeration spaces used for this paradigm are typically not well-defined. Instead, the vast majority of techniques define specific local search heuristics to iteratively go over a set of arbitrarily shaped candidates until certain criteria are met. A simple and popular formulation of this type is called an Upper Level Set (ULS) scan statistic [128, 161, 162]. ULS requires a tessellation of space (e.g., by geographic unit boundary; grid-partitioning), and starts by calculating a test statistic for each cell. Then, a cell belongs to an ULS of threshold t if its test statistic value $\geq t$. All connected components in an ULS form the set of candidate clusters, with potentially arbitrary shapes. A ULS tree is constructed by applying a series of thresholds and significant clusters are merged to generate final results. Similarly, a minimum-spanning-tree (MST) based approach enumerates connected sub-trees that are part of an MST of a tessellated input data, where the edge weights are inferred directly or indirectly by test statistic values [45]. There also exists a FlexScan approach that enumerates all possible connected subsets in a tessellation in a brute-force manner [194, 200]. However, due to the exponential nature of the search space, it can only identify small clusters (e.g., <30 cells [59, 138, 200]). A trajectory based method was developed to relax the requirement of a tessellation [53], where data points are sequentially linked to form a trajectory using test statistic values; clusters are then enumerated by going through connected segments.

Meta-heuristics have also been explored to generate irregularly shaped candidates, and most still require a tessellated input space. A simulated annealing formulation [57] was used to enumerate candidates by iteratively search over neighboring choices, each defined as another candidate that differs by one-cell from the current candidate. As a simulated annealing approach, the probability of moving into an inferior neighboring choice gradually decreases as the iteration number increases. Many genetic algorithm based formulations [58, 175] were also developed to generate arbitrarily shaped candidates, where off-springs are evaluated by a choice of test statistics. In addition, random-walk based method was developed to search free-form regions using a heuristic metric [92].

To reduce over-complicated "tree shapes" of clusters that are generated as a result of the freedom of arbitrary shapes but are spatially not meaningful, regularization functions were developed to penalize such shapes based on geometric compactness defined by various criteria (e.g., ratio between a cluster's area to its convex hull [58, 60], sparsity [140, 179]). The regularization terms were also used in multi-objective formulations to generate pareto sets where clustering results are provided under different levels of shape-regularity constraints [30].

While traditional clustering techniques have provided many efficient schemes to detect irregularly shaped clusters (e.g., DBSCAN [66], OPTICS [14], spectral clustering [202], Chameleon [99], HDBSCAN [29, 152], and so on.), these methods, as introduced earlier, do not incorporate statistical rigor required by domain applications of spatial hotspot detection. Recently, a statistically robust formulation of DBSCAN—Significant DBSCAN—was developed to explicitly control the rate of spurious clusters [211, 214], though it currently does not model the distribution of control points described in Table 1, Section 3.3.1.

3.5.2 Constraints. There are two types of constraints widely used for spatial hotspot detection. The first type concerns mostly about computational feasibility, i.e., to mathematically constrain the enumeration space to a finite and tractable set of candidates (e.g., non-exponential); whereas the second type of constraints is used to improve solution quality.

Computation-driven constraints: Given that even an enumeration space of predefined shapes may have infinite number of candidates, techniques rely on constrained parameterization of shapes to make the enumeration feasible. In the Euclidean space, circles are often constrained by two points—one at the center and the other on the circumference, corresponding to a $O(N^2)$ search space [102]. A center point may come from the set of event incidents (e.g., disease cases), underlying control data (e.g., population points), or an additional input set of candidate centers (e.g., centers of cells in a grid [65]). This two-point definition has a natural explanation in epidemiology or similar domains, where the spread of an event (e.g., an infectious disease) propagates from a center source. Similarly, a ring-shaped cluster is parameterized by two concentric circles [63, 64]. Since the inner circle may have a low density of points (e.g., serial criminals often do not commit crimes too close to their residence locations), its enumeration space uses three-point circles where all three points lie on the circumference. The outer circle is still defined by one-point on the circumference as the center location has been determined by the inner circle; this leads to a $O(N^4)$ search space. For rectangular shapes, candidates are constrained to rectangles whose boundaries on all sides touch at least one data point or grid cell, yielding the same $O(N^4)$ space [145, 146, 204]. Many other similar constraints are used for ellipses, non-orthogonal ellipses, and so on. Since flexibly shaped patterns can hardly be parameterized, there is often no further computation-driven constraints beyond the implicit constraints implied by different heuristic search strategies (e.g., local constraints).

In the network space, shortest-path is a common constraint for linear path enumeration [51, 156, 196]. While this may be a reasonable approximation for people's travel behaviors, many real-world paths such as bus routes do not always follow shortest-paths. Instead, a relaxed simple-path constraint—any path without self-intersections—quickly increases the potential space to $O(N!)$ [199]. Another extension beyond shortest paths is to directly define the enumeration space using real-world trajectories and their connected sub-components [115]. For network-distance based circles [198], rings [62], and so on, the same shortest-path constraint is used to define the space that is reachable from a center (e.g., by a threshold r).

Solution-quality-driven constraints: This type of constraints is developed to further reduce non-meaningful outputs beyond statistical modeling, and is irrespective of computational considerations. One most important constraint is the spatial contiguity constraint, requiring that any two locations inside a cluster are mutually reachable by its inner space. This typically increases the computational cost [127]; in fact, a linear algorithm exists if this constraint is taken out [137]. A relaxed version of this contiguity constraint is the proximity constraint [137, 189–191], which requires segments inside a candidate to be within a local neighborhood defined either by a spatial distance or k-nearest neighbors; the segments are allowed to be spatially disjoint. Similarly, density-contiguity within each hotspot can also be enforced via local constraints [211, 214]. In network space, contiguity constraints are commonly defined by sub-graph connectivity or its relaxed formulations [11, 25]. Another widely used constraint is a maximum population threshold, which eliminates clusters that cover more than 50% of underlying population used in control data (e.g., [102]); otherwise, a cluster is considered as a general phenomenon rather than a "hotspot". As discussed in Section 3.5.1, regularity constraints were also used in arbitrarily-shaped cluster detection to avoid over complicated tree-like patterns [30, 45, 58, 60].

3.5.3 Maximization Algorithms. Statistically, robust clustering techniques typically detect only the best cluster in each round, and remove it before searching for the next cluster to avoid shadowing effects between patterns (detailed in Section 3.6.1). Thus, after an enumeration space is defined, maximization algorithms are used to identify the cluster candidate with the highest test statistic value. For example, test statistics $T(S)$ in Table 2 can be used as objective functions, and S represents a cluster candidate where cluster-specific inputs (e.g., counts of case n and control a for the Poisson-based test statistic) are calculated. Many new computational techniques have been developed by the data mining community to reduce the cost of maximization. In the following, we summarize these acceleration algorithms in three broad categories: exact, approximation, and heuristic algorithms.

Exact algorithms: The vast majority of maximization algorithms fall into this category, which means they can guarantee that the output cluster indeed achieves the maximum test statistic value within an explicitly defined search space (i.e., the cluster is the same as that of a brute-force algorithm). As a result, exact methods mainly focus on pre-defined shapes (Sections 3.5.1 and 3.5.2).

The initial algorithmic improvement was introduced by SaTScan [4]. The algorithm is only applicable to circular-shaped candidate enumeration, which pre-sorts all the points by distance to each circle center. The pre-sorting frees a candidate evaluation step from performing range queries to identify which case or control points are inside the candidate (inferred by the order after sorting). This reduces the overall enumeration and evaluation cost from $O(N^3)$ to $O(N^2 \cdot \log N)$, assuming both cardinalities of center and circumference points are proportional to N. Since then, many significant improvements have been developed to further reduce the cost.

The first major type of strategies uses pruning, i.e., to eliminate subsets of enumeration spaces where no candidate may have the maximum test statistic value. An overlap-kd-tree based approach was first developed for finding square-shaped clusters in a two-dimensional grid [145]. Overlaps are added to original kd-tree partitions so that candidates that span across partitions are not missed. Then, a branch-and-bound pruning strategy is used to enumerate through the tree. Since only the optimal cluster is selected in each round, the existing maximum score is used as thresholds to prune branches with a smaller upper-bound score. Later works have extended the overlap-kd-tree and branch-and-bound strategies to cover rectangular shaped clusters [146], and for multi-dimensional data [150]. Similarly, filter-and-refine strategies were designed for various pre-defined shapes (e.g., circles [65], rings [63, 64], ellipses [197], linear paths [156, 196, 199], and iso-distance sub-networks [198]) in non-tessellated input spaces. Leveraging that calculating the number of points in a sub-grid (e.g., $O(1)$ using an integral image [214]) is much faster than in a continuous space, these techniques typically use a discretized grid space to compute upper bounds of test statistic values of candidates in continuous spaces. A refining step is only carried out to check exact candidates if the upper bounds exceed a threshold (either user-specified or using the current maximum). These bounds are often designed specifically for each shape based on their characteristics. In addition, incremental update rules have also been developed to efficiently compute test statistic scores of neighboring candidates (e.g., $O(1)$ amortized cost [133, 140, 207]); the update rules depend on specific test statistics.

Another widely adopted strategy is based on a **linear-time subset scanning** (**LTSS**) property first introduced in [137] for tessellated data. LTSS only applies to scenarios where components (e.g., grid cells) of a cluster are unconstrained by spatial contiguity, i.e., it can be any arbitrary subset of data elements. By relaxing the contiguity constraint, LTSS states that, for many test statistic functions F, there exists an ordering function G, such that the subset with a cardinality of k always reaches the maximum F score if its members have the top-k individual G scores in the data [137]. Assuming a sorted order of individual components by G has been given, LTSS allows finding the optimal subset in linear time out of an exponential enumeration space. It has been shown that the LTSS property can be satisfied by a variety of test statistic functions [123, 144, 190], including Kulldorff's scan statistic, expectation-based Poisson or Gaussian function, separable exponential families and many quasi-convex functions. The method has also been used in non-parametric scan statistics [36], GraphScan [25], and so on. A few special cases have been proven to make LTSS able to output contiguous subsets as clusters (e.g., non-existence of break-tires [36]). As mentioned above, one key limitation of LTSS is the lack of spatial contiguity in each cluster. To mitigate this, a proximity-constrained extension was developed [189, 191]. Instead of applying LTSS on the entire data, it sequentially goes through all data components (e.g., grid cells) and apply LTSS only on the k-nearest neighbors of each component, or neighbors within its r-radius neighborhood. Discontiguity is still allowed for subsets within each local neighborhood, but this avoids long-distance separation.

In the network space, shortest-path trees were used to efficiently enumerate candidate paths between all pairs of points [156, 196]. Since data points may add a large number of low-degree (two-degree) nodes to a network, dynamic segmentation was employed to construct the shortest-path tree with only original network nodes and then dynamically retrieve shortest paths between data points. For the all-simple-path enumeration space, the number of candidates can be factorial [199]. Thus, more aggressive pruning methods were developed. Instead of directly using network nodes to build a simple-path tree, fragmentation-multi-graph-trees are used, where each tree node represents a sub-network and paths within each node can be dynamically computed and reused. Pruning and filter-and-refine strategies described above for Euclidean space based methods were also used in network-based hotspot detection [198]. To handle high computational cost (e.g., NP-hardness [11, 25], high-order combinatorials [180, 181]) incurred by constraints on sub-graph connectivity, compactness or sparsity, methods often use relaxed reformulations such as semi-definite programming [11] or spectral graph theory based relaxations [180, 181], where results (e.g., detection power) can be rigorously analyzed.

