Annals of the American Association of Geographers

ISSN: 2469-4452 (Print) 2469-4460 (Online) Journal homepage: https://tandfonline.com/loi/raag21

ESCIP: An Expansion-Based Spatial Clustering Method for Inhomogeneous Point Processes

Ting Li, Yizhao Gao & Shaowen Wang

To cite this article: Ting Li, Yizhao Gao & Shaowen Wang (2019): ESCIP: An Expansion-Based Spatial Clustering Method for Inhomogeneous Point Processes, Annals of the American Association of Geographers, DOI: 10.1080/24694452.2019.1625747

To link to this article: https://doi.org/10.1080/24694452.2019.1625747

Published online: 31 Jul 2019.

ESCIP: An Expansion-Based Spatial Clustering Method for Inhomogeneous Point Processes

Ting Li,[*,†] Yizhao Gao,[*,†] and Shaowen Wang[*,†,‡,§,¶]

[*]*Department of Geography and Geographic Information Science, University of Illinois*
[†]*CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois*
[‡]*Department of Urban and Regional Planning, University of Illinois*
[§]*Department of Computer Science, University of Illinois*
[¶]*School of Information Sciences, University of Illinois*

Detecting irregularly shaped spatial clusters within heterogeneous point processes is challenging because the number of potential clusters with different sizes and shapes can be enormous. This research develops a novel method, expansion-based spatial clustering for inhomogeneous point processes (ESCIP), for detecting spatial clusters of any shape within a heterogeneous point process in the context of analyzing spatial big data. Statistical testing is used to find core points—points with neighboring areas that have significantly more cases than the expectation—and an expansion approach is developed to find irregularly shaped clusters by connecting nearby core points. Instead of employing a brute-force search for all potential clusters, as done in the spatial scan statistics, this approach only requires testing a small neighboring area for each potential core point. Moreover, spatial indexing is leveraged to speed up the search for nearby points and the expansion of clusters. The proposed method is implemented with Poisson and Bernoulli models and evaluated for large spatial data sets. Experimental results show that ESCIP can detect irregularly shaped spatial clusters from millions of points with high efficiency. It is also demonstrated that the method outperforms the spatial scan statistics on the flexibility of cluster shapes and computational performance. Furthermore, ESCIP ensures that every subset of a detected cluster is statistically significant and contiguous. *Key Words: cyberGIS, spatial algorithm, spatial analysis, spatial clustering.*

在异质的点过程中侦测不规则形成的空间和集群具有挑战，因为潜在的大小和形状各异的集群可能为数众多。本研究为不同质的点过程（ESCIP）发展一个崭新方法，根据扩张的空间集群，以在分析空间大数据脉络中的异质点过程中侦测任何形状的空间集群。统计检定用来寻找核心点——邻近区域较预期显着具有更多案例的点——并发展一个扩张方法，通过连结附近的核心点，发掘不规则形塑的集群。不同于像空间扫描统计一般运用暴力法搜寻所有潜在的集群，此一法仅需检验每个潜在核心点的小范围邻近面积。此外，空间指标发挥槓杆作用来加速搜寻邻近点和集群的扩张。本文提出的方法同时运用卜瓦松和伯努利模型，并对大型空间数据集进行评估。实验结果显示ESCIP能够以高效能侦测数百万点中不规则形塑的空间集群。实验结果亦证实，该方法在集群形式的弹性和演算表现上，较空间扫描统计表现更好。此外，ESCIP确保每一个侦测到的集群子集在统计上是显着且连续的。*关键词：网络地理信息系统，空间演算，空间分析，空间集群。*

Detectar aglomeraciones espaciales configuradas irregularmente dentro de procesos puntuales heterogéneos representa todo un reto debido a que el número de aglomeraciones potenciales de diferentes tamaños y formas puede ser enorme. Esta investigación desarrolla un nuevo método, la aglomeración espacial basada en la expansión para procesos puntuales no homogéneos (ESCIP) para detectar aglomeraciones espaciales de cualquier forma dentro de un proceso puntual heterogéneo en el contexto de análisis de big data espaciales. Se usa prueba estadística para hallar puntos medulares—puntos con áreas vecinas que significativamente tienen más casos de los esperados—y se desarrolla un enfoque de expansión para hallar aglomeraciones conformadas irregularmente, conectando puntos medulares cercanos. En vez de utilizar una búsqueda de fuerza bruta para todas las aglomeraciones potenciales, como se hace en las estadísticas espaciales de escáner, este enfoque solo requiere poner a prueba una pequeña área vecina por cada punto medular potencial. Aún más, se apalanca la indexación espacial para acelerar la búsqueda de puntos cercanos y la expansión de las

Annals of the American Association of Geographers, 0(0) 2019, pp. 1–18
Initial submission, January 2018; revised submissions, June 2018 and January 2019; final acceptance, March 2019
Color versions of one or more figures in the article can be found online at www.tandfonline.com/raag
Published by Taylor & Francis, LLC.

aglomeraciones. El método propuesto se implementó con modelos Poisson y Bernoulli y se evalúa para conjuntos de datos espaciales grandes. Los resultados experimentales muestran que la ESCIP pueden detectar aglomeraciones espaciales de configuración irregular desde millones de puntos, con alto grado de eficiencia. Se demuestra también que el método supera en desempeño a la estadística espacial de escáner en lo que concierne a flexibilidad de formas del aglomerado y desempeño computacional. Todavía más, ESCIP asegura que cada subconjunto de una aglomeración detectada es estadísticamente significativo y contiguo. *Palabras clave: aglomeración espacial, algoritmo espacial, análisis espacial, cíberSIG.*

Spatial clustering is often used to test whether events are randomly distributed over space or any local excesses can be detected and to test whether such excesses can be reasonably assumed to have occurred by chance (Kulldorff 1999). It has been widely used in many research domains, including spatial epidemiology (Kulldorff and Nagarwalla 1995; Gatrell et al. 1996), crime studies (Eck et al. 2005; Nakaya and Yano 2010), event detection (Cheng and Wicks 2014), astrophysics (Tramacere and Vecchio 2013), movement analysis (Gao, Li, et al. 2018), ecology (Plotkin, Chave, and Ashton 2002), and spatial analysis (Anselin 1995; Rogerson and Yamada 2008). In these areas, it is important not only to discover clusters but also to find the "meaningful" clusters that are not likely to have occurred by chance (Neill and Moore 2004). For example, a large number of disease cases in a region might be caused by certain driving factors like pollutants or simply a large population. Thus, population density, or other underlying intensities that govern the distribution of events if no clusters exist in the study area (null hypothesis), should be considered for cluster detection.

Clustering point-based events has long been a focus of spatial clustering analysis. Many geographical phenomena can be modeled as point-based events, such as disease cases, retail stores, species locations, and people's space–time footprints. Currently, a huge amount of point-based spatial big data, such as Global Positioning System (GPS) tracking records, location-based social media, and volunteered geographic information (Elwood 2008; Flanagin and Metzger 2008), are becoming increasingly available due to technological advances (Kwan 2016). For instance, New York City has been publishing taxi data containing the pickup and drop-off locations for billions of individual trips since 2009. These large point-based data sources offer opportunities to understand fine-scale spatial patterns and potentially the space–time dynamics of the underlying phenomena. At the same time, the computational and data intensities of cluster detection challenge the capability of existing methods and algorithms.

Much effort has been expended for detecting spatial clusters of arbitrary shapes and varying densities in point data sets during the past several decades (Ertöz, Steinbach, and Kumar 2003; Kriegel et al. 2011), including DBSCAN (Ester et al. 1996), DENCLUE (Hinneburg and Keim 1998), OPTICS (Ankerst et al. 1999), DECODE (Pei et al. 2009), CURE (Guha, Rastogi, and Shim 1998), CHAMELEON (Karypis, Han, and Kumar 1999), and MST-DBSCAN (Kuo, Wen, and Sabel 2018). These methods, however, can only identify areas where events are closely packed together but cannot distinguish between clusters that are statistically significant and those that are likely to have occurred by chance in a heterogeneous point process. When the intensity of the underlying point process has a known inhomogeneity under the null hypothesis of no clusters, detecting spatial clusters can be challenging because of the lack of efficient methods to find irregularly shaped clusters that are statistically significant. The spatial scan statistics (Naus 1965; Kulldorff 1997) are commonly used for cluster detection in heterogeneous point processes. They have limited capability, though, for dealing with large data sets and detecting irregularly shaped clusters. There are approaches to identifying irregularly shaped spatial clusters by connecting neighboring regions, using genetic algorithms or other optimization methods (Duczmal and Assunção 2004; Tango and Takahashi 2005; Aldstadt and Getis 2006; Duczmal et al. 2007; Pei et al. 2011; Izakian and Pedrycz 2012; Murray, Grubesic, and Wei 2014; Yin and Mu 2017). These approaches assume, however, that input data are aggregated into a small set of polygons and detect clusters by connecting nearby polygons. Therefore, these methods are not useful for data-intensive point clustering analyses.

This article describes a novel spatial clustering method, expansion-based spatial clustering for inhomogeneous point process (ESCIP), for the following interrelated purposes: first, to detect spatial clusters in a point process with a known heterogeneous intensity; second, to detect clusters with irregular shapes; and third, to handle large spatial data sets with high computational efficiency. The method first identifies points with significant local excess of events as core points and then uses an expansion approach to find irregularly shaped clusters by connecting nearby core points. Instead of a brute-force search for all potential clusters, as done by Kulldorff's (1997) spatial scan statistics, this approach only checks a limited number of search windows to find core points for expansion. Spatial heterogeneity is treated in this method by using an inhomogeneous background to generate the expected spatial distribution of events under the null hypothesis of no clustering. Employing two popular baseline processes (Poisson and Bernoulli), statistical testing is used to find significant core points by taking the background intensity into consideration, rather than purely based on the absolute density of events. Spatial indexing is also leveraged to speed up the search for nearby points and the expansion of clusters.

