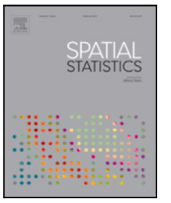
Spatial Statistics 61 (2024) 100839

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



A hypothesis test for detecting spatial patterns in categorical areal data

Stella Self [a,*], Xingpei Zhao [a], Anja Zgodic [a], Anna Overby [b], David White [c], Alexander C. McLain [a,1], Caitlin Dyckman [d,1]

[a] Arnold School of Public Health, University of South Carolina, 921 Assembly Street, Columbia, SC 29208, USA
[b] United States Department of Agriculture, Forest Service, Southern Research Station, Forest Economics and Policy, Research Triangle Park, NC, USA
[c] College of Behavioral, Social and Health Sciences, Clemson University, Epsilon Zeta Dr, Clemson, SC 29634, USA
[d] College of Architecture, Art and Construction, Clemson University, Fernow Street, Clemson, SC 29634, USA

ARTICLE INFO

Keywords:
Cluster detection
Clustering
Categorical areal data
Positive area proportion function

ABSTRACT

The vast growth of spatial datasets in recent decades has fueled the development of many statistical methods for detecting spatial patterns. Two of the most commonly studied spatial patterns are clustering, loosely defined as datapoints with similar attributes existing close together, and dispersion, loosely defined as the semi-regular placement of datapoints with similar attributes. In this work, we develop a hypothesis test to detect spatial clustering or dispersion at specific distances in categorical areal data. Such data consists of a set of spatial regions whose boundaries are fixed and known (e.g., counties) associated with a categorical random variable (e.g. whether the county is rural, micropolitan, or metropolitan). We propose a method to extend the positive area proportion function (developed for detecting spatial clustering in binary areal data) to the categorical case. This proposal, referred to as the categorical positive areal proportion function test, can detect various spatial patterns, including homogeneous clusters, heterogeneous clusters, and dispersion. Our approach is the first method capable of distinguishing between different types of clustering in categorical areal data. After validating our method using an extensive simulation study, we use the categorical positive area proportion function test to detect spatial patterns in Boulder County, Colorado USA biological, agricultural, built and open conservation easements.

1. Introduction

Detecting spatial patterns is important in many fields, including ecology, sociology, epidemiology and environmental planning. Clustering is one of the most common spatial patterns. It can be broadly defined as an excess of events in a particular area (e.g., a large number of households infected with a disease in a neighborhood) or as a group of similar observations in close proximity (e.g., a neighborhood of households with similar incomes). The most appropriate method for detecting spatial patterns often depends on whether each observation is associated with a particular coordinate location (point process data) or a spatial region such as a Census tract or county (areal data). Different spatial patterns can occur at different distances simultaneously. For example, small clusters

* Corresponding author.
E-mail addresses: scwatson@mailbox.sc.edu (S. Self), xingpei@email.sc.edu (X. Zhao), azgodic@email.sc.edu (A. Zgodic), anna.overby@usda.gov (A. Overby), whitedl@clemson.edu (D. White), mclaina@mailbox.sc.edu (A.C. McLain), cdyckma@clemson.edu (C. Dyckman).
[1] Shared Last Author.

https://doi.org/10.1016/j.spasta.2024.100839
Received 26 October 2023; Received in revised form 27 March 2024; Accepted 23 April 2024
Available online 4 May 2024

could cluster together to make larger clusters or could occur at semi-regular intervals to create small distance clustering and large distance dispersion. In many fields, the distance at which spatial patterns manifest is informative about the processes underlying the pattern. Statistical methods that can detect spatial patterns at specific distances are thus highly desirable.

In this work, we develop a hypothesis testing procedure to detect spatial clustering or dispersion in categorical areal data at specific distances. This data type consists of a set of spatial regions whose boundaries are fixed and known, with each region being associated with a (non-ordered) categorical random variable. For example, the set of regions might be all the counties in a particular state with the categorical random variable being whether the county is rural, micropolitan, or metropolitan. In our motivating data application, we have the set of all land parcels in Boulder County, Colorado, some of which are held as conservation easements (CEs). Data on the type of CE (agricultural, environmental, open, etc.) is available. It is of interest to determine if spatial patterns exist between different types of CEs. We treat the areal units (e.g land parcels) as fixed and known; only the categories are considered random.