Approximation and heuristic algorithms: We discuss algorithms in these two categories together as it is hard to draw an exact boundary given the various definitions used by techniques. In computational theory, an approximation algorithm should guarantee that it always returns a solution S_{apx} whose objective function value $F(S_{apx})$ (i.e., test statistic value for hotspot mapping) is within ϵ of the optimal solution $F(S_{opt})$. For hotspot mapping, one challenge in clearly defining "approximation" is that—from an application perspective—it is as important to know "where" and "how large" a hotspot is in addition to a F value; this can be further complicated considering the large number of shapes used. In addition, this problem has broad interests by researchers from statistics and computer science, making the use of "approximation" more diversified.

A major difficulty in approximating the solution of spatial hotspots is due to the combinatorial nature of candidate enumeration in a finite and discrete space of parameters. To bridge this gap, an approximation scheme was provided in [8] to search for the maximum-scoring region over a set of axis-parallel rectangles. This approach requires the scoring function F to be a convex discrepancy function, and one key idea was to simplify the optimization using a set of linear functions to construct an approximation. Using Kulldorff's spatial scan statistic as an example of F, the approach was able to reduce the search cost from $O(N^4)$ to $P(\frac{1}{\epsilon}N^2 \log^2 N)$, where ϵ is the gap to optimality. The linear-function based approximation was further improved in [7] leveraging the fact that only the maximum-scoring region is needed for each round. It also provides a new approximation algorithm for grid-based input data with an $O(g^3 \cdot poly(\log g, \frac{1}{\epsilon}))$ complexity, where g is the number of columns and rows in a $g \times g$ grid. For rectangular shaped hotspot detection, an approximation-version of the exact brand-and-bound algorithm was also given in [146]. This extension uses a relaxed instead of an exact upper-bound and thus allows more aggressive pruning. Probabilistic results were provided to describe how likely each individual relaxed-bound will remain correct, though no distance to global optimality was given. A scalable sampling-based approach was also developed for trajectory data, which established approximation relationship among variable error bounds, cluster shape, sample size, and execution time [122]; it also considered various ways to spatially sample and approximate the input trajectories. In network space, approximation algorithms have also been developed by analyzing and using the submodularity of objective functions [173]. For sparsity constraints, efficient stochastic gradient descent algorithms have been developed, which have linear convergence rates for various objective functions [228]. Another type of approximation algorithms aims at maintaining the optimality with respect to asymptotic detection power, guaranteeing that the approach will give the optimal solution with sufficiently large number of samples [172, 174, 204]. Related results were first achieved for one-dimension data, where regions are defined as intervals [174]. This approximation only needs a $O(N \log N)$ search over the $O(N^2)$ enumeration space. The analysis has been later extended to two-dimensional space [204], where the best candidate in a set of $O(N^4)$ axis-parallel rectangles can be approximated by an economical set of rectangles, with an $O(N \log^2 N)$ cost. Additional assumptions typically used by these results include continuous surface of control data, independence between samples, data types and certain test statistic functions.

Several other methods have a clearer heuristic nature. In ring-shaped hotspot detection, for example, a Delaunay-triangulation was used to generate guesses of ring centers instead of exhaustively going over centers of all three-point circles, reducing the center space from $O(N^3)$ to $O(N)$ [63]. For multinomial-model-based hotspot detection (e.g., hotspots of high-diversity), a heuristic was developed to gradually narrow search volume as candidate sizes grow, based on the observation that candidate scores tend to converge (i.e., stabilize) as their sizes get larger [207]. An empirical sampling based approach was also developed to use only a subset of all data points to perform hotspot mapping [121]. Finally, the metaheuristics reviewed for flexible-shaped hotspot enumeration (e.g., [57, 58, 92]) can also be considered as heuristic-based optimization strategies. While heuristic approaches, in general, will not guarantee solution quality, many have demonstrated high power and performance through empirical results.

3.6 Hypothesis Testing

Hypothesis testing is the final step in each round of hotspot detection, which determines if the optimal cluster is a true hotspot or a spurious pattern.

3.6.1 Region Maximization: A Revisit. There are two major reasons for selecting and testing only the optimal region in a single round [102, 112, 139, 146, 215, 225]. First and most importantly, returning only the optimal region allows robust control over the false positive rate given a significance level α; otherwise, detecting multiple regions (e.g., candidates at different locations or with different sizes) at the same time may be susceptible to multi-testing, i.e., increasing the chance of having a type-I error. As suggested by many studies, this maximization can be viewed as an automatic adjustment of significance levels for different groups of candidates (e.g., candidates of the same size) [172, 174, 204]. The intuition can also be easily explained from a data perspective [207, 215]. Given M datasets generated under a null hypothesis H_0, having a critical value—a score threshold derived to determine a candidate's significance at the α level—on the maximum score makes sure that a spurious cluster can only be falsely reported as a hotspot in smaller than αM datasets. In other words, to get a spurious detection in a dataset under H_0, its maximum score has to pass the threshold on the maximum score. The second reason is that this helps reduce the mutual shadow effect between true hotspots [112, 225]; e.g., eliminating a true hotspot from data will make the signal of a secondary true hotspot statistically stronger.

In order to detect multiple clusters, a significant hotspot will be removed from data (both its case and control points), and the same detection and testing procedure will be repeated on the updated data. While this may increase the risk of reporting a spurious hotspot at the dataset-level, the effect is very slight as the execution of any following round is conditional on the detection of a significant cluster in the previous round (i.e., strict unidirectional dependence between rounds). There have also been many studies trying to increase the detection power for multi-hotspot scenarios by adjusting test statistic functions [112, 225].

Table 4. Statistical Hypothesis Testing

Paradigms	Aspects	Examples
Frequentist	Modeling	• Random locations [64, 102, 146, 211] or fixed location with random values [89, 104, 207] • Scale-independent [102, 156, 194], scale-dependent [174, 204, 213], or unified [215] testing • Bayesian as sub-models [50, 123, 226] • Nonparametric empirical p-values [36, 142]
	Computation	• Monte-Carlo simulation accelerated by bounds [64, 146, 211], shared computation [156, 199], reduction [215], and so on. • Test statistic distribution approximation [39, 40, 75]
Bayesian	Modeling	• Direct priors [140, 149] or priors as parametric distributions [119, 139] • Direct posterior probabilities as results [132, 140, 148, 149] or with added tests (e.g., empirical) [136, 179] • Rank-based [168, 169]
	Computation	• Fast subset sums [137, 143, 179] • Conjugate priors [132]

3.6.2 Two Paradigms. Both frequentist and Bayesian-based paradigms have been explored to test the significance of a cluster. The frequentist approach is used for both spatial and spatiotemporal hotspot detection. While the Bayesian variation was originally proposed and mainly used for the latter (e.g., emerging hotspot detection using time-series data over space), it can still be used if prior information is available. Table 4 shows a summary of the two paradigms.

Frequentist Approach. Frequentist based testing is the classic in this field and still dominates the use in related research and applications, partially because of the tradition and the popularity of the software SaTScan. The frequentist view considers the data as an aggregation of the past, so null hypotheses based on it typically assume an entire dataset is generated by a complete random process (e.g., homogeneous Poisson process). The exact definitions of the hypotheses differ by statistical models discussed in Section 3.4.1.

Figure 4 shows several examples of random point distributions following different point processes in the Euclidean space. Figures 4(a1) and (a2) show example datasets generated by Bernoulli-based **spatial point processes (SPP)** under H_1 and H_0, respectively; where the black and orange points are control and case points, respectively. Under the null hypothesis H_0, each control point has an identical probability of being a case point (e.g., a COVID-19 case). In contrast, H_1 in (a1) has a higher rate inside hotspot regions and lower outside. Figures 4(b1) and (b2) are example point distributions from the continuous Poisson point processes. Here, H_0 assumes a homogeneous underlying probability density, which is also known as the complete spatial randomness. H_1 contains a region with a higher probability density of generating points (i.e., a hotspot). In Figure 4(c2), H_0 of a normal-model-based point process assumes the values observed at all data points (e.g., PM-2.5 readings on air quality sensors) are generated by the same distribution. Instead, a hotspot in H_1 has a higher mean (of the normal model) as shown in Figure 4(c1). Finally, Figures 4(d1) and (d2) show example distributions from a multinomial-model-based point process, where H_1 in (d1) contains a region with significantly higher diversification of point types (e.g., biodiversity) and (d2) contains a random redistribution of the point types over space. Figure 5 further shows examples of Bernoulli-based and Poisson point processes in a network space with linear hotspots.

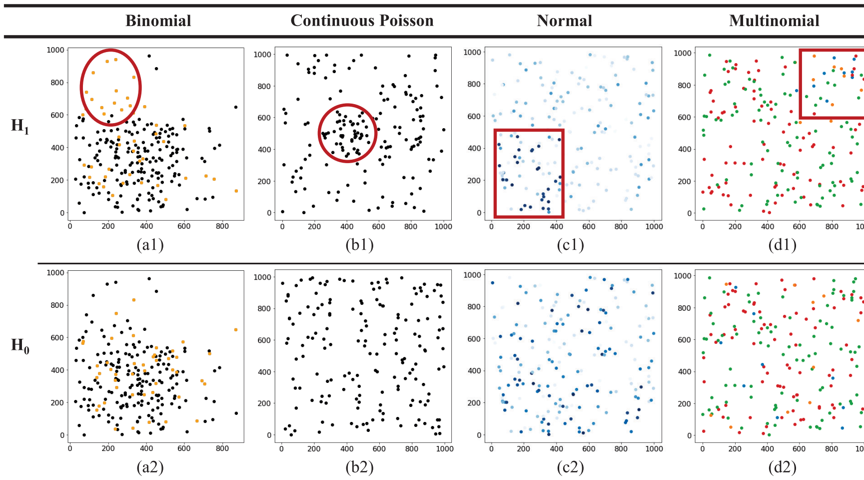
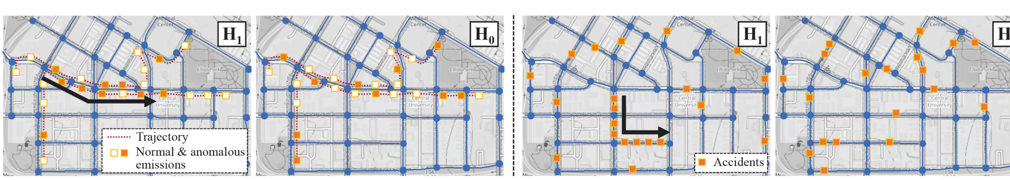


Fig. 4. Example statistical models for SPP in Euclidean space. The top-row shows observed data (i.e., input) with a true hotspot inserted using H_1; and the bottom row shows examples of simulation datasets generated by H_0 in a Monte Carlo trial. Note that SPPs by default model the spatial distribution of points. While in this example point locations are allowed to be fixed (can be random), it can be interpreted as having the space confined to a set of spatial coordinates.



(a) Binomial SPP on a network (b) Poisson SPP on a network

Fig. 5. Examples of SPP in network space.

Based on the hypotheses, the test will produce a p-value for the optimal cluster C^*_{obs} in the observed data D_{obs}: $P_{H_0}(C^*_{obs}, D_{obs}) = P\{F(C^*_{obs}, D_{obs}) \leq F(C^*_{H_0}, D_{H_0})\}$, where F is the test statistic function and $C^*_{H_0}$ is the optimal cluster in data D_{H_0} following H_0.

Given the combined complexity of point processes, test statistics, enumeration spaces, and maximization algorithms, the distribution of $F(C^*_{H_0}, D_{H_0})$ in general is unknown in closed-form, except in over-simplified special cases that are often not useful in practice [146]. Thus, there have been two lines of strategies developed to estimate the p-value.

The most common strategy is to approximate the distribution via Monte Carlo simulation, i.e., generating a large number of randomized datasets (e.g., 999, 9,999) following H_0 and running the same maximization algorithm on all datasets to simulate the distribution of $F(C^*_{H_0}, D_{H_0})$ (e.g., [102, 174, 188, 204]). Figure 4 (bottom row) shows randomization examples generated under H_0, where each can be considered as the simulated dataset in a Monte Carlo trial. Note that the spatial distribution of control points or base locations can either remain the same or be redistributed for Bernoulli, normal and multinomial point processes [102, 104, 146, 207]. In contrast, the control for continuous Poisson and many other processes is assumed to be a continuous surface, so the case point locations need to be redistributed in each trial. This Monte Carlo approach is broadly applicable as long as H_0 is well defined, at the price of greatly increased computational cost.