ESCIP was applied to simulated geolocated households in Chittagong Division, Bangladesh. Experimental results showed that this method can effectively detect irregularly shaped spatial clusters over an inhomogeneous background in large data sets. Compared with spatial scan statistics, this method can detect clusters with more flexible shapes and higher computational performance. Moreover, ESCIP ensures that each detected cluster is self-contiguous and every subset of the cluster is statistically significant.

Related Work

Spatial point clustering has long been an important task in spatial analysis and spatial data mining (Han, Kamber, and Tung 2001). In geographic space, a spatial cluster is usually defined as an excess of events or values in an area (Jacquez 2007). The existence of a cluster can be considered a sign that certain driving factors for events, other than random chance or noise, exist in the area. For instance, clusters with exceptionally high rates of a certain disease might provide clues to the etiology of the disease and could also indicate areas where health care or disease prevention and control measures should be improved (Kulldorff 1999).

Many standard (aspatial) clustering approaches, such as *k*-means (MacQueen 1967) and *k*-medoid (Kaufman and Rousseeuw 1990), are designed to maximize intercluster similarities and minimize intracluster similarities without considering the spatial configuration and distribution of points. Therefore, they should be modified before they are used to assess spatial point clustering to capture important spatial relationships and patterns (Grubesic, Wei, and Murray 2014). For instance, the distance function used in the *k*-means method might distort proximity relationships if spatial outliers exist in the study area (Murray and Estivill-Castro 1998).

Density-based clustering methods, such as DBSCAN (Ester et al. 1996), DENCLUE (Hinneburg and Keim 1998), OPTICS (Ankerst et al. 1999), and DECODE (Pei et al. 2009), assume that the density of points inside a cluster is considerably higher than the density of points outside a cluster (i.e., outliers). In these methods, clusters are defined as dense regions of events in space, usually separated by low-density regions (noise). Density-based clustering methods can find spatial clusters of different sizes and shapes and also filter out noises (Han et al. 2001). DBSCAN (Ester et al. 1996) is one of the most widely used density-based clustering methods. DBSCAN searches for clusters based on the concepts of density connectedness and density reachability. All of the points within the same cluster should be mutually density connected and all of the points not density reachable from any other point should be considered noise (Ester et al. 1996). DBSCAN assumes the underlying intensity that generates events is homogeneous—clusters are areas with spatial density above a uniform threshold. Although many variations of DBSCAN and new approaches such as DECODE have been developed to detect spatial clusters with different densities (Liu, Zhou, and Wu 2007; Pei et al. 2009), these methods still find clusters purely based on the distribution of events, without using any information about underlying intensities that govern the spatial distribution of the events. Therefore, these methods are not suitable for spatially explicit cluster detection in heterogeneous point processes (Murray, Grubesic, and Wei 2014).

Scan-based methods are commonly used for detecting spatial point clusters when such spatial heterogeneity exists (Pei et al. 2011). Examples of these methods include the Geographical Analysis Machine (Openshaw et al. 1987), spatial scan statistics (Naus 1965; Kulldorff and Nagarwalla 1995; Kulldorff 1997), and the methods proposed by Besag and Newell (1991) and by Fotheringham and Zhan (1996). Predefined geographical scanning windows are used to identify areas with elevated or deflated rates of events in scan-based methods (Grubesic, Wei, and Murray 2014). Among all of these methods, the spatial scan statistic by Kulldorff (1997, 1999), along with the software tool SaTScan, stands out for its capability of finding the most likely clusters through likelihood comparison and eliminating intersecting clusters.

Naus (1965) first proposed scan statistics to detect clusters in point processes, which was further extended by Kulldorff (1997) into spatial scan statistics. Using scan windows with varying sizes, spatial scan statistics search for the most likely clusters that cannot be explained by the baseline process under the null hypothesis of complete randomness. The baseline processes include heterogeneous Poisson and Bernoulli processes with known intensity (Kulldorff 1997). The Poisson model handles the number of events with varying intensity such as a known underlying population. The Bernoulli model is usually used to compare the spatial distribution of two types of events (e.g., cases or controls). There are two major limitations of spatial scan statistics. First, computational intensity increases dramatically as a data set grows due to the brute-force search strategy (Pei et al. 2011). Second, the shapes of scanning windows used in spatial scan statistics are fixed (circular or elliptic; Kulldorff et al. 2006), which makes it hard to identify irregularly shaped clusters such as clusters along roads or rivers (Izakian and Pedrycz 2012; Murray, Grubesic, and Wei 2014). Many approaches have been developed to improve the computational performance of scan statistics through approximations (Agarwal et al. 2006) or sampling (Matheny et al. 2016). Although these methods can generate results similar to scan statistics much faster, they are still unable to detect irregularly shaped clusters.

To find irregularly shaped clusters, Tango and Takahashi (2005) proposed FleXScan, a flexibly shaped spatial scan statistic, but FleXScan is more computationally intensive. Duczmal and Assunção (2004) used a simulated annealing strategy to find irregularly shaped clusters, but this method is sensitive to a parameter named fraction value (Pei et al. 2011). Recently, different optimization strategies, such as linear optimization (Murray, Grubesic, and Wei 2014), genetic optimization algorithm (Duczmal et al. 2007), ant colony optimization (Pei et al. 2011), particle swarm optimization (Izakian and Pedrycz 2012), multidirectional optimum ecotope-based algorithm (Aldstadt and Getis 2006), and a hybrid method (Yin and Mu 2017), were also used to find clusters of arbitrary shapes. These methods all assume that input events are aggregated into a small number of polygons and that output clusters are detected by connecting such polygons. Hence, the shape of an output cluster is still limited to combinations of those polygons. Moreover, these methods, which only work for a limited number of polygons, are too complicated and not efficient enough for large point data. Therefore, this article focuses on solving the aforementioned problems of both computational efficiency and limited cluster shapes in spatial point clustering analysis.

Method

Background and the Null Hypothesis

To find clusters of events, it is necessary to define the spatial distribution of events under the null hypothesis of no spatial clusters. Such a distribution can be derived from a point process with a certain underlying intensity. Because the underlying intensity is usually unevenly distributed, the distribution of events should typically also be heterogeneous across space under the null hypothesis. For instance, when detecting disease clusters, it is natural to assume that there is no cluster when the number of disease cases in each area is proportional to the population at risk. Population density usually varies across regions, and the expected number of disease cases should also change accordingly. For instance, 10 disease cases out of a population of 100 are more likely to be a cluster than 10 disease cases out of a population of 10,000.

In addition to a point event data set, another data set is used by ESCIP to provide the underlying intensity that governs the spatial distribution of events if no spatial clusters exist. This data set is referred to as the *background*. A background can be specified in

multiple ways to provide the underlying intensity for the point events. Under the null hypothesis of no clustering, the distribution of events should follow a purely random baseline point process with a known intensity from such a background. Any local excess of events that cannot be explained by the baseline process is identified as a spatial cluster.

One way to account for the background is to directly provide the expected intensity or estimate it from other related phenomena. For instance, in a disease clustering scenario, the expected spatial density of disease cases can be estimated as the product of the overall disease rate and the spatial density of the population at risk. A Poisson model can be used in this situation. For simplicity, when referring to Poisson models, the terms *event* and *case* will be used interchangeably in this article. Another way to model the background is to provide events from another group (controls) as background observations and compare the distribution of the original events (cases) to these controls. For instance, the spatial distribution of people with a certain disease can be compared to the spatial distribution of people without that disease to detect spatial clusters. A Bernoulli model is often used in this situation.

Poisson and Bernoulli Models

To detect statistically significant clusters that are unlikely to have occurred by chance in a point process, it is important to first define the baseline process under the null hypothesis. Theoretically, many point process models can be used for core point detection, including but not limited to Poisson models and Bernoulli models, which are the most popular in spatial point pattern analysis (Kulldorff 1997). This article focuses on these two models.

Poisson Model. In a heterogeneous Poisson model, the number of events n_i in an area i follows a Poisson distribution with an expected value of λ_i; that is, $n_i \sim \text{Poisson}(\lambda_i)$. Under the null hypothesis of no clustering, n_i should follow a Poisson distribution with an expected value of λ_{i0}; that is, $n_i \sim \text{Poisson}(\lambda_{i0})$, where λ_{i0} is the intensity in the area i. A common way to define λ_{i0} is to set it proportional to the number of background observations; that is,

$$\lambda_{i0} = \frac{\text{number of background observations in area } i}{\text{total number of background observations}} \times \text{total number of cases.}$$

Then, a statistical hypothesis test can be used to decide whether area i is a cluster at significance level α, with H_0: $\lambda_i \le \lambda_{i0}$ (not a cluster) and H_a: $\lambda_i > \lambda_{i0}$ (a cluster).

Using the cumulative distribution function (CDF) of a Poisson distribution, the probability of getting c_i or more cases within area i under the null hypothesis can be estimated as

$$P(n_i \ge c_i) = 1 - \sum_{j=0}^{c_i-1} \frac{\lambda_{i0}^{j} e^{-\lambda_{i0}}}{j!}. \quad (1)$$

If the probability is below the significance level α, the area i contains a high concentration of cases and is defined as a cluster.