Several methods exist for detecting spatial clustering in numeric areal data. Popular methods of detecting spatial autocorrelation are global and local Moran's I statistics (Moran, 1950; Anselin, 1995) (positive spatial autocorrelation often manifests as clustering). Global and local Getis Ord statistics can also be used to detect clustering (Getis and Ord, 1992; Anselin, 1995); such an approach is often referred to as a 'hotspot' analysis. Spatial scan statistics can also detect hotspots and clusters. Scan statistics have been developed for many types of data, including binomial (Kulldorff, 1997), count (Kulldorff, 1997), ordinal (Jung et al., 2007), normal (Kulldorff et al., 2009), and time-to-event data (Huang et al., 2007). Bayesian spatial scan statistics have also been developed for (zero-inflated) binomial and count data (Neill et al., 2005; Cançado et al., 2017).

Methods for detecting spatial patterns in categorical areal data are much more limited than in the continuous case. The multinomial scan statistic can be used to detect clustering in categorical areal data, but it cannot detect dispersion (Jung et al., 2010). The positive area proportion function can detect clustering and dispersion at specific distances in binary areal data (e.g. categorical areal data with only two categories) (Self et al., 2023). Several methods exist for detecting spatial patterns in categorical point process data, including Neyman–Scott process models (Neyman and Scott, 1958) and multivariate point process models (Majumder et al., 2022). However, these models assume randomness is present in both the locations of the observations and the observed categories, making them inherently inappropriate for categorical areal data.

In this work, we extend the positive area proportion function and the associated hypothesis test for spatial patterns to categorical areal data with an arbitrary number of categories. Our method can detect a variety of spatial patterns at specific distances, including dispersion, clusters composed of a single category, and clusters composed of multiple categories. Unlike the spatial scan statistic, our method can identify what *type* of clustering is present. For example, our method can distinguish between clusters consisting of category A units surrounded by category B units versus clustering consisting of category A units surrounded by category C units versus clusters of category A units alone. As different types of clustering often have different implications in practice, this information is a valuable contribution to the detection of spatial patterns. Our method is the first hypothesis testing procedure for categorical areal data that is capable of distinguishing between different types of clustering and identifying the distances at which spatial patterns are occurring. We present the categorical positive areal proportion function in Section 2, assess its performance via a simulation study in Section 3 and apply it to the Boulder County data in Section 4. Section 5 contains concluding remarks.

2. Methodology

2.1. Categorical areal processes

Consider a fixed set of areal units, where each falls into exactly one of $K \geq 2$ unordered categories. We assume the boundaries of the areal units are fixed and known and only the assignment of categories is random. Formally, we define a *categorical areal process* (CAP) $\mathcal{A}$ on a Borel set $\mathscr{A} \subset \mathbb{R}^2$ as a collection of N disjoint Borel sets $a_1, a_2, \ldots, a_N$ whose union covers $\mathscr{A}$ and a random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)'$, where $Y_i = k$ if a_i is in category k. We say a CAP is *stationary* if for any $k \in \{1, 2, \ldots, K\}$, $P(Y_i = k) = p_k$ for all i and *independent* if $Y_1, Y_2, \ldots, Y_N$ are mutually independent.

CAPs can exhibit a variety of spatial patterns, which can be classified as clustering or dispersion. Some examples are depicted in Fig. 1. Loosely, clustering is an excess of areal units of a particular category (or categories) in part(s) of the study area. Clustering can occur at different distances. For example, we might have small clusters randomly spread throughout the study area or small clusters which are themselves clustered together. Clusters may consist of areal units in a single category (homogeneous clusters, see Fig. 1b and c) or multiple categories (heterogeneous clusters, see Fig. 1d, e and f). Dispersion occurs when areal units in a particular category occur at semi-regular distances, that is, they are further apart and more regularly spaced than what would be expected under a stationary, independent process (see Fig. 1g). A CAP may exhibit multiple spatial patterns. For example, one category may exhibit clustering while another exhibits dispersion. Alternately, the same category may exhibit clustering at small distances and dispersion at large distances (that is, the clusters may occur at regularly spaced intervals). We wish to develop a hypothesis test which can detect different spatial patterns in different categories and at different distances. To accomplish this, we extend to the positive area proportion function (PAPF) developed by Self et al. (2023) to categorical data.