In addition to the computational enhancements discussed in Section 3.5.3, which can be used as efficient sub-routines in Monte Carlo trials, a variety of dedicated algorithms have been developed to reduce the simulation cost (e.g., [64, 146, 150, 196]). One convenient characteristic that has been leveraged by many approaches is that we do not need to know the exact maximum value $F(C^*_{H_0}, D_{H_0})$ for each random dataset D_{H_0}; instead, to calculate the p-value $P_{H_0}(C^*_{obs}, D_{obs})$, all we need is whether $F(C^*_{obs}, D_{obs}) \leq F(C^*_{H_0}, D_{H_0})$ in the trials. Using this characteristic, the maximum score from the observed data $F(C^*_{obs}, D_{obs})$ is directly used as the bound in both branch-and-bound [150] or filter-and-refine algorithms [64, 115, 156, 196, 199] to aggressively narrow search spaces. In addition, once a candidate is found to have a greater score than $F(C^*_{obs}, D_{obs})$, the rest of the trial can be skipped. This also allows the use of non-exact but low-cost data structures during the simulation. As an example, a discrete-DBSCAN was used to replace the original DBSCAN for Monte Carlo trials to pre-estimate upper and lower bounds of $F(C^*_{H_0}, D_{H_0})$ [211, 214]; the original DBSCAN is only re-applied for a subset of trials where the bounds are insufficient to deduce the relationship. Another widely adopted trick is early-termination [4]. As its name suggests, this method terminates the whole Monte Carlo simulation once the results from completed trials are sufficient to conclude the significance. This is extremely efficient if a detection is insignificant. For example, for a significance level $\alpha = 0.01$, the minimum number of trials needed to accept H_0 is 9 out of 999. It was also shown in [214] that early-termination is highly efficient for general hotspot detection formulations.

Shared computation has also been used for tasks that remain stationary across Monte Carlo trials. For example, many expensive operations for candidate enumeration and maximization (Section 3.5.3) only need to be pre-computed once, such as the sorting of points [88, 89, 104, 207], and shortest-path trees [156, 196] or fragmentation-multi-graph trees [199] constructed for underlying road networks. The possibility of using only a small set of samples from the entire dataset to construct upper bounds of p-values has been explored. Specifically, a reduction algorithm was developed in [215]. It shows that the original significance-enforced clustering problem can be mapped to a series of reduced problems (e.g., from N to $\lceil \rho N \rceil$ points) where the p-values of reduced clusters in these problems can be used to upper-bound the p-values of clusters in the original problem. Moreover, in scenarios where significance testing needs to be performed simultaneously for a collection of clusters, a dual-convergence algorithm was used to gradually narrow the gap between upper and lower bounds in pruning as the simulation progresses [211, 214].

In addition to Monte Carlo simulation, another line of efforts has explored a variety of mathematical approximations of the distribution of $F(C^*_{H_0}, D_{H_0})$. Similar to many of the approximation algorithms summarized in Section 3.5.3, the approximation schemes here often aim to asymptotically approximate the distribution, i.e., narrowing the error bound as the number of samples gradually grows to infinity. Related work started on one-dimensional data [74, 131] or fixed-cluster-size (e.g., equal-size 1D or 2D windows [38, 39]) due to the complexity of analysis, and was later extended to two- and multi-dimension spaces for several types of test statistics (e.g., maximum normalized scan score) [40, 76, 77]. Strategies have also been explored to select best-performing parameters in the approximation models [39], and many approximations have demonstrated good performance through comparisons with simulated results [39, 76]. As discussed in Section 1, a detailed summary of these mathematical approximations has been provided in [75]. Compared to the Monte Carlo based techniques, accurate approximations of test statistic distributions in closed-forms may greatly reduce the computational overhead. On the other hand, Monte Carlo simulation tends to be more stable (e.g., fewer assumptions) and more flexible (e.g., can be generally applied to hotspot detection problems). Some efforts have also tried to combine Monte Carlo simulation with mathematical approximation as a middle ground between the two types of approaches [4]. Finally, as discussed in Section 3.4.3, p-values are commonly calculated in a scale-independent manner (i.e., the same critical value for all hotspot sizes), but scale-aware methods tend to be more robust [174, 204, 215].

Bayesian Approach: Bayesian based testing schemes were developed to better model uncertainty during hotspot detection by incorporating informative prior information (e.g., expected number of observations in a spatial region, shape regularity, and size) when available. As explained in [139], prior information may also be added into frequentist approaches as constraints, weights, or penalty functions in test statistics [30, 172, 190], but the Bayesian approach offers more flexibility (e.g., priors as continuous distributions). Moreover, a fundamental difference between the two frameworks is that the frequentist approach tests the significance of clusters by checking if H_0 can be rejected, whereas the Bayesian method (e.g., [132]) compares the posterior probabilities of the null and a set of different alternative hypotheses. In the multivariate Bayesian scan statistic [149], for example, each alternative hypothesis H_1 is defined by a pair of a spatial footprint S (i.e., a candidate region) and an event type E (e.g., an increase in pharmacy visits may be caused by a disease outbreak, promotional sales of OTC medicines, and so on.). A posterior probability is calculated for each hypothesis as $P(H_0|D) = \frac{P(D|H_0)\cdot P(H_0)}{P(D)}$ and $P(H_1(S,E)|D) = \frac{P(D|H_1(S,E))\cdot P(H_1(S,E))}{P(D)}$, where D is the observed data, H_0 is the null hypothesis and H_1 is the alternative hypothesis for a region S and event E (there is typically a large number of (S,E) pairs).

Informative prior distributions and their parameters over space are the most critical components to calculate the large number of posteriors in Bayesian scan statistics [132, 140, 148, 149]. Uninformative priors may reduce the Bayesian approach to certain versions of frequentist approaches such as the spatial scan statistic where candidates are ranked by their likelihood ratios [139] (many other types of test statistic functions exist as summarized in Section 3.4.2).

Original forms of Bayesian scan statistics methods [140, 149] estimate the priors by directly learning their distributions (e.g., a prior conditional probability of an event occurring in each region) from past data within a relevant temporal window. Due to the large number of candidate regions (Section 3.5.1), this nonparametric learning requires a large amount of training data to work. To mitigate this issue, some of the prior distributions are further abstracted to parametric distributions that are governed by a small number of parameters [119], which can be learned well with a smaller amount of training data. The downside is that the training is done through a computationally expensive expectation-maximization process [139]. In addition, the Bayesian approach normalizes the individual posteriors using the overall probability of data. This is again costly as $P(D) = P(D|H_0)P(H_0) + \sum_{S,E} P(D|H_1(S,E))P(H_1(S,E))$, which means all combinations of candidate regions and events need to be enumerated to calculate the summation. As a result, this method is constrained to simple regular shapes (e.g., circles, rectangles).

A remedy is the fast subset sums approach [143, 179], which is an alternative to the above region-based Bayesian scan statistics. This method focuses more on individual locations and subsets of locations rather than contiguous regions. It was found in [136] that the computationally-expensive terms in estimating the final posterior probability for each individual location *loc*, such as the sums of posteriors of all location-subsets that contain *loc*, can be exactly computed without enumerating through all the subsets. The result becomes a map of posterior probabilities on individual locations. To improve the spatial proximity of locations with high posterior probabilities, a generalized fast subset sums method was developed [179], in which the proximity and regularity can be controlled by a new sparsity parameter. Another computational enhancement used in Bayesian approaches is the construction of a conjugate prior [132]; otherwise expensive simulations such as **Markov Chain Monte Carlo** (**MCMC**) are needed to calculate the posterior.

The randomization test summarized in the frequentist approach is not used here for significance testing. Instead, the posterior probabilities (e.g., a regular-shaped cluster with highest posterior for

H_1 or a map of posterior probabilities) are directly given to users for decision making. To enforce a false positive rate more explicitly, historical and synthetic data were used in [136, 179] to select a proper threshold on the posterior probabilities or its sums, where the number of false positives under different thresholds can be empirically evaluated.

Many variations have also been developed. A Bayesian network model was developed to combine the spatial-based multivariate Bayesian scan statistic [140] and the population-based PANDA [43] to better capture details in individual data with spatial consideration [95]. A temporal extension was later proposed [94], where the increase of density is modeled using a linear function. A rank-based spatial clustering algorithm uses the Bayesian scan statistic [168, 169] to compute posterior probabilities on individual locations, which were then used to search for clusters. A dedicated Bernoulli version of the Bayesian scan statistic was developed to improve performance on binary labeled points (e.g., a crime case or not, which is not aggregated) [170].

Integrated Bayesian-frequentist based approaches have been explored. A grid-based Bayesian variable window approach first uses Bayesian to define the test statistic and then employs a frequentist based randomization test to perform significance testing [226]. Similarly, methods such as anomalous group detection [50] and fast generalized subset scan [123] use a Bayesian network to evaluate and select the best candidates, and randomization tests to compute critical values.

3.7 Evaluation

One additional benefit of this statistically-robust clustering formulation is that ground truth hotspots are mathematically well-defined using point-process-based hypotheses (Section 3.6). This allows rigorous synthetic data generation, where true hotpots with varying properties can be inserted into point processes to generate validation datasets (e.g., Figures 4 and 5). Such datasets have been widely used to evaluate results from different clustering approaches [6, 60, 87, 107, 129, 134, 172] under a variety of assumptions (e.g., data size, effect size, spatial scale). Typical metrics used in the literature include detection power (i.e., 1-β where β is the Type-II error) [60, 107, 134]; rate of spurious results (i.e., Type-I error) or a total number of spurious clusters detected [149, 215]; precision, recall and F1-scores [207, 213, 215]; and the number of days in advance in detecting an emerging hotspot in the spatio-temporal case [95, 140, 141, 149, 179]. While counting the number of spurious detections in H_0-data is straightforward, measuring if a true hotspot is successfully detected can be tricky. Different strategies have been used to conduct this. For tessellated data (e.g., grid or polygon inputs), the success is often measured by the overlap ratio (i.e., intersection over union—IoU) between the true pattern and detection [95, 179, 191]. This can also be generalized to the IoU between point sets [211]. At the object level, IoU may be less robust. For example, the IoU between two circles decreases quickly even with small differences in center locations and radii [215]. Thus, separate criteria on center accuracy—which may be important in applications (e.g., source of infection or pollution [102]; locations of serial criminals [64])—and size similarity have been used as replacements [207, 215]. To make the validation process more holistic, incompatible assumptions have been used to generate synthetic data and test the sensitivity of techniques (e.g., regular vs. irregular shapes [45, 60, 200], normal vs. non-normal processes for continuous-value based clustering [88, 89, 104]). In addition to quantitative evaluation, qualitative evaluation (e.g., real or simulated case studies, visual inspection) is also commonly employed [60, 63, 156, 162, 196, 211, 214].

In general, many empirical results (e.g., [146, 174, 204, 213, 215]) have shown that spatial hotspots are easier to detect with larger data sizes, effect sizes, and spatial footprints; fewer number of true hotspots; less uncertainty (e.g., due to aggregation or positional inaccuracy); regular shapes, and so on. Related theoretical results have also been developed (e.g., detectability of a clustering signal) [34, 166, 204]. Validation using real-world datasets (e.g., crime cases, traffic accidents, COVID-19

Table 5. Examples of Open Datasets for Evaluation

Category	Examples of open datasets
Simulated data with ground truth	Northeastern US benchmark [2, 25–27, 55, 107, 186]; New York City disease outbreaks benchmark [108, 200]; arbitrarily-shaped (with customizable generation code) [211, 214], and so on.
Real data with known ground truth	Battle of the water sensor networks [25–27, 37, 158]; New York Bronx Legionnaire's disease outbreak [63]; Disease infection at airports (social network data) [187], and so on.
Real data without known ground truth*	New York cancer incidence [5, 24, 79], New York birth defects [5, 185], Sudden Infant Death Syndrome (SIDS) occurrences in North Carolina [47, 127]; Metro Transit bus trajectories with engine measurements [115]; Los Angeles highway traffic [26, 27]; Orlando pedestrians fatalities [196], and so on.