In this article, the same likelihood ratio function as used by Kulldorff's (1997) spatial scan statistics is used to evaluate results and compare detected clusters. The likelihood function of a cluster is defined as L_C/L_0, where L_C is the maximum likelihood of each scanning window to be a cluster and L_0 is the maximum likelihood when there is no cluster. In a Poisson model, the maximum likelihood ratio function of one cluster with an expected number of cases λ_{i0} and observed number of cases c_i is proportional to Equation 2 as described by Kulldorff (2015):

$$\left(\frac{c_i}{\lambda_{i0}}\right)^{c_i} \left(\frac{C - c_i}{C - \lambda_{i0}}\right)^{C - c_i}, \quad (2)$$

where C is the total number of cases in the entire study area, $C - c_i$ is the number of observed cases outside the cluster, and $C - \lambda_{i0}$ is the expected number of cases outside the cluster.

Bernoulli Model. In a Bernoulli model, each individual point can be in one of two states; for example, people with or without a certain disease, young or old people, or daytime or nighttime events. Individuals in one of the two states are defined as cases and the others as controls. Controls can be considered as a specific type of background observations in this article. In a Bernoulli model, the probability that any individual point is a case is independent of its location and the existence of any other points. Each point in an area i follows a Bernoulli distribution with probability p_i of being a case and probability $1 - p_i$ of being a control. Therefore, the number of cases n_i in an area i with N_i total points should follow a binomial distribution that is the

sum of N_i Bernoulli trials with the same probability p_i; that is, $n_i \sim$ Binomial(N_i, p_i), where $N_i =$ number of cases + number of controls in area i and the expectation of n_i is $N_i p_i$. Under the null hypothesis of no clustering, each individual should have the same probability of being a case regardless of its geographical location and thus p_i should be the same for all areas; that is, $p_i = p_0$. Similar to a Poisson model, clusters in a Bernoulli model can be detected through a statistical hypothesis test with H_0: $p_i \le p_0$ (not a cluster) and H_a: $p_i > p_0$ (a cluster), where $p_0 = \frac{\text{total number of cases}}{\text{total number of cases + total number of controls}} = \frac{C}{N}$ in the entire region. Using the CDF of a binomial distribution, the probability of getting c_i or more cases within area i under the null hypothesis can be estimated as

$$P(n_i \ge c_i) = 1 - \sum_{j=0}^{c_i - 1} \binom{N_i}{j} {p_0}^j (1-p_0)^{N_i - j}. \quad (3)$$

In a Bernoulli model, the maximum likelihood function L_C of one cluster with an observed number of cases c_i can be derived as (Kulldorff 2015)

$$\left(\frac{c_i}{N_i}\right)^{c_i} \left(\frac{N_i - c_i}{N_i}\right)^{N_i - c_i} \left(\frac{C - c_i}{N - N_i}\right)^{C - c_i} \left(\frac{(N - N_i) - (C - c_i)}{N - N_i}\right)^{(N - N_i) - (C - c_i)}. \quad (4)$$

where C is the total number of cases, N_i is the combined number of cases and controls within the cluster, and N is the combined number of cases and controls in the entire study area. Because the maximum likelihood if there is no cluster L_0 is constant for any cluster in a Bernoulli model, L_C can be directly used to find the most likely clusters.

Finding Clusters over an Inhomogeneous Background

Core Points and Spatial Clusters. Assuming there are some true clusters in the study area, the principle of ESCIP is first to identify all the point events (including both cases and background observations) in these clusters as core points. To detect core points in any of these clusters, this article uses a small circular search window with radius ε near each point event. A point event is defined as a core point if the number of cases c_i within distance ε is unlikely to be generated by chance under the null hypothesis of no clustering: The probability of having an equal number of or more cases is less than the given significance level α in the baseline process. Specifically, a core point should satisfy $P(n_i \ge c_i) \le \alpha$, with the λ_{i0} in Equation 1 or the N_i in Equation 3 calculated from the number of background observations (or controls) in the area within radius ε.

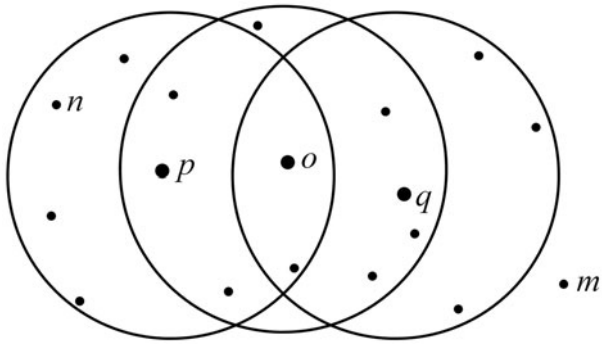

Figure 1. Illustration of reachability. Points o, p, and q are three core points that have passed the statistical hypothesis testing using an intensity derived from the background. Points n and m are not core points, although n is in p's search window and contributes to the testing of p for a core point. Points p and q are directly reachable from point o. Point p and q are reachable from each other. Points o, p, and q are in the same cluster.

Modified from DBSCAN (Ester et al. 1996), additional concepts and definitions are used in this article and an illustration is shown in Figure 1.

- *Directly reachable:* Two core points are directly reachable from each other if they are within distance ε from each other.
- *Reachable:* A core point q is said to be reachable from p if there is a path of core points $p_1, p_2, \ldots, p_n$, with $p_1 = p$ and $p_n = q$, where p_i and p_{i+1} are directly reachable from each other.
- *Cluster:* A cluster consists of a core point and all core points reachable from it. All core points in a cluster are reachable from each other and it does not matter which core point is chosen as the starting point for expansion. Furthermore, these clusters are only potential because their statistical significance is not guaranteed. Only the clusters that passed statistical significance testing are considered as final clusters, which are described later in this section.

Cluster Detection Procedure. ESCIP has three major phases: identifying core points, detecting clusters by expanding core points, and statistical inferences through Monte Carlo simulation. In the first phase, the number of cases and background observations are counted separately in a small circular search window with radius ε from each input

observation. Then, statistical testing is used to decide whether this point is a core point at significance level α using either Equation (1) or Equation (3) based on the selected point process model. For each point i in the data set, if the probability of getting c_i or more cases in the window under the null hypothesis is less than a given significance level α—that is, $P(n_i \geq c_i) \leq \alpha$—then point i is identified as a core point.

In the second phase, an expansion-based approach is used to detect clusters by connecting the core points identified in the first phase. This phase continuously selects an unvisited core point (starting point) and expands it to form a new cluster. During this process, once a core point c is visited, it is added to the current cluster; the algorithm then visits all of the core points that are unvisited and directly reachable from c; the process continues until all core points reachable from the starting point are visited and then moves on to the next starting point to form the next cluster. At the end of this phase, each core point should belong to one and only one cluster, and no noncore points should belong to any cluster. For instance, point n in Figure 1 will not be counted in the cluster containing point o, p, and q. None of these clusters intersect with each other and any two of them are at least distance ε apart. This expansion process is deterministic, and the input and processing orders have no influence on the clustering results. Once a cluster is detected from the expansion procedure, the numbers of cases and background observations in the entire cluster are counted and its likelihood function (e.g., Equation 2 or Equation 4) is calculated. A cluster with a higher likelihood is considered as more significant and important in this article. Clusters with the top N highest likelihood are identified as the top N clusters.

The third phase evaluates the statistical significance of detected clusters through Monte Carlo simulation. This phase estimates the p value of each detected spatial cluster by comparing its likelihood with the maximum likelihood from simulated data sets, which is a standard procedure for the scan statistics (Kulldorff 1997). In each simulated replication, a new point data set is generated from the null hypothesis that there are no spatial clusters. Then the ESCIP cluster detection procedure is applied to the simulated data set, and the maximum likelihood of all detected clusters is recorded. After all replications, the p value of each original cluster is calculated as Equation 5, where N_{Higher} is the number of replications with a higher maximum likelihood and $N_{\text{Replication}}$ is the total number of replications. Clusters with a high estimated p value are considered to be insignificant because a cluster with a higher likelihood is possible if the data set is completely random and hence are filtered out.

$$p = \frac{N_{\text{Higher}} + 1}{N_{\text{Replication}} + 1}. \tag{5}$$

The random point generation ensures that the total numbers of cases and background observations are kept constant across replications. The processes for generating random points are different for Poisson models and Bernoulli models. In a Poisson model, background observations are kept the same. In each replication, cases are randomly sampled from background observations such that each background observation is equally likely to be selected as a case. In a Bernoulli model, random point data sets are generated through random labeling. In each replication, the point locations of the case and background combined data set are fixed, and a fixed number of case labels the same as the original data set are randomly assigned to points in the combined data set.

ESCIP requires two primary input parameters: search radius ε and significance level α. The choice of these parameters needs to ensure that statistical testing can be effectively conducted in most search windows to identify core points. It is difficult to test whether a point is a core point or not with only five observations in its search window. With a larger data volume and denser observations, a smaller search radius is enough to ensure that enough observations exist in any search window and, as a result, finer scale spatial clustering patterns can be revealed. The influence of significance level is limited, especially when the number of observations in a search window is large and the ratios in and outside clusters are noticeably different. In our experiments, commonly used significance levels such as 0.05, 0.01, and 0.005 tend to generate similar clusters with only minor differences. The search radius ε also defines the maximum reachable distance in the expansion phase. With a larger search radius, resulting clusters will have more smoothed boundaries and nearby smaller clusters are more likely to be merged into

larger ones, which might be preferred if a user only needs an overview of the general spatial patterns. A smaller search radius, assuming that it is still sufficient to check for core points, will reveal fine-scale spatial patterns by detecting highly irregularly shaped clusters. Because both the spatial distribution of observations and users' analysis requirements vary, having a single way to decide the best search radius is impossible. Users are suggested to explore different search radii to compare resulting spatial patterns.