2.2. Review of the positive area proportion function

The PAPF was developed for binary areal data, which is a special case of a CAP with $K = 2$. For $\mathscr{S} \subset \mathscr{A}$, allow $A(\mathscr{S})$ to be the area of $\mathscr{S}$, and define $\mathcal{N}_k(\mathscr{S}) = \sum_{i=1}^{N} A(\mathscr{S} \cap a_i) I(Y_i = k)$. Thus, $\mathcal{N}_k(\mathscr{S})$ is the amount of area in $\mathscr{S}$ which falls into areal units of

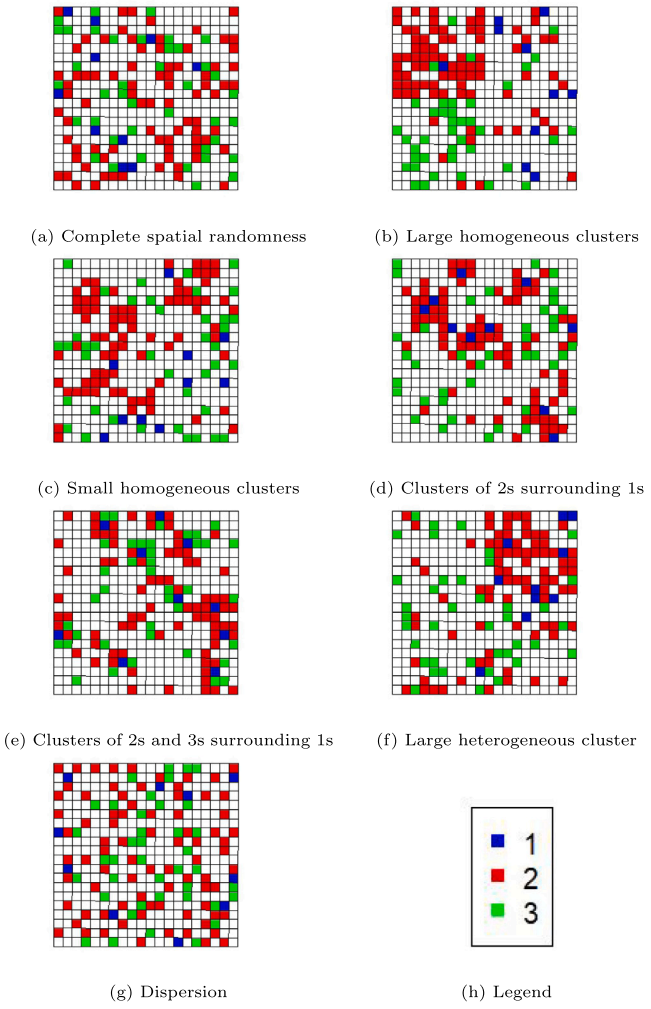


Fig. 1. The figure displays examples of data generated on study area $\mathscr{A}_1$ under each data generation mechanism.

category k. For a CAP with $K = 2$, the PAPF for an areal unit a_i at a distance $r > 0$ is defined by Self et al. (2023) as

$$M(r,i) = E\left\{ \frac{\mathscr{N}_1[c(\boldsymbol{\ell}_i, r) \cap a_i^c] + A[c(\boldsymbol{\ell}_i, r) \cap a_i]}{A[c(\boldsymbol{\ell}_i, r) \cap \mathscr{A}]} \cdot \left(\frac{\mathscr{N}_1(\mathscr{A})}{A(\mathscr{A})} \right)^{-1} | n_1(\boldsymbol{Y}) > 0 \right\},$$

where $\boldsymbol{\ell}_i$ is the centroid of a_i, $c(\boldsymbol{\ell}, r)$ denotes the circle centered at $\boldsymbol{\ell}$ with radius r, and $n_k(\boldsymbol{Y}) = \sum_{i=1}^{N} I(Y_i = k)$. The first term is the proportion of the circle of radius r centered at $\boldsymbol{\ell}_i$ falling into $\mathscr{A}$ that falls into a_i or falls into areal units with $Y_j = 1$. Using the terminology of Self et al. (2023), in which areal units with $Y_i = 1$ were referred to as *positive units*, $M(r,i)$ is the proportion of the circle of radius r centered at $\boldsymbol{\ell}_i$ which either falls into unit a_i or is positive, rescaled by the total proportion of the study area which is positive. Note that, if part of $c(\boldsymbol{\ell}_i, r)$ falls outside of $\mathscr{A}$ then the proportion is taken relative that part of the circle falling inside the study area. The *positive area proportion function at a distance r* is defined as

$$M(r) = E\left[\frac{1}{n_1(\boldsymbol{Y})} \sum_{i=1}^{N} M_1(r,i) I(Y_i = 1) | n_1(\boldsymbol{Y}) > 0 \right]$$

which is simply the expected value of the average of $M(r,i)$ taken over all positive units.