*The score of the best significant cluster and computational time are often used to compare methods, and domain knowledge is used to identify explanations and surprises.

symptoms) often require deep involvement of domain experts. Many studies have used combinations of real-world base data (e.g., spatial domains such as city boundaries, county maps, and road networks [60, 87, 102, 107, 156, 196, 199]; control data such as population or traffic volume [115, 140, 146]) and synthetic event data (e.g., disease cases) in evaluation. Moreover, to further improve synthetic data's representativeness of real-world phenomenons, studies have gone beyond simple statistical assumptions to generate more realistic distributions (e.g., using imitation learning), such as synthetic population and contact networks [23, 82, 227], human mobility maps [18], trajectories [159], and so on. These strategies for realistic spatial data generation can help better estimate the performance of techniques in real scenarios. Finally, we list examples of open datasets across various domains that can be leveraged for evaluation and comparison in Table 5.

4 FUTURE OPPORTUNITIES

There are many opportunities lying ahead to: (1) further enhance existing hotspot detection techniques, (2) incorporate statistical rigor into general clustering methods, (3) harness emerging problems and new forms of spatial data, and (4) explore synergistic integration with learning for mutual benefits.

4.1 Improving Solution Quality, Flexibility, and Efficiency

As discussed in Section 3.4.2, the design of a test statistic is a critical component that directly determines the solution quality. Moreover, the design is non-trivial as localized test statistics (e.g., likelihood ratios [64, 102, 145], density ratios [156, 184, 196, 199], and entropy measures [207]) can be easily biased when being used to compare and rank candidates in hotspot detection, especially across scales [34, 172, 174, 204]. While approaches have been developed to correct the bias and restore the optimal detection power (e.g., [34, 172, 174, 204]), most of them require sophisticated analysis and have only been explored in special cases of hotspot mapping (e.g., Gaussian white noises, one-dimensional intervals, rectangular regions). This limits the methods' adoption in general scenarios summarized in Section 3, despite their importance. An interesting direction would be to explicitly explore the tradeoff between flexibility and optimality, and develop more generally applicable frameworks that may not necessarily fully restore the optimal power. In addition, recent methods that try to correct the bias by incorporating spatial-nondeterminism-awareness are less dependent on the specification of problems [207, 213, 215], but they do introduce additional computational complexity and require more efficient algorithms and data structures. Scalability-wise, existing algorithmic building blocks (Section 3.5.3) have not yet considered applications in the context of spatial big data [67, 165, 221]. Enabling frameworks of large-scale applications need to be developed to model additional complexity (e.g., I/O, network communication, indexing) in distributed environments (e.g., Apache Sedona [221]).

4.2 More Holistic Modeling

The tradeoff between detection power and robustness against spurious patterns has not been sufficiently explored. While statistical rigor is certainly important to suppress false alarms—especially in real-world applications where resource is limited, missing one true hotspot in many cases may turn out to be more costly in certain scenarios (e.g., the spread of COVID-19). Thus, to better inform decision making, an integration of the two indicators may be needed [140, 215]. This integration can be non-trivial as the relative weights of detection power and false positive rate may be nonstationary during different stages of an event (e.g., disease spread) and can be affected by additional dynamic factors (e.g., cost of remedy, impact, policies, forecasts, geographical heterogeneity, local demographics). Consideration of these complex factors in hotspot recommendation may require measures beyond a single test statistic and simple criteria based on significance levels. These new modeling will also require new computational frameworks.

Moreover, the vast majority of techniques use the simplifying assumption that there is a single hotspot in the input data for H_1 [112, 139]. Although this does not necessarily limit the use of the approaches in detecting arbitrary number of clusters [4, 64, 207, 215] (e.g., by performing rounds of detection until the best candidate is no longer significant, as discussed in Section 3.6.1), it may cause the methods to work sub-optimally due to shadow-effects between multiple hotspots [225]. For example, existence of other hotspots increases the baseline score outside one true hotspot, making it harder to confirm its significance. While several methods have been developed to mitigate this issue [112, 225], they are again limited to certain special cases and have not been widely adopted (e.g., not explored for Bayesian modeling [139]). There are opportunities to more broadly consider this issue in hotspot detection and develop general frameworks to address it.

The uncertainty in input datasets also has not been well modeled in general except a few efforts (e.g., [25, 90, 155]). Given that existing datasets in many problems (e.g., disease surveillance) are often provided at an aggregated level (e.g., by county or zip code) due to privacy concerns, such positional uncertainty of individual points can substantially affect solution quality [52, 117, 120]. Explicit techniques (e.g., Dempster– Shafer based modeling) are needed in future research to improve the robustness of the techniques in face of uncertainty.

4.3 Expanding to Broad Problems and Emerging Spatial Data

A major advantage of the wide variety of techniques summarized in this article is their statistical robustness, which allows explicit control of the rate of spurious detections in the output. This modeling has also been used recently to improve the statistical robustness and automate parameter selection of general clustering techniques. Significant DBSCAN [211, 214], for example, can effectively filter out a large number of spurious clusters generated by DBSCAN in data containing heavy-noise and clusters of varying densities. The incorporation of statistical modeling can also greatly reduce efforts in parameter selection. Similarly, a significant spatial data partitioning approach based on non-parallel split lines has been developed [72]. However, in general, statistically robust formulations have not been explored for most general clustering and data partitioning approaches (e.g., HDBSCAN [29, 152], spectral clustering [202], deep clustering [125], and many others [217]). Many opportunities lie ahead to incorporate statistical rigor into these broad techniques.

In addition, there are many newly emerging spatial data types and problems where hotspot detection techniques may be valuable. As an example, location-enriched social network data (e.g., geo-tagged tweets) open up opportunities for monitoring disease spread and many other events (e.g., disasters) using novel formulations of hotspot detection [187, 188]. In the context of COVID-19, business points-of-interest data and their visits (e.g., SafeGraph [3]) can also be leveraged in hotspot detection to generate timely warning of high-risk regions and inform policy making. The use of these new types of datasets in hotspot detection is also challenging due to well-known issues including data quality (e.g., positional accuracy and uncertainty [52, 117]), non-stationarity (e.g., algorithm dynamics, which are considered as major factors that cause erroneous predictions in Google Flu Trends in 2013 [110]) and representativeness (e.g., selection bias [100]).

Many new types of data are also available in the network space. First, traditional GPS trajectories are currently underutilized in hotspot detection despite their potential importance (e.g., transportation safety, environment, smart cities, and public health) [15]. Second, attribute-enriched trajectories, including on-board-diagnostic data (i.e., vehicle trajectories with hundreds of real-time engine measurements [113, 114]) and connected vehicle data [12], also provide lots of new possibilities in novel hotspot detection formulations (e.g., routes with abnormally high emission or energy consumption [115]). Also, many events in urban areas happen and propagate through road networks rather than the Euclidean space, so current research that are mostly focused on the Euclidean space may be expanded to the network space to take the full advantage of new high-resolution datasets. Advancements in these directions may require new statistical formulations (e.g., physics-aware point processes) and computational frameworks.

New opportunities also exist to go beyond existing hotspot formulations to address emerging societal problems. For example, the lack of diversity in many domains has raised major concerns in recent years on the resilience of our communities (e.g., low crop diversity in agriculture poses high risk to global food security [164]). Research landscape in spatial hotspot detection can be broadened to help address these critical issues faced by our society [207].

4.4 Synergistic Integration with Learning

One fundamental property of spatial data is spatial heterogeneity, i.e., data distributions are non-stationary over space, violating the common identical distribution assumption under machine learning methods. Recently, incorporation of statistically-robust formulations (e.g., multivariate scan statistics) in deep learning models has shown promising results in tackling spatial data heterogeneity and substantially improving model performances [210, 212].

On the other hand, statistically robust clustering techniques can also leverage advances from related communities to improve its capacity. For example, the growing maturity of deep learning [111], especially in computer vision and natural language processing, may provide a useful way to further enrich the variety of datasets that can be used for hotspot detection. While it can be risky to directly use deep learning to find hotspots due to concerns such as non-interpretability and overfitting issues [31], these models can help preprocess large volumes of visual and text information from remote sensing data (satellite or aerial imagery, LiDAR point clouds) [93, 208], social platforms (e.g., Twitter, Flickr), traffic cameras, connected vehicles, and so on, to generate valuable new data sources for hotspot detection, which would otherwise be tedious, time-consuming and not scalable. Moreover, learning approaches (e.g., generative adversarial networks) may be leveraged to generate more realistic simulation data under various scenarios to improve evaluation [18, 159, 227].

From a computational perspective, data-driven approaches may also be used to reduce computation load and speed up the enumeration process [118] in scenarios where computing resource is limited on the application-side or real-time performance (e.g., interaction) is needed. Machine learning methods, for example, may be used to heuristically prune low-potential enumeration spaces or recommend a list of high-potential regions for the search to prioritize [17, 223]. Other settings such as reinforcement learning can also be helpful in this scenario.

To reduce the need of training data (e.g., in Bayesian hotspot detection techniques [140, 149]), novel transfer learning methods [195] can be explored to reuse relevant parameters based on similarity between domain distributions across spatial locations, time periods, data, or problems.

5 CONCLUSIONS

In this article, we provided an overview of statistically robust clustering techniques for spatial hotspot mapping. Given the wide application of hotspot mapping techniques in societal applications and the potential importance (e.g., infectious disease surveillance for COVID-19 scenarios), the goal is to present a taxonomy of the vast literature in different key sub-fields (e.g., statistical modeling, maximization algorithms, significance testing) to trigger more thoughts from the broad computing communities.[1] On the other hand, the statistically robust formulations may provide new paths for improving general clustering (e.g., removing spurious clusters) and learning techniques (e.g., heterogeneity); similarly, the validation strategies described (e.g., statistical definition based ground-truth generation) can be potentially used to expand existing evaluation methods for general clustering. To serve these purposes, we selected the terms *clustering* and *hotspot* mainly due to their common use and adoption in the computing community.[2]

We hope that the taxonomies and summaries of techniques provided in this article can serve as a stepping stone for generating new ideas and approaches (e.g., future directions discussed in Section 4) in computing research, and help practitioners select appropriate paradigms based on their data and application needs.

REFERENCES

[1] 2017. National Cancer Institute, Surveillance Research Program. Retrieved 30 October, 2021 from https://surveillance.cancer.gov//.

[2] 2021. Northeastern US benchmark. Retrieved 30 October, 2021 from https://www.satscan.org/datasets/nebenchmark/index.html.

[3] 2021. SafeGraph. Retrieved 30 October, 2021 from https://www.safegraph.com/.

[4] 2021. SaTScan. Retrieved 30 October, 2021 from https://www.satscan.org/.

[5] 2021. SaTScan datasets. Retrieved 30 October, 2021 from https://www.satscan.org/datasets.html.

[6] Geir Aamodt, Sven O. Samuelsen, and Anders Skrondal. 2006. A simulation study of three methods for detecting disease clusters. *International Journal of Health Geographics* 5, 1 (2006), 1–11.

[7] Deepak Agarwal, Andrew McGregor, Jeff M. Phillips, Suresh Venkatasubramanian, and Zhengyuan Zhu. 2006. Spatial scan statistics: Approximations and performance study. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 24–33.

[8] Deepak Agarwal, Jeff M. Phillips, and Suresh Venkatasubramanian. 2006. The hunting of the bump: On maximizing statistical discrepancy. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm*. 1137–1146.