Although it is technically possible for a subset of an expanded cluster to have a higher likelihood ratio and higher statistical significance than the entire cluster, it is unlikely to miss important clusters for the following reasons. First, a larger cluster tends to have a higher likelihood than its subset and, as a result, methods based on likelihood tend to overestimate the spatial extents of clusters (Assuncao et al. 2006; Tango and Takahashi 2012). Second, even if the entire cluster has a lower likelihood than a subset of it, the entire cluster is not likely to be insignificant. This is because each core point constituting the entire cluster needs to be significant based on Equation 1 or Equation 3, and the combination of these significant parts should have a higher significance. Third, although exceptions might happen because significance testing based on Equation 1 or Equation 3 is weaker than significance testing based on simulation, realistically it is likely impossible for clusters with thousands or even millions of points that ESCIP is intended to detect.

Computation and Implementation

With the aforementioned expansion approach, ESCIP only needs to check one small search window at each location. Without any optimization or spatial indexing, the computational complexity for finding all of the points within a search window is $O(n)$, where n is the total number of input points. As a result, the total time complexity of ESCIP without Monte Carlo simulation is $O(n^2)$, as such searching is necessary for each input point to test for core points and to expand clusters. Spatial indexing can be leveraged, however, to greatly speed up the search of nearby points and the expansion of clusters, as the search radius is usually much smaller than the extent of a study area. In this article, grid-based indexing is used. Square blocks of the same size as the search radius ε are used to cover the entire study area, and each point is indexed based on the block in which it falls. With this spatial indexing, all potential reachable points from a point p are either in p's block or its eight directly neighboring blocks (nine blocks in total). The complexity is hence reduced to $O(ns)$, where s is the average number of points in the nine blocks, and s is significantly smaller than n.

In comparison, spatial scan statistics require a brute-force search of all potential clusters—scanning windows with different radii need to be tested at each location. Hence, the complexity of spatial scan statistics is much higher. Suppose there are n input points, m potential cluster centers ($m = n$, if input points are also used as cluster centers), and l scanning windows with different radii at each center; its total time complexity without Monte Carlo simulation is $O(nml)$. Because the maximum radius of scanning windows is usually comparable to the extent of a study area, the effect of spatial indexing to improve performance is limited. No matter how points are spatially indexed, spatial scan statistics need to calculate the distances between each cluster center to a large proportion of all input points. As a consequence, ESCIP has a significant computational performance advantage over spatial scan statistics. ESCIP with spatial indexing support is implemented using C for both Bernoulli and Poisson models and will be published as open-source software.

Case Study

To evaluate the effectiveness of ESCIP for detecting significant spatial clusters with any shape in large point data sets, a case study using both baseline processes is carried out in this article. The data set used in this study was generated by Ehlschlaeger et al. (2016), and it contains simulated individual households in Chittagong Division, Bangladesh. There are 5,984,314 individual households in the data set. Each household is geolocated and has socioeconomic and infrastructural attributes including electricity availability and the number of people.

Clustering Households with Electricity

The first experiment detects spatial clusters of households with electricity in Chittagong Division, Bangladesh. This experiment provides an example of

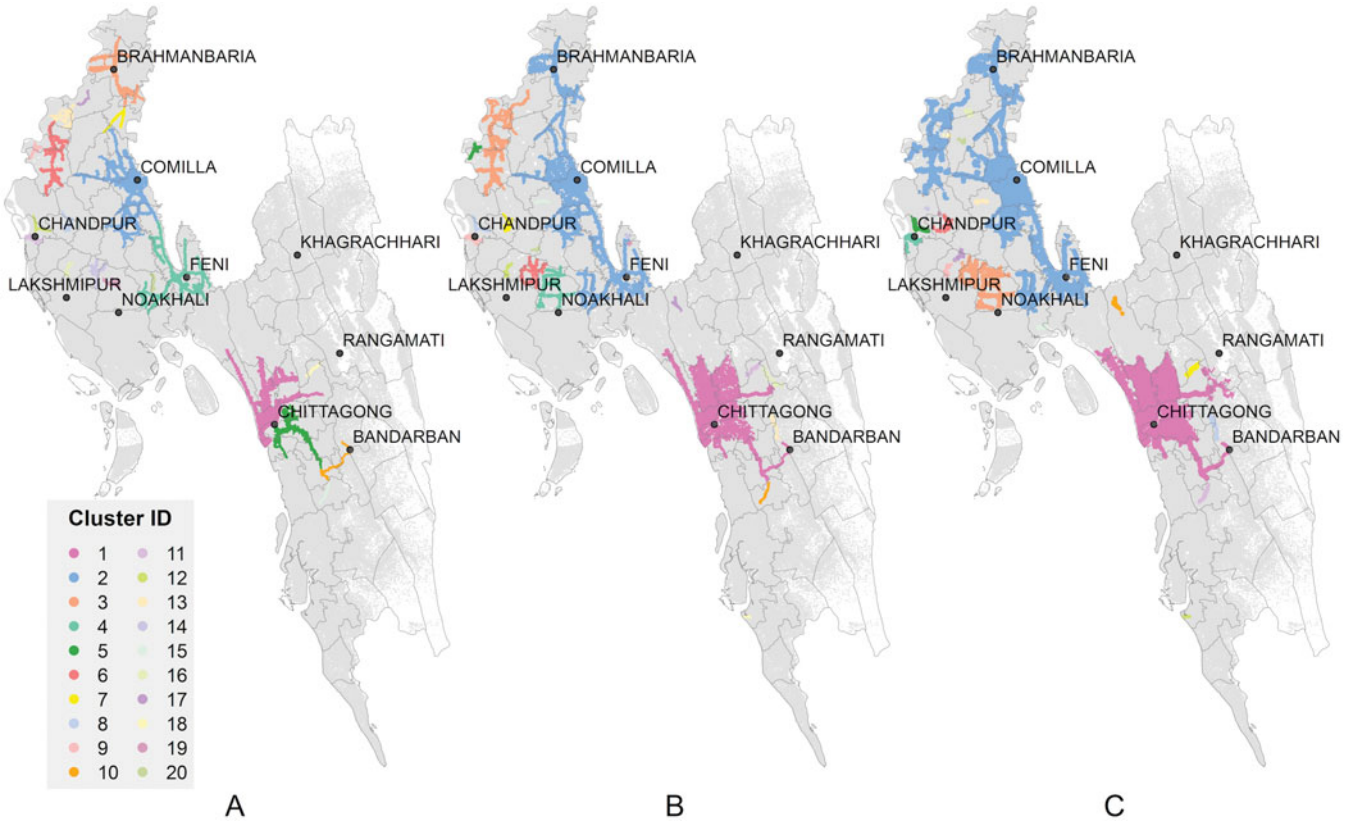

Figure 2. Clusters of households with electricity in Chittagong Division: (A) 250-m search radius; (B) 500-m search radius; (C): 1,000-m search radius.

how to use a Bernoulli model to detect clusters from a point process in a case–control study, where each point can be either a case (household with electricity) or a control (household without electricity). Among the 5,984,314 simulated individual households in Chittagong Division, 3,802,479 are cases and 2,165,437 are controls with respect to electricity availability. The remaining 16,398 households have unknown electricity availability and are not included in this analysis. A search radius of 500 m and a significance level of 0.01 are used in the experiment. The search radius of 500 m is a valid choice because it is much smaller than the extent of Chittagong Division (roughly 220 km by 380 km), and there are at least hundreds of households within most search windows. The top twenty clusters in terms of highest likelihood are shown in Figure 2B. Households in different clusters are colored and households not in any of the twenty clusters are shown in gray. The summary statistics of each cluster are shown in Table 1, including the number of cases and controls, the expected number of cases and controls, and the log-likelihood. The p values of these twenty clusters estimated from Monte Carlo simulation with ninety-nine replications are all 0.01.

The most likely cluster is around Chittagong City, the largest city in the study area and the major coastal seaport city in Bangladesh. There are more than 1 million households in this cluster, and 92.6 percent of them have access to electricity. This cluster covers the most developed areas in Chittagong Division with good infrastructure. The second cluster expands from Brahmanbaria to Feni and includes Comilla, which is the second largest city in the study area. These regions are connected into this cluster by relatively more developed areas near highways. Many remaining clusters are in the western part of the study area, and none of them are in the less developed mountainous regions in the east.