2.3. The category k-j proportion function

We define the *categorical k-j proportion function for unit i at a distance r* as

$$M_{kji}(r) = E\left\{ \frac{\mathscr{N}_j[c(\boldsymbol{\ell}_i, r)]}{A[c(\boldsymbol{\ell}_i, r) \cap \mathscr{A}]} \cdot \left(\frac{\mathscr{N}_j(\mathscr{A})}{A(\mathscr{A})} \right)^{-1} \Big| Y_i = k \right\},$$

where we define $0/0 = 0$ (as $\mathcal{N}_j(\mathscr{A})$ may be 0). We will use $M_{kji}(r)$ to assess clustering of category j relative to category k by considering M_{kji} for areal units a_i in category k. For such units, the first term is the proportion of the circle of radius r centered at $\boldsymbol{\ell}_i$ which falls into units in category j, where the proportion is taken relative to the area of the circle falling inside the study area.

Define the *categorical k-j proportion function at a distance r* as

$$M_{kj}(r) = E\left[\frac{1}{n_k(\boldsymbol{Y})}\sum_{i=1}^{N} M_{kji}(r) I(Y_i = k)\right],$$

where we again take $0/0 = 0$. Thus $M_{kj}(r)$ is the expected value of the average of $M_{kji}(r)$ over all units a_i where $Y_i = k$. When $k = j$, $M_{kk}(r)$ quantifies the proportion of the area near a unit in category k that is also in category k. Large values of $M_{kk}(r)$ suggest homogeneous clusters of category k and small values indicate that dispersion of category k units. When $k \neq j$, $M_{kj}(r)$ quantifies the amount of area in category j near category k units and can be used to detect heterogeneous clusters composed of units in categories k and j. (See Section 3 for more details).

For a CAP realization with $\boldsymbol{Y} = \boldsymbol{y}$ and $y_i = k$, we can estimate $M_{kji}(r)$ with

$$\widehat{M}_{kji}(r, \boldsymbol{y}) = \frac{\mathcal{N}_j[c(\boldsymbol{\ell}_i, r)]}{A[c(\boldsymbol{\ell}_i, r) \cap \mathscr{A}]} \cdot \left(\frac{\mathcal{N}_j(\mathscr{A})}{A(\mathscr{A})}\right)^{-1}.$$

Note that $E\left[\widehat{M}_{kji}(r, i, \cdot) | Y_i = k\right] = M_{kji}(r, i)$. We can then estimate $M_{kj}(r)$ with

$$\widehat{M}_{kj}(r, \boldsymbol{y}) = \frac{1}{n_k(\boldsymbol{y})}\sum_{i=1}^{N} \widehat{M}_{kji}(r, \boldsymbol{y}) I(y_i = k).$$

Allow $M^{(0)}_{kji}(r)$ and $M^{(0)}_{kj}(r)$ to denote $M_{kji}(r)$ and $M_{kj}(r)$ for a stationary, independent process, respectively. It can be shown that for a stationary, independent process, $E[\widehat{M}_{kj}(r)] = M^{(0)}_{kj}(r)$. See Web Appendix A for additional details.

2.4. Hypothesis testing using the categorical positive area proportion function at specific distances

The categorical positive area proportion function (CPAPF) can be used to detect spatial patterns in categorical areal data. Here we focus our attention on detecting clustering and dispersion. When clustering consists of an excess of a single type of areal unit, we will refer to the clustering as homogeneous. See Fig. 1b, c, Web Figure 1b and Web Figure 1c for examples of homogeneous clustering.