[9] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. 1998. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*. 94–105.

[10] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery* 29, 3 (2015), 626–688.

[11] Cem Aksoylar, Lorenzo Orecchia, and Venkatesh Saligrama. 2017. Connected subgraph detection with mirror descent on SDPs. In *Proceedings of the International Conference on Machine Learning*. PMLR, 51–59.

[12] Reem Y. Ali, Venkata M. V. Gunturi, Shashi Shekhar, Ahmed Eldawy, Mohamed F. Mokbel, Andrew J. Kotz, and William F. Northrop. 2015. Future connected vehicles: Challenges and opportunities for spatio-temporal computing. In *Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 1–4.

[13] Raid Amin, Erich K. Ritter, and Laura Cossette. 2012. A geospatial analysis of shark attack rates for the coast of California: 1994–2010. *Journal of Environment and Ecology* 3, 1 (2012), 246–255.

[1] Short summaries of the key steps are available in the online supplementary appendix for quick references.
[2] Scan statistic has also been widely used in statistics communities and applications.

[14] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: Ordering points to identify the clustering structure. In *Proceedings of the ACM Sigmod Record*, Vol. 28. ACM, 49–60.

[15] R. M. Assunção, R. C. S. N. P. Souza, and M. O. Prates. 2020. New frontiers for scan statistics: Network, trajectory, and text data. In *Handbook of Scan Statistics*. Glaz J. and Koutras M. (Eds.), Springer.

[16] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. 2018. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys* 51, 4 (2018), 83.

[17] Maria-Florina Balcan, Travis Dick, Tuomas Sandholm, and Ellen Vitercik. 2018. Learning to branch. In *Proceedings of the International Conference on Machine Learning*. PMLR, 344–353.

[18] Han Bao, Xun Zhou, Yingxue Zhang, Yanhua Li, and Yiqun Xie. 2020. Covid-gan: Estimating human mobility responses to covid-19 pandemic through spatio-temporal conditional generative adversarial networks. In *Proceedings of the 28th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 273–282.

[19] Brett R. Bayles, Shyam M. Thomas, Gregory S. Simmons, Elizabeth E. Grafton-Cardwell, and Mathew P. Daugherty. 2017. Spatiotemporal dynamics of the Southern California Asian citrus psyllid (Diaphorina citri) invasion. *PLoS One* 12, 3 (2017), e0173226.

[20] Asa Ben-Hur, David Horn, Hava T. Siegelmann, and Vladimir Vapnik. 2001. Support vector clustering. *Journal of Machine Learning Research* 2, Dec (2001), 125–137.

[21] Robert H. Berk and Douglas H. Jones. 1979. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 47, 1 (1979), 47–59.

[22] Pavel Berkhin. 2006. A survey of clustering data mining techniques. In *Grouping Multidimensional Data*. Kogan J., Nicholas C., and Teboulle M. (Eds.), Springer, 25–71.

[23] Keith R. Bisset, Jiangzhuo Chen, Xizhou Feng, V. S. Anil Kumar, and Madhav V. Marathe. 2009. EpiFast: A fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *Proceedings of the 23rd International Conference on Supercomputing*. 430–439.

[24] Francis P. Boscoe, Thomas O. Talbot, and Martin Kulldorff. 2016. Public domain small-area cancer incidence data for New York State, 2005-2009. *Geospatial Health* 11, 1 (2016), 304.

[25] 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.

[26] Jose Cadena, Feng Chen, and Anil Vullikanti. 2017. Near-optimal and practical algorithms for graph scan statistics. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 624–632.

[27] Jose Cadena, Feng Chen, and Anil Vullikanti. 2019. Near-optimal and practical algorithms for graph scan statistics with connectivity constraints. *ACM Transactions on Knowledge Discovery from Data* 13, 2 (2019), 1–33.

[28] Jose Cadena, David Falcone, Achla Marathe, and Anil Vullikanti. 2019. Discovery of under immunized spatial clusters using network scan statistics. *BMC Medical Informatics and Decision Making* 19, 1 (2019), 1–14.

[29] Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. 2015. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data* 10, 1 (2015), 1–51.

[30] André L. F. Cançado, Anderson R. Duarte, Luiz H. Duczmal, Sabino J. Ferreira, Carlos M. Fonseca, and Eliane C. D. M. Gontijo. 2010. Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. *International Journal of Health Geographics* 9, 1 (2010), 1–17.

[31] Davide Castelvecchi. 2016. Can we open the black box of AI? *Nature News* 538, 7623 (2016), 20.

[32] Alfred Castro-Ginard, C. Jordi, Xavier Luri, Francesc Julbe, Mario Morvan, L. Balaguer-Núñez, and Tristan Cantat-Gaudin. 2018. A new method for unveiling open clusters in Gaia-New nearby open clusters confirmed by DR2. *Astronomy & Astrophysics* 618 (2018), A59. https://www.aanda.org/articles/aa/full_html/2018/10/aa33390-18/aa33390-18.html.

[33] Veronique Chadillon-Farinacci, Philippe Apparicio, and Carlo Morselli. 2015. Cannabis cultivation in Quebec: Between space–time hotspots and coldspots. *International Journal of Drug Policy* 26, 3 (2015), 311–322.

[34] Hock Peng Chan and Guenther Walther. 2013. Detection with the scan and the average likelihood ratio. *Statistica Sinica* 23, 1 (2013), 409–428.

[35] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Computing Surveys* 41, 3 (2009), 1–58.

[36] Feng Chen and Daniel B. Neill. 2014. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1166–1175.

[37] Feng Chen and Baojian Zhou. 2016. A generalized matching pursuit approach for graph-structured sparsity. In *International Joint Conference on Artificial Intelligence*. 1389–1395.

[38] Jie Chen and Joseph Glaz. 1996. Two-dimensional discrete scan statistics. *Statistics & Probability Letters* 31, 1 (1996), 59–68.

[39] Jie Chen and Joseph Glaz. 2002. Approximations for a conditional two-dimensional scan statistic. *Statistics & Probability Letters* 58, 3 (2002), 287–296.

[40] Jie Chen and Joseph Glaz. 2009. Approximations for two-dimensional variable window scan statistics. In *Scan Statistics*, Joseph Glaz, Vladimir Pozdnyakov, and Sylvan Wallenstein (Eds.). Springer, 109–128.

[41] Jin Chen, Robert E. Roth, Adam T. Naito, Eugene J. Lengerich, and Alan M. MacEachren. 2008. Geovisual analytics to enhance spatial scan statistic interpretation: An analysis of US cervical cancer mortality. *International Journal of Health Geographics* 7, 1 (2008), 1–18.

[42] Marlize Coleman, Michael Coleman, Aaron M. Mabuza, Gerdalize Kok, Maureen Coetzee, and David N. Durrheim. 2009. Using the SaTScan method to detect local malaria clusters for guiding malaria control programmes. *Malaria Journal* 8, 1 (17 Apr 2009), 68. DOI: https://doi.org/10.1186/1475-2875-8-68

[43] Gregory F. Cooper, John N. Dowling, John D. Levander, et al. 2007. A Bayesian algorithm for detecting CDC Category A outbreak diseases from emergency department chief complaints. *Advances in Disease Surveillance* 2, 2 (2007), 45.

[44] 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. https://www.sciencedirect.com/science/article/pii/S1877584520300332.

[45] Marcelo Azevedo Costa, Renato Martins Assunçao, and Martin Kulldorff. 2012. Constrained spanning tree algorithms for irregularly-shaped spatial clustering. *Computational Statistics & Data Analysis* 56, 6 (2012), 1771–1783.

[46] Marcelo Azevedo Costa and Martin Kulldorff. 2009. Applications of spatial scan statistics: A review. *Scan Statistics: Methods and Applications*, Joseph Glaz, Vladimir Pozdnyakov, and Sylvan Wallenstein (Eds.). Birkhäuser Boston, 129–152. DOI: 10.1007/978-0-8176-4749-0_6

[47] Noel Cressie and Ngai H. Chan. 1989. Spatial modeling of regional variables. *Journal of the American Statistical Association* 84, 406 (1989), 393–401.

[48] Diego F. Cuadros, Anngie Hernandez, Maria F. Torres, Diana M. Torres, Adam J. Branscum, and Diego F. Rincon. 2017. Vector transmission alone fails to explain the potato yellow vein virus epidemic among potato crops in Colombia. *Frontiers in Plant Science* 8 (2017), 1654. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5622202/.

[49] Lionel Cucala. 2016. A Mann–Whitney scan statistic for continuous data. *Communications in Statistics-Theory and Methods* 45, 2 (2016), 321–329.

[50] Kaustav Das, Jeff Schneider, and Daniel B. Neill. 2009. *Detecting Anomalous Groups in Categorical Datasets*. Vol. 12. Carnegie Mellon University, School of Computer Science, Machine Learning.

[51] Daniel P. De Oliveira, Daniel B. Neill, James H. Garrett Jr, and Lucio Soibelman. 2011. Detection of patterns in water distribution pipe breakage using spatial scan statistics for point events in a physical network. *Journal of Computing in Civil Engineering* 25, 1 (2011), 21–30.

[52] Eric Delmelle, Coline Dony, Irene Casas, Meijuan Jia, and Wenwu Tang. 2014. Visualizing the impact of space-time uncertainties on dengue fever patterns. *International Journal of Geographical Information Science* 28, 5 (2014), 1107–1127.

[53] Christophe Demattei, Nicolas Molinari, and Jean-Pierre Daurès. 2007. Arbitrarily shaped multiple spatial cluster detection for case event data. *Computational Statistics & Data Analysis* 51, 8 (2007), 3931–3945.

[54] Achintya N. Dey, Peter Hicks, Stephen Benoit, and Jerome I. Tokars. 2010. Automated monitoring of clusters of falls associated with severe winter weather using the BioSense system. *Injury Prevention* 16, 6 (2010), 403–407.

[55] Raphaella Carvalho Diniz, Pedro O. S. 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.

[56] David Donoho and Jiashun Jin. 2015. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science* 30, 1 (2015), 1–25.

[57] 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.

[58] Luiz Duczmal, André L. F. Cançado, Ricardo H. C. Takahashi, and Lupercio F. Bessegato. 2007. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis* 52, 1 (2007), 43–52.

[59] Luiz Duczmal, Anderson Ribeiro Duarte, and Ricardo Tavares. 2009. Extensions of the scan statistic for the detection and inference of spatial clusters. In *Scan Statistics*, Joseph Glaz, Vladimir Pozdnyakov, and Sylvan Wallenstein (Eds.). Springer, 153–177.

[60] Luiz Duczmal, Martin Kulldorff, and Lan Huang. 2006. Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics* 15, 2 (2006), 428–442.

[61] Lutz Dumbgen and Vladimir G. Spokoiny. 2001. Multiscale testing of qualitative hypotheses. *Annals of Statistics* 29, 1 (2001), 124–152.

[62] Emre Eftelioglu, Yan Li, Xun Tang, Shashi Shekhar, James M. Kang, and Christopher Farah. 2016. Mining network hotspots with holes: A summary of results. In *Annual International Conference on Geographic Information Science*. Springer, 51–67.

[63] Emre Eftelioglu, Shashi Shekhar, James M. Kang, and Christopher C. Farah. 2016. Ring-shaped hotspot detection. *IEEE Transactions on Knowledge and Data Engineering* 28, 12 (2016), 3367–3381.

[64] Emre Eftelioglu, Shashi Shekhar, Dev Oliver, Xun Zhou, Michael R. Evans, Yiqun Xie, James M. Kang, Renee Laubscher, and Christopher Farah. 2014. Ring-shaped hotspot detection: A summary of results. In *Proceedings of the 2014 IEEE International Conference on Data Mining*. IEEE, 815–820.

[65] Emre Eftelioglu, Xun Tang, and Shashi Shekhar. 2015. Geographically robust hotspot detection: A summary of results. In *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop*. IEEE, 1447–1456.