Furthermore, the clustering results using a 500-m search radius (Figure 2B) are compared with the results using a 250-m search radius (Figure 2A) and a 1,000-m search radius (Figure 2C). The top twenty clusters are shown in these figures. High similarities exist in these three clustering results—cluster regions such as Chittagong City, Comilla, Feni, and

Table 1. Summary statistics of the top twenty clusters of households with electricity

Cluster ID	# Cases	# Controls	Exp. cases	Exp. controls	Log-likelihood
1	1,031,258	82,392	709,566	404,084	−3,609,583.084
2	861,968	126,496	629,803	358,661	−3,747,524.134
3	179,871	36,506	137,865	78,512	−3,888,668.637
4	77,889	22,079	63,695	36,273	−3,904,431.174
5	15,212	1,823	10,854	6,181	−3,906,297.989
6	35,402	9,032	28,311	16,123	−3,906,516.601
7	22,180	4,288	16,864	9,604	−3,906,583.941
8	23,917	5,168	18,532	10,553	−3,906,791.628
9	24,443	5,682	19,194	10,931	−3,907,012.823
10	16,931	4,363	13,568	7,726	−3,907,953.600
11	11,407	2,401	8,798	5,010	−3,908,014.941
12	9,503	2,294	7,517	4,280	−3,908,414.919
13	8,035	1,815	6,276	3,574	−3,908,457.387
14	2,908	211	1,987	1,132	−3,908,460.962
15	12,425	3,527	10,164	5,788	−3,908,463.395
16	3,304	394	2,356	1,342	−3,908,579.389
17	6,781	1,597	5,338	3,040	−3,908,618.758
18	6,517	1,648	5,202	2,963	−3,908,711.598
19	1,461	83	984	560	−3,908,794.533
20	2,664	421	1,966	1,119	−3,908,815.541

Brahmanbaria are consistent on all three maps. There are two major differences between the results. First, the shapes of clusters are more smoothed and have fewer holes as a result of a larger search radius. Nearby clusters that are separated when the search radius is small are sometimes combined into larger clusters when the search radius is larger. The second major difference is that more points are identified as core points, and the sizes of clustering areas are larger when the search window increases. A major reason for the difference is the spatial distribution of observations. Because urban areas with higher population density tend to have higher electricity accessibility, spatial clusters of households with electricity are usually in areas with a higher point density. With a larger search window radius, the influence of the high-density areas will be extended farther into low-density areas, and more observations near the cluster boundaries will be identified as core points. If the phenomena being studied are more likely to be in less populous regions, we can expect that fewer core points can be identified and clusters shrink slightly when a larger search radius is used.

To validate the efficiency and effectiveness of ESCIP, its result is compared with the spatial scan statistics. Because the scale of the example problem is beyond the capability of a single-desktop environment, existing software toolkits, such as SaTScan, cannot be used. Thus, in this article, an efficient spatial scan statistic is implemented in C using a cyberGIS approach (Wang and Armstrong 2009; Wang 2010; Wang, Liu, and Padmanabhan 2016). There is a trade-off between the complexity of the scanning windows and the computational intensity in spatial scan statistics. Although having more flexible scanning windows can potentially detect clusters better, it dramatically increases the total number of scanning windows and consequently the computing time. For instance, circular scanning windows with the same center can only vary by size, whereas elliptical ones need to cover any combination of eccentricities, angles, and sizes (Kulldorff et al. 2006). Hence, clustering results and computing times need to be evaluated simultaneously. The spatial scan statistics implementation in this article uses circular scanning windows that are centered at input data points. Numbers of a regular interval (1 km, 2 km, … , 50 km) are used as scanning window radii, which requires much fewer scanning windows than using distances between circle centers to every other point. Further, this implementation is parallelized using OpenMP to improve performance and ensure that it can be finished within a reasonable amount of time. The result of the spatial scan statistics and a

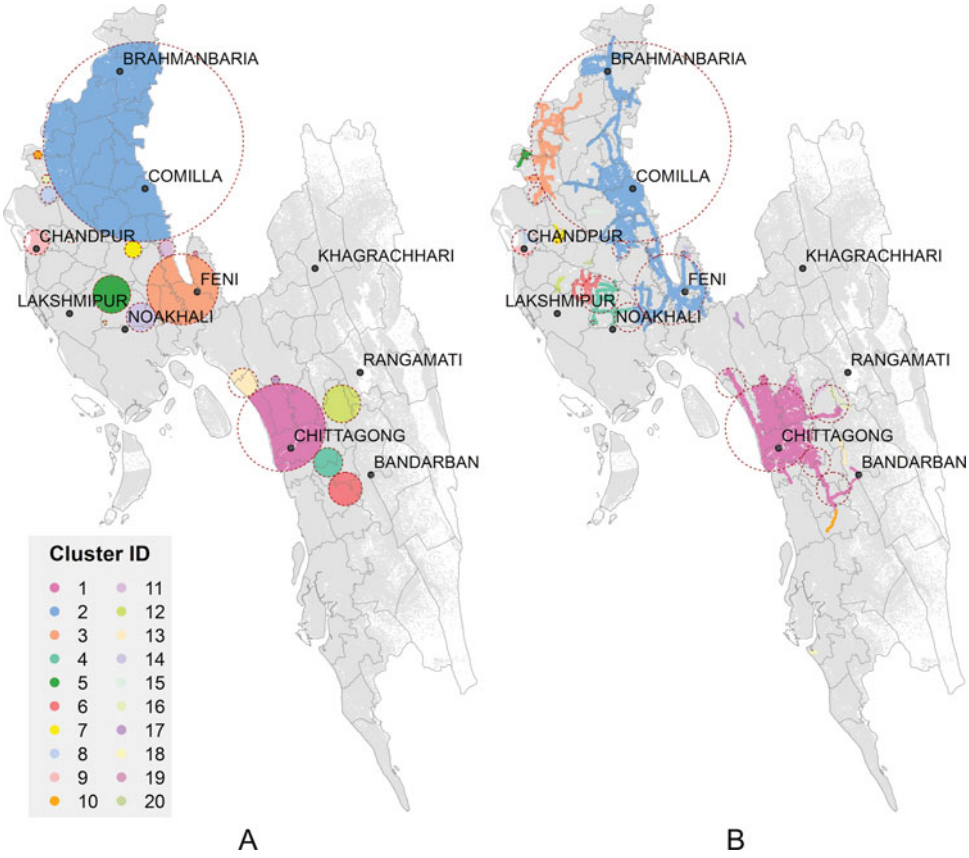

Figure 3. (A) Result of the spatial scan statistics. The dashed circles are the extents of clusters and households in different clusters are colored differently. (B) Comparison between ESCIP clusters (in colors) and scan statistics clusters (extent in dashed circles). ESCIP = expansion-based spatial clustering for inhomogeneous point process.

Table 2. Performance comparison between ESCIP and spatial scan statistics

Percentage of the original data used	Number of input points	Running time		
		ESCIP	Spatial scan	Parallel spatial scan (OpenMP 20 cores)
5.00	299,216	2.714 s	4,570.785 s	349.11 s
10.00	598,432	4.545 s	18,650.286 s	1,356.64 s
20.00	1,196,863	12.415 s	72,811.625 s (>20 hr)	5,343.87 s
50.00	2,992,157	54.044 s	N/A	33,254.154 s
100.00	5,984,314	227.861 s (<4 min)	N/A	136,184.257 s (>37 hr)

Note: ESCIP = expansion-based spatial clustering for inhomogeneous point process.

comparison between the two methods can be found in Figure 3. A comparison of the running times is also shown in Table 2, where subsets (from 5 percent to 100 percent) of the original data are used as the input. The tests were conducted using a computing node with two Intel Xeon E5-2660 processors with ten cores (a total of twenty cores) and 256 GB of RAM.

As shown in Figure 3, there is a significant overlap between the results generated by the two methods. Most clusters of ESCIP fall into spatial scan statistics' clusters but cover much less area with more accurate and flexible shapes. One problem of spatial scan statistics is that all of the points within the scan window are assigned to the same cluster. It is often found that some parts in large clusters have a lower than expected case intensity, even though the overall likelihood of the cluster is high. It shows that the spatial contiguity of a cluster could be violated in spatial scan statistics. For example, cluster 2 of spatial scan statistics contains both large areas with low electricity availability to the west of

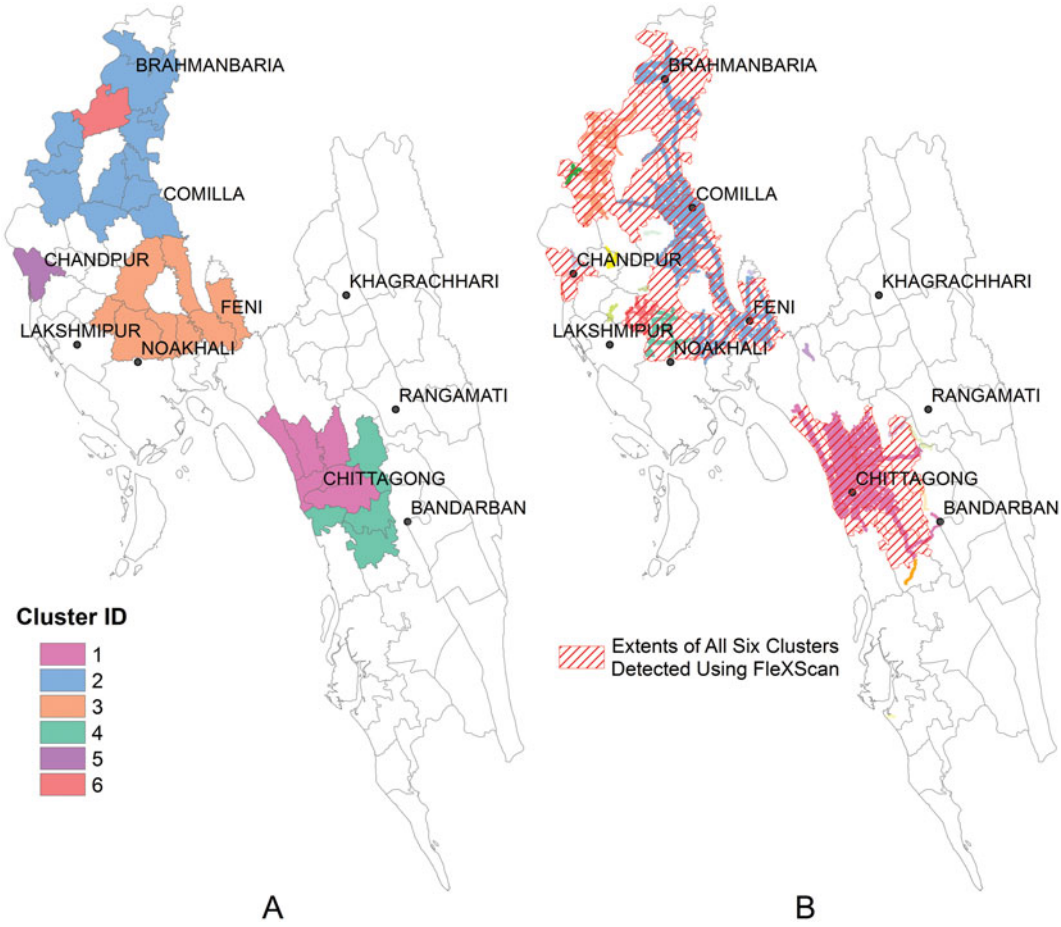

Figure 4. (A) Clusters detected by FleXScan; (B) Comparison between ESCIP and FleXScan. ESCIP = expansion-based spatial clustering for inhomogeneous point process.