Consider testing the null hypothesis of a stationary, independent CAP versus the alternative hypothesis that the process exhibits homogeneous clustering with respect to category k at a distance r. Define:

$$T_k(r, \boldsymbol{y}) = \widehat{M}_{kk}(r, \boldsymbol{y}) - M^{(0)}_{kk}(r, \boldsymbol{n}),$$

where $\boldsymbol{n} = [n_1(\boldsymbol{y}), \ldots, n_K(\boldsymbol{y})]'$,

$$M^{(0)}_{kj}(r, \boldsymbol{n}) = E\left[\frac{1}{n_k(\boldsymbol{Y})}\sum_{i=1}^{N} M^{(0)}_{ijk}(r, \boldsymbol{n}) I(Y_i = k) | \boldsymbol{n}(\boldsymbol{Y}) = \boldsymbol{n}\right]$$

and

$$M^{(0)}_{kji}(r, \boldsymbol{n}) = E\left\{\frac{\mathcal{N}_j[c(\boldsymbol{\ell}_i, r)]}{A[c(\boldsymbol{\ell}_i, r) \cap \mathscr{A}]} \cdot \left(\frac{\mathcal{N}_j(\mathscr{A})}{A(\mathscr{A})}\right)^{-1} | Y_i = k, n(\boldsymbol{Y}) = \boldsymbol{n}\right\}$$

for a stationary independent CAP. Note that $M^{(0)}_{kj}(r, \boldsymbol{n})$ is analogous to $M^{(0)}_{kj}(r)$, conditional on a fixed number of areal units in each category with $Y_i = k$. This additional conditioning reduces the variability in the null distribution of the test statistic. When there is homogeneous clustering in category k, there will be more category k units in close proximity to other category k units than expected under a stationary, independent process, inflating the value of $\widehat{M}_{kk}(r, \cdot)$. Thus an upper-tailed test may be used to detect homogeneous clustering.

Similarly, a lower tailed test using the same test statistic can be used to detect dispersion in category k. We say that a CAP exhibits dispersion with respect to category k if areal units in category k are further apart or more regularly spaced than would be expected under a stationary, independent process. In general, there will be less area in category k units near other category k units for a dispersed process, reducing the value of $\widehat{M}_{kk}(r, \cdot)$.

The CPAPF can also be used to detect heterogeneous clustering, defined as an excess of two or more categories in the same part of the study area. Heterogeneous clustering can take many forms. For example, there might be an excess of areal units in two different categories evenly mixed across the same portion of the study area (see Fig. 1f) or an excess of areal units of one category surrounding units in another (see Fig. 1d). Suppose we wish to test a null hypothesis of a stationary, independent process versus an alternative hypothesis of heterogeneous clustering involving categories k and j. Consider the following test statistic

$$T_{kj}(r, \boldsymbol{y}) = \widehat{M}_{kj}(r, \boldsymbol{y}) - M^{(0)}_{kj}(r, \boldsymbol{n}).$$

Note that in general, $T_{kj}(r, \boldsymbol{y}) \neq T_{jk}(r, \boldsymbol{y})$. When there is an excess of both category k and category j units in the same part of the study area (e.g. Fig. 1f), both $T_{kj}(r, \cdot)$ and $T_{jk}(r, \cdot)$ tend to be inflated. However, when there is an excess of category j surrounding category k (see Fig. 1d), we would expect T_{kj} and T_{jj} to be more inflated than T_{jk}, though the later may still exhibit some inflation.

The distribution of $T_{kj}(r, \boldsymbol{Y})$ for a stationary, independent CAP does not belong to a known distributional family but can be estimated with Monte Carlo simulations. As we condition on $\boldsymbol{n}(\boldsymbol{Y})$, the data for the Monte Carlo procedure would be generated under the so-called *random labeling hypothesis*, by randomly reassigning the y_i values to the areal units. While $M_{kj}^{(0)}(r, \boldsymbol{n})$ can be computed exactly, doing so involves computing N sums with $\binom{N}{\boldsymbol{n}}$ terms each. Furthermore, computing the terms in each sum involves computing the area of polygon intersections, which is generally computationally expensive. As an alternative, we propose approximating $M_{kj}^{(0)}(r, \boldsymbol{n})$ using the same set of Monte Carlo simulations used to approximate the distribution of $T_{kj}(r, \boldsymbol{Y})$. That is, we take

$$\widehat{M}_{kj}^{(0)}(r, \boldsymbol{n}) = \frac{1}{G} \sum_{g=1}^{G} \widehat{M}_{kj}(r, \boldsymbol{y}^{(g)})$$

where G is the number of Monte Carlo simulations and $\boldsymbol{y}^{(g)}$ is the data generated for the gth simulation.