[66] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Vol. 96. 226–231.

[67] Michael R. Evans, Dev Oliver, KwangSoo Yang, Xun Zhou, Reem Y. Ali, and Shashi Shekhar. 2019. Enabling spatial big data via CyberGIS: Challenges and opportunities. In *CyberGIS for Geospatial Discovery and Innovation*. Wang S. and Goodchild M. (Eds.), Springer, 143–170.

[68] Songlin Fei. 2010. Applying hotspot detection methods in forestry: A case study of chestnut oak regeneration. *International Journal of Forestry Research* 2010 (2010). https://www.hindawi.com/journals/ijfr/2010/815292/.

[69] Ronald E. Gangnon and Murray K. Clayton. 2001. A weighted average likelihood ratio test for spatial clustering of disease. *Statistics in Medicine* 20, 19 (2001), 2977–2987.

[70] Ronald E. Gangnon and Murray K. Clayton. 2004. Likelihood-based tests for localized spatial clustering of disease. *Environmetrics: The Official Journal of the International Environmetrics Society* 15, 8 (2004), 797–810.

[71] Peng Gao, Diansheng Guo, Ke Liao, Jennifer J. Webb, and Susan L. Cutter. 2013. Early detection of terrorism outbreaks using prospective space–time scan statistics. *The Professional Geographer* 65, 4 (2013), 676–691.

[72] Jean Gaudart, Belco Poudiougou, Stéphane Ranque, and Ogobara Doumbo. 2005. Oblique decision trees for spatial pattern detection: Optimal algorithm and application to malaria risk. *BMC Medical Research Methodology* 5, 1 (2005), 1–11.

[73] Joseph Glaz and Markos V. Koutras. 2019. *Handbook of Scan Statistics*. Springer.

[74] Joseph Glaz and Joseph I. Naus. 1991. Tight bounds and approximations for scan statistic probabilities for discrete data. *The Annals of Applied Probability* 1, 2 (1991), 306–318.

[75] Joseph Glaz, Vladimir Pozdnyakov, and Sylvan Wallenstein. 2009. *Scan Statistics: Methods and Applications*. Springer.

[76] Joseph Glaz and Zhenkui Zhang. 2004. Multiple window discrete scan statistics. *Journal of Applied Statistics* 31, 8 (2004), 967–980.

[77] Joseph Glaz and Zhenkui Zhang. 2006. Maximum scan score-type statistics. *Statistics & Probability Letters* 76, 13 (2006), 1316–1322.

[78] Sharon K. Greene, Eric R. Peterson, Deborah Kapell, et al. 2016. Daily reportable disease spatiotemporal cluster detection, New York City, New York, USA, 2014–2015. *Emerging Infectious Diseases* 22, 10 (2016), 1808.

[79] George Grekousis. 2021. Local fuzzy geographically weighted clustering: A new method for geodemographic segmentation. *International Journal of Geographical Information Science* 35, 1 (2021), 152–174.

[80] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. 1998. CURE: An efficient clustering algorithm for large databases. *ACM Sigmod Record* 27, 2 (1998), 73–84.

[81] Anjum Hajat, Charlene Hsia, and Marie S. O'Neill. 2015. Socioeconomic disparities and air pollution exposure: A global review. *Current Environmental Health Reports* 2, 4 (2015), 440–450.

[82] Kirk Harland, Alison Heppenstall, Dianna Smith, and Mark H. Birkin. 2012. Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation* 15, 1 (2012). https://www.jasss.org/15/1/1.html.

[83] Marco Helbich. 2012. Beyond postsuburbia? Multifunctional service agglomeration in Vienna's urban fringe. *Tijdschrift Voor Economische en Sociale Geografie* 103, 1 (2012), 39–52.

[84] U. L. F. Hjalmars, Martin Kulldorff, GÖRAN GUSTAFSSON, and Neville Nagarwalla. 1996. Childhood leukaemia in Sweden: Using GIS and a spatial scan statistic for cluster detection. *Statistics in Medicine* 15, 7–9 (1996), 707–715.

[85] Alexander Hohl, Eric M. Delmelle, Michael R. Desjardins, and Yu Lan. 2020. Daily surveillance of COVID-19 using the prospective space-time scan statistic in the United States. *Spatial and Spatio-Temporal Epidemiology* 34 (2020), 100354.

[86] Lan Huang, Martin Kulldorff, and David Gregorio. 2007. A spatial scan statistic for survival data. *Biometrics* 63, 1 (2007), 109–118.

[87] Lan Huang, Linda W. Pickle, and Barnali Das. 2008. Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. *Statistics in Medicine* 27, 25 (2008), 5111–5142.

[88] Lan Huang, Ram C. Tiwari, Linda W. Pickle, and Zhaohui Zou. 2010. Covariate adjusted weighted normal spatial scan statistics with applications to study geographic clustering of obesity and lung cancer mortality in the United States. *Statistics in Medicine* 29, 23 (2010), 2410–2422.

[89] Lan Huang, Ram C. Tiwari, Zhaohui Zou, Martin Kulldorff, and Eric J. Feuer. 2009. Weighted normal spatial scan statistic for heterogeneous population data. *Journal of the American Statistical Association* 104, 487 (2009), 886–898.
[90] Yan Huang and Jason W. Powell. 2012. Detecting regions of disequilibrium in taxi services under uncertainty. In *Proceedings of the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 139–148.
[91] Vandana Pursnani Janeja and Vijayalakshmi Atluri. 2005. Ls 3: A linear semantic scan statistic technique for detecting anomalous windows. In *Proceedings of the 2005 ACM Symposium on Applied Computing*. ACM, 493–497.
[92] Vandana P. Janeja and Vijayalakshmi Atluri. 2008. Random walks to identify anomalous free-form spatial scan windows. *IEEE Transactions on Knowledge and Data Engineering* 20, 10 (2008), 1378–1392.
[93] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. 2017. Incremental dual-memory lstm in land cover prediction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 867–876.
[94] Xia Jiang and Gregory F. Cooper. 2010. A Bayesian spatio-temporal method for disease outbreak detection. *Journal of the American Medical Informatics Association* 17, 4 (2010), 462–471.
[95] Xia Jiang, Daniel B. Neill, and Gregory F. Cooper. 2010. A Bayesian network model for spatial event surveillance. *International Journal of Approximate Reasoning* 51, 2 (2010), 224–239.
[96] Inkyung Jung and Ho Jin Cho. 2015. A nonparametric spatial scan statistic for continuous data. *International Journal of Health Geographics* 14, 1 (2015), 1–6.
[97] Inkyung Jung, Martin Kulldorff, and Otukei John Richard. 2010. A spatial scan statistic for multinomial data. *Statistics in Medicine* 29, 18 (2010), 1910–1918.
[98] Robert J. Kaminski, Eric S. Jefferis, and Chanchalat Chanhatasilpa. 2000. A spatial analysis of American police killed in the line of duty. In *Atlas of Crime: Mapping the Criminal Landscape*, Linda Turnbull, Elaine Hallisey Hendrix, and Borden D. Dent (Eds.). Oryx Press, Phoenix, AZ, 212–220.
[99] George Karypis, Eui-Hong Han, and Vipin Kumar. 1999. Chameleon: Hierarchical clustering using dynamic modeling. *Computer* 32, 8 (1999), 68–75.
[100] Nishant Kishore, Mathew V. Kiang, Kenth Engø-Monsen, Navin Vembar, Andrew Schroeder, Satchit Balsari, and Caroline O. Buckee. 2020. Measuring mobility to monitor travel and physical distancing interventions: A common framework for mobile phone data analysis. *The Lancet Digital Health* 2, 11 (2020), E622–E628.
[101] Julia Krolik, Gerald Evans, Paul Belanger, Allison Maier, Geoffrey Hall, Alan Joyce, Stephanie Guimont, Amanda Pelot, and Anna Majury. 2014. Microbial source tracking and spatial analysis of E. coli contaminated private well waters in southeastern Ontario. *Journal of Water and Health* 12, 2 (2014), 348–357.
[102] Martin Kulldorff. 1997. A spatial scan statistic. *Communications in Statistics-Theory and Methods* 26, 6 (1997), 1481–1496.
[103] Martin Kulldorff. 1999. Spatial scan statistics: Models, calculations, and applications. In *Scan Statistics and Applications.* Glaz J. and Balakrishnan N. (Eds.), Springer, 303–322.
[104] Martin Kulldorff, Lan Huang, and Kevin Konty. 2009. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics* 8, 1 (2009), 58.
[105] Martin Kulldorff, Lan Huang, Linda Pickle, and Luiz Duczmal. 2006. An elliptic spatial scan statistic. *Statistics in Medicine* 25, 22 (2006), 3929–3943.
[106] Martin Kulldorff, Farzad Mostashari, Luiz Duczmal, W. Katherine Yih, Ken Kleinman, and Richard Platt. 2007. Multivariate scan statistics for disease surveillance. *Statistics in Medicine* 26, 8 (2007), 1824–1833.
[107] Martin Kulldorff, Toshiro Tango, and Peter J. Park. 2003. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis* 42, 4 (2003), 665–684.
[108] Martin Kulldorff, Z. Zhang, J. Hartman, R. Heffernan, L. Huang, and F. Mostashari. 2004. Benchmark data and power calculations for evaluating disease outbreak detection methods. *Morbidity and Mortality Weekly Report* 53 (2004), 144–151. https://pubmed.ncbi.nlm.nih.gov/15714644/.
[109] Anaïs Ladoy, Onya Opota, Pierre-Nicolas Carron, Idris Guessous, Séverine Vuilleumier, Stéphane Joost, and Gilbert Greub. 2021. Size and duration of COVID-19 clusters go along with a high SARS-CoV-2 viral load: A spatio-temporal investigation in Vaud state, Switzerland. *Science of The Total Environment* 787 (2021), 147483. https://www.sciencedirect.com/science/article/pii/S0048969721025547.
[110] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: Traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.
[111] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[112] Xiao-Zhou Li, Jin-Feng Wang, Wei-Zhong Yang, Zhong-Jie Li, and Sheng-Jie Lai. 2011. A spatial scan statistic for multiple clusters. *Mathematical Biosciences* 233, 2 (2011), 135–142.
[113] Yan Li, Pratik Kotwal, Pengyue Wang, Yiqun Xie, Shashi Shekhar, and William Northrop. 2020. Physics-guided energy-efficient path selection using on-board diagnostics data. *ACM Transactions on Data Science* 1, 3 (2020), 1–28.

[114] Yan Li, Shashi Shekhar, Pengyue Wang, and William Northrop. 2018. Physics-guided energy-efficient path selection: A summary of results. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems,* 99–108.

[115] Yan Li, Yiqun Xie, Pengyue Wang, Shashi Shekhar, and William Northrop. 2020. Significant lagrangian linear hotspot discovery. In *Proceedings of the 13th ACM SIGSPATIAL International Workshop on Computational Transportation Science.* 1–10.

[116] Jennifer Lord and Shamarial Roberson. 2020. Investigation of geographic disparities of pre-diabetes and diabetes in Florida. *BMC Public Health* 20, 1 (2020), 1–15.

[117] Lan Luo. 2013. Impact of spatial aggregation error on the spatial scan analysis: A case study of colorectal cancer. *Geospatial Health* 8, 1 (2013), 23–35.

[118] Wenjian Luo, Ruikang Yi, Bin Yang, and Peilan Xu. 2018. Surrogate-assisted evolutionary framework for data-driven dynamic optimization. *IEEE Transactions on Emerging Topics in Computational Intelligence* 3, 2 (2018), 137–150.

[119] Maxim Makatchev and Daniel B. Neill. 2008. *Learning Outbreak Regions in Bayesian Spatial Scan Statistics.* Technical Report. Carnegie Mellon University.

[120] Nicholas Malizia. 2013. Inaccuracy, uncertainty and the space-time permutation scan statistic. *PloS One* 8, 2 (2013), e52034.