Comilla and areas outside of the study areas without data. This problem can be resolved through our method as each small part of a final cluster must pass a local statistical test. With the expansion-based approach only connecting those areas that have passed the test, every part of a detected cluster is significant and contiguous. As a result, the low-intensity parts can be eliminated effectively. Another problem of spatial scan statistics is that the shape of the scanning window is fixed, which limits the shape of output clusters. Some important parts outside the scan window might be excluded because of the limited window shapes. For instance, cluster 1 of ESCIP represents a connected region with high electricity availability near Chittagong City. Spatial scan statistics detect it as multiple clusters (clusters 1, 4, 6, 12, 13, and 17), however, due to the limitation of the circular shape. Moreover, the maximum of log-likelihood values in our approach (−3,609,583.084) is larger than that in the spatial scan statistics approach (−3,628,583.79), although our method does not expand clusters to maximize likelihood values. It demonstrated that even using the criteria of scan statistics, ESCIP detects clusters in a more effective way than the scan statistics approach.

ESCIP is also much more computationally efficient than spatial scan statistics. Table 2 shows that ESCIP is thousands of times faster than spatial scan statistics. When more than half of the original data set is used, the spatial scan statistics cannot finish within two days (the running time is marked as N/A in Table 2). A parallel spatial scan statistic with twenty cores (forty threads) barely finishes the experiments in two days, and ESCIP is still hundreds of times faster. Monte Carlo simulation was not conducted for spatial scan statistics because it was too slow for spatial scan statistics to finish the hundreds of simulations.

The result from FleXScan, a well-known irregularly shaped spatial scan statistic, by Tango and Takahashi (2005), is also shown in Figure 4 for comparison. As mentioned in the related work, existing

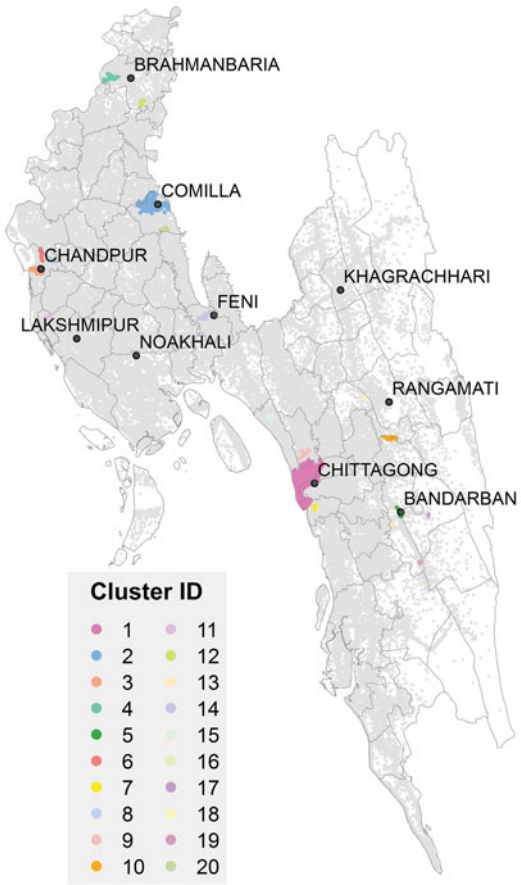

Figure 5. Clusters of single-person households in Chittagong Division.

approaches to irregularly shaped spatial scan statistics can only work with a limited number of aggregated polygons. Thus, the individual-level households are aggregated into case and control counts at the *upazila* (subdistrict) level before using FleXScan. FlexScan identified most upazilas that intersect with clusters detected by ESCIP, but it cannot accurately depict the fine spatial details of these clusters due to aggregation. By aggregating data into finer spatial units, FleXScan might potentially detect more flexibly shaped clusters, but it is difficult because of the computational complexity. In our experiment, FlexScan could not finish within a week if aggregation scales finer than upazila were used.

Clustering Single-Person Households

The second experiment detects spatial clusters of single-person households in Chittagong Division, Bangladesh. This experiment provides an example of how to use a Poisson model to detect clusters with a spatially varying population density. There are 269,871 single-person households in the data set, which is roughly 4.5 percent of all households. The experiment was also conducted using a search radius of 500 m and a significance level of 0.01.

Table 3. Summary statistics of the top twenty clusters of single-person households

Cluster ID	# Cases	# Exp. cases	Log-likelihood ratio
1	65,207	27,656.7	21,451.689
2	7,300	4,184.8	964.994
3	1,315	718.7	198.814
4	872	464.1	142.342
5	284	102.2	108.542
6	750	431.8	96.072
7	392	201.3	70.612
8	499	309.9	48.667
9	215	104.2	44.934
10	146	60.2	43.491
11	172	83.4	35.934
12	355	219.5	35.188
13	70	24.4	28.099
14	364	250.6	22.524
15	84	36.7	22.307
16	49	16.2	21.365
17	39	11.8	19.493
18	17	2.3	19.306
19	59	24.4	17.501
20	169	104.2	16.951

Figure 5 shows the top twenty clusters of single-person households. Table 3 shows the number of cases, the expected number of cases, and the log-likelihood ratio (Equation 2) of each cluster. The p values of these twenty clusters estimated from Monte Carlo simulation with ninety-nine replications are all 0.01. The two most noticeable clusters are in the two largest urban areas (Chittagong and Comilla) in this region. Many other smaller clusters are either in or around towns such as Feni, Chandpur, Brahmanbaria, and Bandarban. None of the clusters expand to large areas connecting different regions. This pattern is different from that of households with electricity, which has many large interregion clusters with hundreds of thousands of households. Households usually have electricity if they are in regions with electrical infrastructure. As a result, households with electricity are often distributed around big cities and wealthy areas and have a strong clustering pattern. It is uncommon for a household to not have electricity if all of its neighbors have electricity. Single-person households usually do not have such a strong clustering pattern

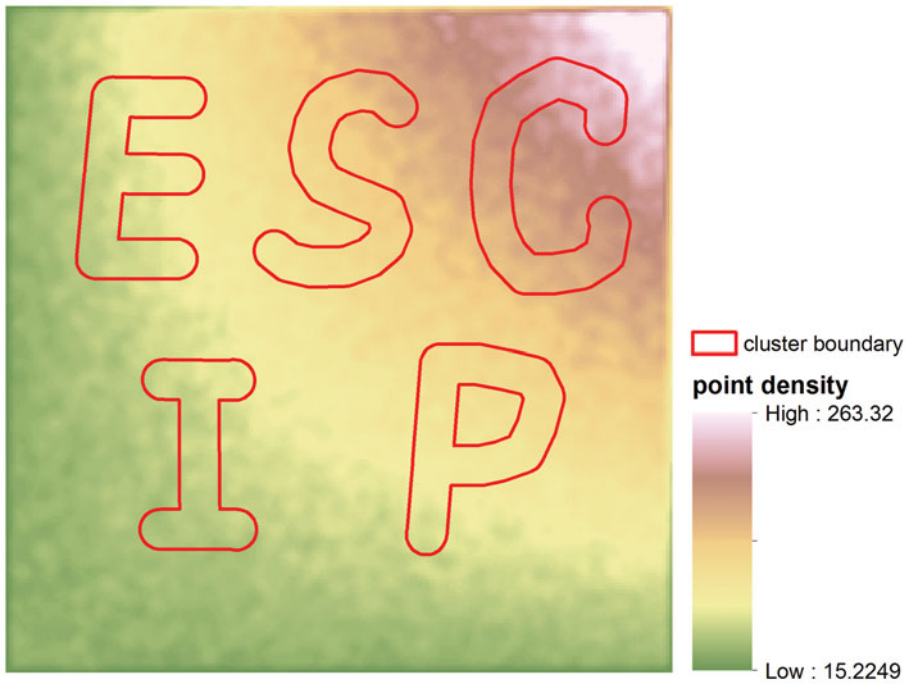

Figure 6. Simulated point density and the shapes of five clusters. ESCIP = expansion-based spatial clustering for inhomogeneous point process.

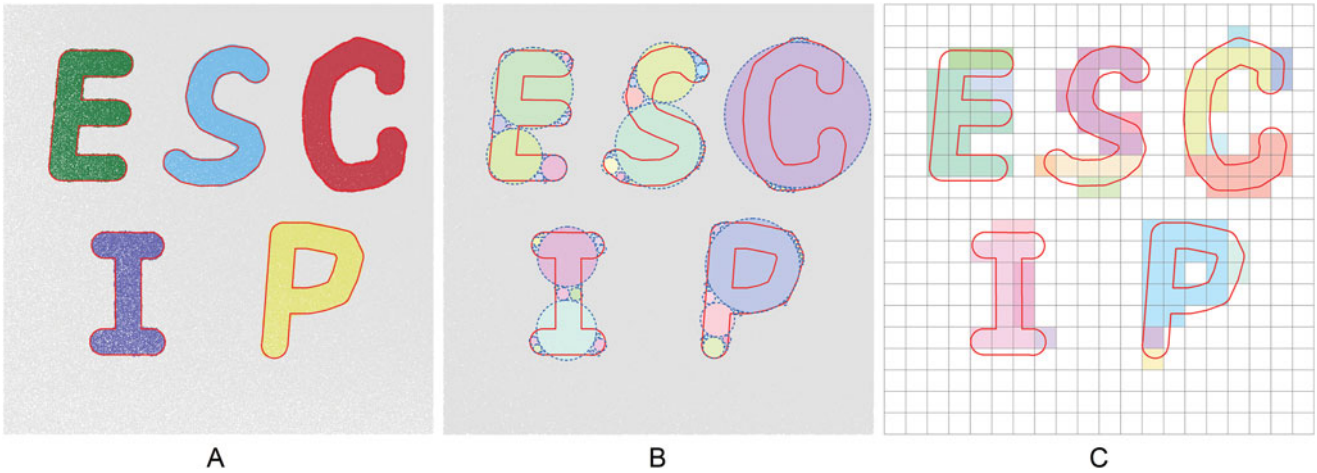

Figure 7. Clustering results of the simulated clusters: (A) result of ESCIP; (B) result of spatial scan statistic; (C) result of FleXScan using fishnet cells. ESCIP = expansion-based spatial clustering for inhomogeneous point process.

because it is more common for a single-person household to be surrounded by multiperson households.

Evaluation Based on a Simulated Data Set with Known Clusters

To further evaluate the performance of ESCIP, another data set with known spatial clusters is randomly generated. This data set contains 1,000,000 individual points that are unevenly distributed in a square region with side length 100—point density increases gradually from the bottom-left corner to the top-right corner. Five clusters are created with the shapes of the letters E, S, C, I, and P, respectively. The density of points and the shapes of the clusters are shown in Figure 6. Points within any cluster have a probability of 0.5 to be a case; points outside have a probability of 0.1.

Figure 7 shows the clustering result of the three methods on this simulated data set. The result of ESCIP (Figure 7A) was generated using a search window of radius 1 and a significance level of 0.01.

Table 4. Clustering performance of three methods using a simulated data set with known clusters

	# Clusters	TP (%)	FP (%)	TN (%)	FN (%)	Accuracy (%)	Precision (%)	Recall (%)
ESCIP	5	21.24	1.89	76.86	0.01	98.10	91.83	99.96
Spatial scan statistics	99	20.38	14.23	64.52	0.87	84.90	58.88	95.89
FleXScan (fishnet)	22	19.71	9.43	69.32	1.54	89.02	67.63	92.74

Notes: TP = true positive; FP = false positive; TN = true negative; FN = false negative.

ESCIP found five clusters in total, each of which corresponds to one letter. All of these five clusters have a p value of 0.01 based on Monte Carlo simulation with ninety-nine replications. For spatial scan statistics, circular scanning windows centered at data points are used with radii 0.5, 1.0, 1.5, …, 25.0. Figure 7B shows the clusters detected by the spatial scan statistic. As the cluster shape of the spatial scan statistic is limited by circular windows, it fails to capture the shape of the true clusters. Each cluster, except cluster C, is separated into several large circular clusters along with many small ones that are hardly visible on the map. For FleXScan, individual input points are aggregated into the case and population counts using a 5×5 fishnet grid, which results in 400 square regions. Twenty-two significant clusters are found using FleXScan, as shown in Figure 7C. Although these clusters depict the rough shapes of the five simulated clusters, their shapes are limited by the fishnet grids and thus the results cannot accurately reflect the extents of the five letters. In addition, each of the five simulated clusters is partitioned into multiple ones in FleXScan's result. Finally, it is not realistic to further improve the spatial granularity when using FleXScan, as the computation cannot finish within weeks with a finer fishnet grid.

Table 4 compares the clustering performance of ESCIP, spatial scan statistic, and FleXScan. The TP, FP, TN, and FN in Table 4 stand for true positive, false positive, true negative, and false negative, respectively. The top ninety-nine clusters are used for spatial scan statistics result because the overall accuracy is maximized with ninety-nine clusters. ESCIP achieves the best accuracy, precision, and recall among all methods. The errors mostly occur along the edge of the simulated cluster. The result of spatial scan statistics includes a large number of non-cluster points and thus has a higher false positive rate than other methods. As a result, the precision of spatial scan statistics is low. Compared with spatial scan statistics, FleXScan's result has better shapes and therefore higher precision, yet is still inferior to ESCIP.

Discussion and Conclusion

This article describes ESCIP, an expansion-based spatial clustering method for efficiently detecting spatial clusters of flexible shapes within heterogeneous point processes. ESCIP is established on two baseline process models: Poisson and Bernoulli. A case study with two experiments was conducted to evaluate ESCIP using large individual-level geospatial data. Spatial clusters of single-person households and households with electricity were detected from simulated geolocated households in Chittagong Division, Bangladesh. Experimental results showed that ESCIP can efficiently detect clusters of flexible shapes in point processes with known heterogeneous intensity. Compared to spatial scan statistics, ESCIP has three primary advantages. First, ESCIP can detect clusters with more accurate and flexible shapes. Second, the computational performance of ESCIP is high. Third, each cluster detected by ESCIP is contiguous and every part of it is statistically significant.

The experimental results also showed that ESCIP can efficiently detect spatial patterns from large geospatial data sets. Spatial clustering results can be generated within several minutes given a data set with nearly 6 million individual households, each having a different location. With such high efficiency, this article demonstrated that it is possible to directly analyze large-scale geographic phenomena using individual-level records, without having to aggregate them into predefined areal units. The increasing availability of large spatial data sets provides tremendous opportunities for the ESCIP approach. For instance, ESCIP may potentially be applied to detect abnormal high concentrations of massive geolocated social media posts for event detections (Gao, Wang, et al. 2018).

ESCIP can be further improved in multiple aspects. One aspect is to better assess the influence of parameters, specifically significance level and search radius, and to provide methods or guidelines for optimal parameter choices. Systematic evaluation needs to be conducted based on application-specific characteristics for developing generalizable methods and

desirable guidelines. Another future research direction is to expand from spatial to spatiotemporal clustering analysis. Finding clusters with flexible spatiotemporal shapes requires innovative approaches to expanding clusters in both space and time, which is challenging and thus requires exciting research. Finally, it is important to incorporate more spatial or spatiotemporal point process models beyond Poisson and Bernoulli for detecting clusters with different null hypotheses. New models such as space–time permutation (Kulldorff et al. 2005) might be considered to evaluate whether and how spatial patterns change over time.

Acknowledgments

The authors are grateful for help and support from Dr. Sara McLafferty and Dr. Charles R. Ehlschlaeger. The authors also acknowledge insightful comments on earlier drafts received from Editor Ling Bian and anonymous reviewers. The research outcome benefits from helpful critique and feedback from Rebecca Vandewalle and other members of the CyberInfrastructure and Geospatial Information Laboratory at the University of Illinois at Urbana–Champaign, which are greatly appreciated.

Funding

This research is based in part on work supported by the U.S. National Science Foundation under grant numbers 1443080, 1743184, and 1833225. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Agarwal, D., A. McGregor, J. M. Phillips, S. Venkatasubramanian, and Z. Zhu. 2006. Spatial scan statistics: Approximations and performance study. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, ed. L. Ungar, M. Craven, D. Gunopulos, and T. Eliassi-Rad, 24–33. New York: ACM.

Aldstadt, J., and A. Getis. 2006. Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis* 38 (4):327–43. doi: 10.1111/j.1538-4632.2006.00689.x.

Ankerst, M., M. M. Breunig, H. P. Kriegel, and J. Sander. 1999. OPTICS: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD international conference on management of data*, 49–60. New York: ACM.

Anselin, L. 1995. Local indicators of spatial association—LISA. *Geographical Analysis* 27 (2):93–115. doi: 10.1111/j.1538-4632.1995.tb00338.x.

Assuncao, R., M. Costa, A. Tavares, and S. Ferreira. 2006. Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine* 25 (5):723–42. doi: 10.1002/sim.2411.

Besag, J., and J. Newell. 1991. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 154 (1):143–55. doi: 10.2307/2982708.

Cheng, T., and T. Wicks. 2014. Event detection using Twitter: A spatio-temporal approach. *PLoS ONE* 9 (6):e97807. doi: 10.1371/journal.pone.0097807.

Duczmal, L., and R. Assunção. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* 45 (2):269–86. doi: 10.1016/S0167-9473(02)00302-X.

Duczmal, L., A. L. Cançado, R. H. Takahashi, and L. F. Bessegato. 2007. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis* 52 (1):43–52. doi: 10.1016/j.csda.2007.01.016.

Eck, J., S. Chainey, J. Cameron, and R. Wilson. 2005. *Mapping crime: Understanding hotspots*. Washington, DC: National Institute of Justice.

Ehlschlaeger, C. R., Y. Gao, J. D. Westervelt, R. C. Lozar, M. V. Drigo, J. A. Burkhalter, C. L. Baxter, M. D. Hiett, N. R. Myers, and E. R. Hartman. 2016. Mapping neighborhood scale survey responses with uncertainty metrics. *Journal of Spatial Information Science* 2016 (13):103–30.