[121] Michael Matheny, Raghvendra Singh, Liang Zhang, Kaiqiang Wang, and Jeff M. Phillips. 2016. Scalable spatial scan statistics through sampling. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* 1–10.

[122] Michael Matheny, Dong Xie, and Jeff M. Phillips. 2020. Scalable spatial scan statistics for trajectories. *ACM Transactions on Knowledge Discovery from Data* 14, 6 (2020), 1–24.

[123] Edward McFowland, Skyler Speakman, and Daniel B. Neill. 2013. Fast generalized subset scan for anomalous pattern detection. *The Journal of Machine Learning Research* 14, 1 (2013), 1533–1561.

[124] Harvey J. Miller and Jiawei Han. 2009. *Geographic Data Mining and Knowledge Discovery.* CRC press.

[125] Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, and Jun Long. 2018. A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access* 6 (2018), 39501–39514. https://ieeexplore.ieee.org/abstract/document/8412085.

[126] Ruth Minamisava, Simonne S. Nouer, Otaliba L. de Morais Neto, Lícia Kamila Melo, and Ana Lucia S. S. Andrade. 2009. Spatial clusters of violent deaths in a newly urbanized region of Brazil: Highlighting the social disparities. *International Journal of Health Geographics* 8, 1 (2009), 1–10.

[127] Shin-ichi Minato, Jun Kawahara, Fumio Ishioka, et al. 2019. A fast algorithm for combinatorial hotspot mining based on spatial scan statistic. In *Proceedings of the 2019 SIAM International Conference on Data Mining.* 91–99.

[128] Reza Modarres and G. P. Patil. 2007. Hotspot detection with bivariate data. *Journal of Statistical Planning and Inference* 137, 11 (2007), 3643–3654.

[129] Jacklin F. Mosha, Hugh J. W. Sturrock, Brian Greenwood, Colin J. Sutherland, Nahla B. Gadalla, Sharan Atwal, Simon Hemelaar, Joelle M. Brown, Chris Drakeley, Gibson Kibiki, and T. Bousema. 2014. Hot spot or not: A comparison of spatial statistical methods to predict prospective malaria infections. *Malaria Journal* 13, 1 (2014), 53.

[130] Tomoki Nakaya and Keiji Yano. 2010. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS* 14, 3 (2010), 223–239.

[131] Joseph I. Naus. 1966. Some probabilities, expectations and variances for the size of largest clusters and smallest intervals. *Journal of the American Statistical Association* 61, 316 (1966), 1191–1199.

[132] Daniel Neill, Andrew Moore, and Gregory Cooper. 2006. A Bayesian spatial scan statistic. In *Proceedings of the 18th International Conference on Neural Information Processing Systems.* MIT Press, 1003–1010.

[133] Daniel B. Neill. 2006. Detection of spatial and spatio-temporal clusters. In *Tech Rep CMU-CS-06-142, PhD thesis.* Carnegie Mellon University.

[134] Daniel B. Neill. 2009. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics* 8, 1 (2009), 1–16.

[135] Daniel B. Neill. 2009. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting* 25, 3 (2009), 498–517.

[136] Daniel B. Neill. 2011. Fast Bayesian scan statistics for multivariate event detection and visualization. *Statistics in Medicine* 30, 5 (2011), 455–469.

[137] Daniel B. Neill. 2012. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74, 2 (2012), 337–360.

[138] Daniel B. Neill. 2015. *Subset Scanning for Event and Pattern Detection.* Springer, Cham, 1–10. DOI: https://doi.org/10.1007/978-3-319-23519-6_1547-1

[139] Daniel B. Neill. 2018. *Bayesian Scan Statistics.* Springer, New York, NY, 1–21. DOI: https://doi.org/10.1007/978-1-4614-8414-1_28-1

[140] Daniel B. Neill and Gregory F. Cooper. 2010. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning* 79, 3 (2010), 261–282.
[141] Daniel B. Neill, Gregory F. Cooper, Kaustav Das, Xia Jiang, and Jeff Schneider. 2009. Bayesian network scan statistics for multivariate pattern detection. In *Scan Statistics*, Joseph Glaz, Vladimir Pozdnyakov, and Sylvan Wallenstein (Eds.). Springer, 221–249.
[142] Daniel B. Neill and Jeff Lingwall. 2007. A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance* 4, 106 (2007), 570.
[143] Daniel B. Neill and Y. Liu. 2011. Generalized fast subset sums for Bayesian detection and visualization. In *Proceedings of the International Society for Disease Surveillance Conference 2010: Enhancing the Synergy Between Research, Informatics, and Practice in Public Health.* Taylor & Francis.
[144] Daniel B. Neill, Edward McFowland III, and Huanian Zheng. 2013. Fast subset scan for multivariate event detection. *Statistics in Medicine* 32, 13 (2013), 2185–2208.
[145] Daniel B. Neill and Andrew W. Moore. 2004. A fast multi-resolution method for detection of significant spatial disease clusters. In *Proceedings of the Advances in Neural Information Processing Systems.* 651–658.
[146] Daniel B. Neill and Andrew W. Moore. 2004. Rapid detection of significant spatial clusters. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 256–265.
[147] Daniel B. Neill and Andrew W. Moore. 2006. Methods for detecting spatial and spatiotemporal clusters. In *Handbook of Biosurveillance*, M. M. Wagner, A. W. Moore, and R. Aryel (Eds.). Elsevier, Amsterdam, 243–254.
[148] Daniel B. Neill, Andrew W. Moore, and Gregory F. Cooper. 2006. A Bayesian scan statistic for spatial cluster detection. *Advances in Disease Surveillance* 1 (2006), 55. http://faculty.washington.edu/lober/www.isdsjournal.org/htdocs/articles/214.pdf.
[149] Daniel B. Neill, Andrew W. Moore, and Gregory F. Cooper. 2007. A multivariate Bayesian scan statistic. *Advances in Disease Surveillance* 2 (2007), 60. https://faculty.washington.edu/lober/www.isdsjournal.org/htdocs/articles/824.pdf.
[150] Daniel B. Neill, Andrew W. Moore, Francisco Pereira, and Tom M. Mitchell. 2004. Detecting significant multidimensional spatial clusters. In *Advances in Neural Information Processing Systems.* 969–976.
[151] Daniel B. Neill, Andrew W. Moore, Maheshkumar Sabhnani, and Kenny Daniel. 2005. Detection of emerging space-time clusters. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining.* 218–227.
[152] Antonio Cavalcante Araujo Neto, Joerg Sander, Ricardo J. G. B. Campello, and Mario A. Nascimento. 2017. Efficient computation of multiple density-based clustering hierarchies. In *Proceedings of the 2017 IEEE International Conference on Data Mining.* 991–996.
[153] Raymond T. Ng and Jiawei Han. 2002. CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering* 14, 5 (2002), 1003–1016.
[154] Jalil Noroozi, Alireza Naqinezhad, Amir Talebi, Moslem Doostmohammadi, Christoph Plutzar, Sabine B. Rumpf, Zahra Asgarpour, and Gerald M. Schneeweiss. 2019. Hotspots of vascular plant endemism in a global biodiversity hotspot in Southwest Asia suffer from significant conservation gaps. *Biological Conservation* 237 (2019), 299–307. https://www.sciencedirect.com/science/article/pii/S000632071930268X?via%3Dihub.
[155] Fernando L. P. Oliveira, André L. F. Cançado, Luiz H. Duczmal, and Anderson R. Duarte. 2012. Assessing the outline uncertainty of spatial disease clusters. In *Public Health—Methodology, Environmental and Systems Issues*, J. Maddock (Ed.). In *Tech*, 51–66.
[156] Dev Oliver, Shashi Shekhar, James M. Kang, Renee Laubscher, Veronica Carlan, and Abdussalam Bannur. 2013. A k-main routes approach to spatial network activity summarization. *IEEE Transactions on Knowledge and Data Engineering* 26, 6 (2013), 1464–1478.
[157] Carmen Vega Orozco, Marj Tonini, Marco Conedera, and Mikhail Kanveski. 2012. Cluster recognition in spatial-temporal sequences: The case of forest fires. *Geoinformatica* 16, 4 (2012), 653–673.
[158] Avi Ostfeld, James G. Uber, Elad Salomons, Berry J. W., Hart W. E., Phillips C. A., Watson J. P., Dorini G., Jonkergouw P., Kapelan Z., di Pierro F. 2008. The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management* 134, 6 (2008), 556–568.
[159] Menghai Pan, Weixiao Huang, Yanhua Li, Xun Zhou, and Jun Luo. 2020. xGAIL: Explainable generative adversarial imitation learning for explainable human decision analysis. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 1334–1343.
[160] M. Parimala, Daphne Lopez, and N. C. Senthilkumar. 2011. A survey on density based clustering algorithms for mining large spatial databases. *International Journal of Advanced Science and Technology* 31, 1 (2011), 59–66.
[161] G. P. Patil, R. Modarres, W. L. Myers, and P. Patankar. 2006. Spatially constrained clustering and upper level set scan hotspot detection in surveillance geoinformatics. *Environmental and Ecological Statistics* 13, 4 (2006), 365–377.
[162] Ganapati P. Patil and Charles Taillie. 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* 11, 2 (2004), 183–197.

[163] Paola Perchinunno, Francesco D. d'Ovidio, and Francesco Rotondo. 2019. Identification of "Hot Spots" of inner areas in Italy: Scan statistic for urban planning policies. *Social Indicators Research* 143, 3 (2019), 1299–1317.

[164] Dafydd Pilling, Julie Bélanger, and Irene Hoffmann. 2020. Declining biodiversity for food and agriculture needs urgent global action. *Nature Food* 1, 3 (2020), 144–147.

[165] Sushil K. Prasad, Danial Aghajarian, Michael McDermott, Shah D., Mokbel M., Puri S., Rey S. J., Shekhar S., Xe Y., Vatsavai R. R., and Wang F. 2017. Parallel processing over spatial-temporal datasets from geo, bio, climate and social science communities: A research roadmap. In *Proceedings of the 2017 IEEE International Congress on Big Data*. IEEE, 232–250.

[166] Katharina Proksch, Frank Werner, and Axel Munk. 2018. Multiscale scanning in inverse problems. *Annals of Statistics* 46, 6B (2018), 3569–3602.

[167] Jing Qian, Venkatesh Saligrama, and Yuting Chen. 2014. Connected sub-graph detection. In *Proceedings of the Artificial Intelligence and Statistics*. PMLR, 796–804.

[168] Jialan Que and Fu-Chiang Tsui. 2008. A multi-level spatial clustering algorithm for detection of disease outbreaks. In *Proceedings of the AMIA Annual Symposium Proceedings*, Vol. 2008. American Medical Informatics Association, 611.

[169] Jialan Que and Fu-Chiang Tsui. 2011. Rank-based spatial clustering: An algorithm for rapid outbreak detection. *Journal of the American Medical Informatics Association* 18, 3 (2011), 218–224.

[170] Simon Read. 2011. *A Bayesian Approach to the Bernoulli Spatial Scan Statistic*. Technical Report.

[171] Kurt H. Riitters and John W. Coulston. 2005. Hot spots of perforated forest in the eastern United States. *Environmental Management* 35, 4 (2005), 483–492.

[172] Camilo Rivera and Guenther Walther. 2013. Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *Scandinavian Journal of Statistics* 40, 4 (2013), 752–769.

[173] Polina Rozenshtein, Aris Anagnostopoulos, Aristides Gionis, and Nikolaj Tatti. 2014. Event detection in activity networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1176–1185.

[174] Kaspar Rufibach and Guenther Walther. 2010. The block criterion for multiscale inference about a density, with applications to other multiscale problems. *Journal of Computational and Graphical Statistics* 19, 1 (2010), 175–190.