Elwood, S. 2008. Volunteered geographic information: Key questions, concepts and methods to guide emerging research and practice. *GeoJournal* 72 (3–4):133–35. doi: 10.1007/s10708-008-9187-z.

Ertöz, L., M. Steinbach, and V. Kumar. 2003. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 2003 SIAM international conference on data mining*, ed. D. Barbara and C. Kamath, 47–58. Philadelphia, PA: SIAM.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining*, ed. E. Simoudis, J. Han, and U. Fayyad, 226–31. Menlo Park, CA: AAAI Press.

Flanagin, A. J., and M. J. Metzger. 2008. The credibility of volunteered geographic information. *GeoJournal* 72 (3–4):137–48. doi: 10.1007/s10708-008-9188-y.

Fotheringham, A. S., and F. B. Zhan. 1996. A comparison of three exploratory methods for cluster detection in spatial point patterns. *Geographical Analysis* 28 (3):200–18. doi: 10.1111/j.1538-4632.1996.tb00931.x.

Gao, Y., T. Li, S. Wang, M. H. Jeong, and K. Soltani. 2018. A multidimensional spatial scan statistics approach to movement pattern comparison. *International Journal of Geographical Information Science* 32 (7):1304–25. https://doi.org/10.1080/13658816.2018.1426859.

Gao, Y., S. Wang, A. Padmanabhan, J. Yin, and G. Cao. 2018. Mapping spatiotemporal patterns of events

using social media: A case study of influenza trends. *International Journal of Geographical Information Science* 32 (3):425–49. doi: 10.1080/13658816.2017.1406943.

Gatrell, A. C., T. C. Bailey, P. J. Diggle, and B. S. Rowlingson. 1996. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers* 21 (1):256–74. doi: 10.2307/622936.

Grubesic, T. H., R. Wei, and A. T. Murray. 2014. Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers* 104 (6):1134–56. doi: 10.1080/00045608.2014.958389.

Guha, S., R. Rastogi, and K. Shim. 1998. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD international conference on management of data*, ed. A. Tiwary and M. Franklin, 73–84. New York: ACM.

Han, J., M. Kamber, and A. K. H. Tung. 2001. Spatial clustering methods in data mining: A survey. In *Geographic data mining and knowledge discovery*, ed. H. J. Miller and J. Han, 188–217. Didcot, UK: Taylor & Francis.

Hinneburg, A., and D. A. Keim. 1998. An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of the fourth international conference on knowledge discovery and data mining*, R. Agrawal and P. Stolorz, 58–65. Menlo Park, CA: AAAI Press.

Izakian, H., and W. Pedrycz. 2012. A new PSO-optimized geometry of spatial and spatio-temporal scan statistics for disease outbreak detection. *Swarm and Evolutionary Computation* 4:1–11. doi: 10.1016/j.swevo.2012.02.001.

Jacquez, G. M. 2007. Spatial cluster analysis. In *The handbook of geographic information science*, ed. J. P. Wilson and A. S. Fotheringham, 395–416. Hoboken, NJ: Blackwell.

Karypis, G., E. H. Han, and V. Kumar. 1999. Chameleon: Hierarchical clustering using dynamic modeling. *Computer* 32 (8):68–75.

Kaufman, L. R., and P. J. Rousseeuw. 1990. *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: Wiley.

Kriegel, H. P., P. Kröger, J. Sander, and A. Zimek. 2011. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (3):231–40. doi: 10.1002/widm.30.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics: Theory and Methods* 26 (6):1481–96. doi: 10.1080/03610929708831995.

Kulldorff, M. 1999. Spatial scan statistics: Models, calculations, and applications. In *Scan statistics and applications*, ed. J. Glaz and N. Balakrishnan, 303–22. Boston: Birkhäuser.

Kulldorff, M. 2015. SaTScan user guide for version 9.4. Accessed January 5, 2018. http://www.satscan.org/.

Kulldorff, M., R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari. 2005. A space–time permutation scan statistic for disease outbreak detection. *PLoS Medicine* 2 (3):e59. doi: 10.1371/journal.pmed.0020059.

Kulldorff, M., L. Huang, L. Pickle, and L. Duczmal. 2006. An elliptic spatial scan statistic. *Statistics in Medicine* 25 (22):3929–43. doi: 10.1002/sim.2490.

Kulldorff, M., and N. Nagarwalla. 1995. Spatial disease clusters: Detection and inference. *Statistics in Medicine* 14 (8):799–810. doi: 10.1002/sim.4780140809.

Kuo, F. Y., T. H. Wen, and C. E. Sabel. 2018. Characterizing diffusion dynamics of disease clustering: A modified space–time DBSCAN (MST-DBSCAN) algorithm. *Annals of the American Association of Geographers* 108 (4):1168–86. doi: 10.1080/24694452.2017.1407630.

Kwan, M. P. 2016. Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers* 106 (2):274–82.

Liu, P., D. Zhou, and N. Wu. 2007. VDBSCAN: Varied density based spatial clustering of applications with noise. In *2007 international conference on service systems and service management*, ed. J. Chen, 1–4. Piscataway, NJ: IEEE.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability:Vol. 1. Statistics*, ed. L. M. Le Cam and J. Neyman, 281–97. Berkeley: University of California Press.

Matheny, M., R. Singh, L. Zhang, K. Wang, and J. M. Phillips. 2016. Scalable spatial scan statistics through sampling. In *Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems*, ed. M. Ali, S. Newsam, S. Ravada, M. Renz, and G. Trajcevski, 20. New York: ACM.

Murray, A. T., and V. Estivill-Castro. 1998. Cluster discovery techniques for exploratory spatial data analysis. *International Journal of Geographical Information Science* 12 (5):431–43. doi: 10.1080/136588198241734.

Murray, A. T., T. H. Grubesic, and R. Wei. 2014. Spatially significant cluster detection. *Spatial Statistics* 10:103–16. doi: 10.1016/j.spasta.2014.03.001.

Nakaya, T., and K. 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):223–39. doi: 10.1111/j.1467-9671.2010.01194.x.

Naus, J. L. 1965. Clustering of random points in two dimensions. *Biometrika* 52 (1–2):263–66. doi: 10.1093/biomet/52.1-2.263.

Neill, D. B., and A. W. Moore. 2004. Rapid detection of significant spatial clusters. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, ed. R. Kohavi, J. Gehrke, W. DuMouchel, and J. Ghosh, 256–65. New York: ACM.

Openshaw, S., M. Charlton, C. Wymer, and A. Craft. 1987. A Mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information System* 1 (4):335–58. doi: 10.1080/02693798708927821.

Pei, T., A. Jasra, D. J. Hand, A. X. Zhu, and C. Zhou. 2009. DECODE: A new method for discovering clusters of different densities in spatial data. *Data Mining and Knowledge Discovery* 18 (3):337–69. doi: 10.1007/s10618-008-0120-3.

Pei, T., Y. Wan, Y. Jiang, C. Qu, C. Zhou, and Y. Qiao. 2011. Detecting arbitrarily shaped clusters using ant colony optimization. *International Journal of Geographical Information Science* 25 (10):1575–95. doi: 10.1080/13658816.2010.533674.

Plotkin, J. B., J. Chave, and P. S. Ashton. 2002. Cluster analysis of spatial patterns in Malaysian tree species. *The American Naturalist* 160 (5):629–44. doi: 10.1086/342823.

Rogerson, P., and I. Yamada. 2008. *Statistical detection and surveillance of geographic clusters*. Boca Raton, FL: CRC.

Tango, T., and K. Takahashi. 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 4 (1):11. doi: 10.1186/1476-072X-4-11.

Tango, T., and K. Takahashi. 2012. A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. *Statistics in Medicine* 31 (30): 4207–18.

Tramacere, A., and C. Vecchio. 2013. γ-Ray DBSCAN: A clustering algorithm applied to Fermi-LAT γ-ray data—I. Detection performances with real and simulated data. *Astronomy & Astrophysics* 549:A138. doi: 10.1051/0004-6361/201220133.

Wang, S. 2010. A cyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers* 100 (3):535–57.

Wang, S., and M. P. Armstrong. 2009. A theoretical approach to the use of cyberinfrastructure in geographical analysis. *International Journal of Geographical Information Science* 23 (2):169–93. doi: 10.1080/13658810801918509.

Wang, S., Y. Liu, and A. Padmanabhan. 2016. Open cyberGIS software for geospatial research and education in the big data era. *SoftwareX* 5:1–5. doi: 10.1016/j.softx.2015.10.003.

Yin, P., and L. Mu. 2017. A hybrid method for fast detection of spatial disease clusters in irregular shapes. *GeoJournal* 83 (4): 693–705. https://doi.org/10.1007/s10708-017-9799-2.

TING LI received her MS in geography in 2018 from the University of Illinois at Urbana–Champaign, Urbana, IL 61801. E-mail: tingli3@illinois.edu. Her research interests include cyberGIS and spatial clustering analysis.

YIZHAO GAO received his PhD in Geography in 2018 from the University of Illinois at Urbana–Champaign, Urbana, IL 61801. E-mail: ygao29@illinois.edu. He is a software engineer at Google. His research interests include cyberGIS, high-performance computing, spatial analysis, and spatial data science.

SHAOWEN WANG is Professor and Head of the Department of Geography and Geographic Information Science at the University of Illinois at Urbana–Champaign, Urbana, IL 61801. E-mail: shaowen@illinois.edu. His research interests focus on geographic information science and systems (GIS), advanced cyberinfrastructure and cyberGIS, complex environmental and geospatial problems, computational and data sciences, high-performance and distributed computing, and spatial analysis and modeling.