[175] Ritvik Sahajpal, G. V. Ramaraju, and Vibhor Bhatt. 2004. Applying niching genetic algorithms for multiple cluster discovery in spatial analysis. In *Proceedings of the International Conference on Intelligent Sensing and Information Processing*. IEEE, 35–40.

[176] Daniel M. Saman, Henry P. Cole, Melvin L. Myers, Daniel I. Carey, Susan C. Westneat, and Agricola Odoi. 2012. A spatial cluster analysis of tractor overturns in Kentucky from 1960 to 2002. *PLoS One* 7, 1 (2012), e30532.

[177] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* 10, 5 (1998), 1299–1319.

[178] Gorm E. Shackelford, Peter R. Steward, Richard N. German, Tim G. Benton and Steven M. Sait. 2015. Conservation planning in agricultural landscapes: Hotspots of conflict between agriculture and nature. *Diversity and Distributions* 21, 3 (2015), 357–367.

[179] Kan Shao, Yandong Liu, and Daniel B. Neill. 2011. A generalized fast subset sums framework for bayesian event detection. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*. IEEE, 617–625.

[180] James Sharpnack, Alessandro Rinaldo, and Aarti Singh. 2015. Detecting anomalous activity on networks with the graph Fourier scan statistic. *IEEE Transactions on Signal Processing* 64, 2 (2015), 364–379.

[181] James Sharpnack, Aarti Singh, and Alessandro Rinaldo. 2013. Changepoint detection over graphs with the spectral scan statistic. In *Proceedings of the Artificial Intelligence and Statistics*. PMLR, 545–553.

[182] Shashi Shekhar, Zhe Jiang, Reem Ali, Emre Eftelioglu, Xun Tang, Venkata Gunturi, and Xun Zhou. 2015. Spatiotemporal data mining: A computational perspective. *ISPRS International Journal of Geo-Information* 4, 4 (2015), 2306–2338.

[183] Xiaobei Shen and Wei Jiang. 2014. Multivariate normal spatial scan statistic for detecting the most severe cluster of a disease. *Journal of Management Analytics* 1, 2 (2014), 130–145.

[184] Lei Shi and Vandana P. Janeja. 2009. Anomalous window discovery through scan statistics for linear intersecting paths (SSLIP). In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 767–776.

[185] Ivair R. Silva, Luiz Duczmal, and Martin Kulldorff. 2021. Confidence intervals for spatial scan statistic. *Computational Statistics & Data Analysis* 158 (2021), 107185. https://www.sciencedirect.com/science/article/pii/S0167947321000190?casa_token=UDz05b-gsO8AAAAA:RHLZZeV5oFW0y7jQweZ1I7VzWHD-qmzhVeBJ8evfvcNMBeMVCZVeHutBL_sfvOVSm4ClEncY.

[186] Changhong Song and Martin Kulldorff. 2003. Power evaluation of disease clustering tests. *International Journal of Health Geographics* 2, 1 (2003), 1–8.

[187] Roberto C. S. N. P. Souza, Renato M. Assunção, Daniel B. Neill, and Wagner Meira Jr. 2019. Detecting spatial clusters of disease infection risk using sparsely sampled social media mobility patterns. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 359–368.

[188] Roberto C. S. N. P. Souza, Renato M. Assunção, Derick M. Oliveira, Daniel B. Neill, and Wagner Meira Jr. 2019. Where did I get dengue? Detecting spatial clusters of infection risk with social network data. *Spatial and Spatio-Temporal Epidemiology* 29 (2019), 163–175. https://www.sciencedirect.com/science/article/abs/pii/S1877584517301715.

[189] Skyler Speakman, Edward McFowland III, and Daniel B. Neill. 2015. Scalable detection of anomalous patterns with connectivity constraints. *Journal of Computational and Graphical Statistics* 24, 4 (2015), 1014–1033.

[190] Skyler Speakman, Sriram Somanchi, Edward McFowland III, and Daniel B. Neill. 2016. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics* 25, 2 (2016), 382–404.

[191] Skyler Speakman, Yating Zhang, and Daniel B. Neill. 2013. Dynamic pattern detection with temporal consistency and connectivity constraints. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining*. IEEE, 697–706.

[192] B. D. Spindler, S. R. Chipps, R. A. Klumb, and Michael C. Wimberly. 2009. Spatial analysis of pallid sturgeon Scaphirhynchus albus distribution in the Missouri River, South Dakota. *Journal of Applied Ichthyology* 25, suppl. 2 (2009), 8–13. http://www.pallidsturgeon.org/wp-content/uploads/2018/01/K5Spindler-et-al-2009.pdf.

[193] Barbara Szonyi, Indumathi Srinath, Andy Schwartz, Alfonso Clavijo, and Renata Ivanek. 2012. Spatio-temporal epidemiology of Tritrichomonas foetus infection in Texas bulls based on state-wide diagnostic laboratory data. *Veterinary Parasitology* 186, 3–4 (2012), 450–455.

[194] Kunihiko Takahashi, Martin Kulldorff, Toshiro Tango, and Katherine Yih. 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics* 7, 1 (2008), 1–14.

[195] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A survey on deep transfer learning. In *Proceedings of the International Conference on Artificial Neural Networks*. Springer, 270–279.

[196] Xun Tang, Emre Eftelioglu, Dev Oliver, and Shashi Shekhar. 2017. Significant linear hotspot discovery. *IEEE Transactions on Big Data* 3, 2 (2017), 140–153.

[197] Xun Tang, Emre Eftelioglu, and Shashi Shekhar. 2015. Elliptical hotspot detection: A summary of results. In *Proceedings of the 4th International ACM SIGSPATIAL Workshop on Analytics for Big Geospatial Data*. 15–24.

[198] Xun Tang, Emre Eftelioglu, and Shashi Shekhar. 2017. Detecting isodistance hotspots on spatial networks: A summary of results. In *Proceedings of the International Symposium on Spatial and Temporal Databases*. Springer, 281–299.

[199] Xun Tang, Jayant Gupta, and Shashi Shekhar. 2019. Linear hotspot discovery on all simple paths: A summary of results. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 476–479.

[200] Toshiro Tango and Kunihiko Takahashi. 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 4, 1 (2005), 1–15.

[201] Anh Trang, Sanjeev Agarwal, Phillip Regalia, Thomas Broach, and Thomas Smith. 2007. A patterned and unpatterned minefield detection in cluttered environments using Markov marked point process. In *Proceedings of the Detection and Remediation Technologies for Mines and Minelike Targets XII*, Vol. 6553. International Society for Optics and Photonics, 655313.

[202] Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17, 4 (2007), 395–416.

[203] Jonathan Wakefield and Albert Kim. 2013. A Bayesian model for cluster detection. *Biostatistics* 14, 4 (2013), 752–765.

[204] Guenther Walther. 2010. Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statistics* 38, 2 (2010), 1010–1033.

[205] Wei Wang, Jiong Yang, and Richard Muntz. 1997. STING: A statistical information grid approach to spatial data mining. In *Proceedings of the International Conference on VLDB*, Vol. 97. 186–195.

[206] Craig R. Warden. 2008. Comparison of Poisson and Bernoulli spatial cluster analyses of pediatric injuries in a fire district. *International Journal of Health Geographics* 7, 1 (2008), 1–17.

[207] Yiqun Xie, Han Bao, Yan Li, and Shashi Shekhar. 2020. Discovering spatial mixture patterns of interest. In *Proceedings of the 28th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 608–617.

[208] Yiqun Xie, Han Bao, Shashi Shekhar, and Joseph Knight. 2018. A TIMBER framework for mining urban tree inventories using remote sensing datasets. In *Proceedings of the 2018 IEEE International Conference on Data Mining*. IEEE, 1344–1349.

[209] Yiqun Xie, Emre Eftelioglu, Reem Ali, Xun Tang, Yan Li, Ruhi Doshi, and Shashi Shekhar. 2017. Transdisciplinary foundations of geospatial data science. *ISPRS International Journal of Geo-Information* 6, 12 (2017), 395.

[210] Yiqun Xie, Erhu He, Xiaowei Jia, Han Bao, Xun Zhou, Rahul Ghosh, and Praveen Ravirathinam. 2021. A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity. In *Proceedings of the IEEE International Conference on Data Mining*.

[211] Yiqun Xie, Xiaowei Jia, Han Bao, Xun Zhou, and Shashi Shekhar. 2021. Significant DBSCAN+: Statistically robust density-based clustering. *ACM Transactions on Intelligent Systems and Technology* 12, 5 (2021).

[212] Yiqun Xie, Xiaowei Jia, Han Bao, Xun Zhou, Jia Yu, Rahul Ghosh, and Praveen Ravirathinam. 2021. A self-adaptive and model-agnostic deep learning framework for spatially heterogeneous datasets. In *Proceedings of the ACM SIGSPATIAL International Conference on Advancements in Geographic Information Systems*.

[213] Yiqun Xie and Shashi Shekhar. 2019. A nondeterministic normalization based scan statistic (NN-scan) towards robust hotspot detection: A summary of results. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM.
[214] 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.
[215] Yiqun Xie and Shashi Shekhar. 2020. A unified framework for robust and efficient hotspot detection in smart cities. *ACM Transactions on Data Science* 1, 3 (2020), 1–29.
[216] Yiqun Xie, Xun Zhou, and Shashi Shekhar. 2021. Discovering interesting sub-paths with statistical significance from spatio-temporal datasets. *ACM Transactions on Intelligent Systems and Technology* 11, 1 (2021).
[217] Dongkuan Xu and Yingjie Tian. 2015. A comprehensive survey of clustering algorithms. *Annals of Data Science* 2, 2 (2015), 165–193.
[218] Rui Xu and Donald Wunsch. 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks* 16, 3 (2005), 645–678.
[219] Xin Xu, Yuan Zhao, Siyou Xia, and Xinlin Zhang. 2019. Investigation of multi-scale spatio-temporal pattern of oldest-old clusters in China on the basis of spatial scan statistics. *PloS One* 14, 7 (2019), e0219695.
[220] Masatoshi Yoshida, Yuji Naya, and Yasushi Miyashita. 2003. Anatomical organization of forward fiber projections from area TE to perirhinal neurons representing visual long-term memory in monkeys. *Proceedings of the National Academy of Sciences* 100, 7 (2003), 4257–4262.
[221] Jia Yu, Zongsi Zhang, and Mohamed Sarwat. 2019. Spatial data management in apache spark: The geospark perspective and beyond. *GeoInformatica* 23, 1 (2019), 37–78.
[222] April M. Zeoli, Jesenia M. Pizarro, Sue C. Grady, and Christopher Melde. 2014. Homicide as infectious disease: Using public health methods to investigate the diffusion of homicide. *Justice Quarterly* 31, 3 (2014), 609–632.
[223] Jianyuan Zhai and Fani Boukouvala. 2019. Data-driven spatial branch-and-bound algorithms for black-box optimization. In *Proceedings of the Computer Aided Chemical Engineering*. Vol. 47. Elsevier, 71–76.
[224] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Record* 25, 2 (1996), 103–114.
[225] Zhenkui Zhang, Renato Assunção, and Martin Kulldorff. 2010. Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics* 2010 (2010). https://www.hindawi.com/journals/jps/2010/642379/.
[226] Zhenkui Zhang and Joseph Glaz. 2008. Bayesian variable window scan statistics. *Journal of Statistical Planning and Inference* 138, 11 (2008), 3561–3567.
[227] Liang Zhao, Jiangzhuo Chen, Feng Chen, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2015. Simnest: Social media nested epidemic simulation via online semi-supervised deep learning. In *Proceedings of the 2015 IEEE International Conference on Data Mining*. 639–648.
[228] Baojian Zhou, Feng Chen, and Yiming Ying. 2019. Stochastic iterative hard thresholding for graph-structured sparsity optimization. In *Proceedings of the International Conference on Machine Learning*. PMLR, 7563–7573.

Received March 2021; revised August 2021; accepted September 2021