2.5. A global hypothesis test for type I error rate control

In many applications, the optimal radius for assessing spatial patterns may be unknown. The particular categories involved in the spatial patterns may also be unknown. In these cases, one could apply the CPAPF test at a variety of radii and/or to a variety of different categories. However this approach introduces a multiple testing problem and the potential for an inflated type I error rate. As an alternative, we propose a global test for clustering across a range of radii and/or categories.

Let $\mathscr{R} = \{r_1, \ldots, r_R\}$ be the set of radii under consideration and $\mathscr{K} = \{(k_1, j_1), \ldots (k_L, j_L)\}$ be the set of ordered pairs denoting the category pairs of interest. We then wish to use $T_{kj}(r, \cdot y)$ to assess spatial patterns for all $r \in \mathscr{R}$ and all $(k, j) \in \mathscr{K}$. Define

$$T_C(\boldsymbol{y}, \mathscr{R}, \mathscr{K}) = \max_{r \in \mathscr{R}, (k,j) \in \mathscr{K}} \left\{ \frac{T_{kj}(r, \boldsymbol{y})}{S_{kjr}} \right\}$$

and

$$T_D(\boldsymbol{y}, \mathscr{R}, \mathscr{K}) = \min_{r \in \mathscr{R}, (k,j) \in \mathscr{K}} \left\{ \frac{T_{kj}(r, \boldsymbol{y})}{S_{kjr}} \right\},$$

where $S_{kjr} = \text{var}[T_{kj}(r, \cdot)]^{1/2}$ for a stationary, independent CAP conditional on $\boldsymbol{n}(\boldsymbol{Y}) = \boldsymbol{n}$. We can estimate the null distribution of $T_C(\cdot, \mathscr{R}, \mathscr{K})$ by generating data under the random labeling hypothesis, computing $T_{kj}(r, \cdot)/S_{kjr}$ for each $(k, j) \in \mathscr{K}$ and each $r \in \mathscr{R}$ and taking the maximum over all (k, j) and all r for each generated $\boldsymbol{y}$. Note that S_{jkr} can be estimated from the same Monte Carlo procedure used to estimate $M_{kj}^{(0)}(r, \cdot)$. The null distribution of $T_D(\cdot, \mathscr{R}, \mathscr{K})$ can be estimated similarly by taking the minimum rather the maximum. We can use an upper-tailed test on $T_C(\boldsymbol{y}, \mathscr{R}, \mathscr{K})$ to detect clustering and a lower-tailed test on $T_D(\boldsymbol{y}, \mathscr{R}, \mathscr{K})$ to detect dispersion. Testing for both clustering and dispersion introduces a multiple testing problem which can be controlled for by performing both tests and applying a Bonferroni correction for two tests.

3. Simulation study

3.1. Simulation specifications

We compare the performance of the CPAPF method to the that of a multinomial scan statistic using a simulation study. We also consider dichotomizing the data and applying global Moran's I statistic and Getis–Ord general G statistic to detect certain types of clustering. We consider $K = 4$ categories and two study areas:

$\mathscr{A}_1$: A 20 by 20 regular grid.

$\mathscr{A}_2$: 738 contiguous counties along the Eastern Seaboard of the United States (US).

We chose $\mathscr{A}_1$ to assess the performance of our method when the areal units are regularly shaped and spaced, while study area $\mathscr{A}_2$, consisting of all the counties in Middle and South Atlantic states, including New Jersey, New York, Delaware, Pennsylvania, Maryland, West Virginia, Virginia, South Carolina, North Carolina, Georgia, and Florida, plus the District of Columbia, assesses the performance on the method when the areal units are irregularly shaped and the study area is non-convex. We consider seven different types of spatial patterns, arising from seven different data generation mechanisms (DGMs):

(a) A stationary, independent process (the null hypothesis)
(b) Large homogeneous clusters of 1s, 2s and 3s
(c) Small homogeneous clusters of 2s
(d) Clusters of 2s surrounding 1s
(e) Clusters of 2s and 3s surrounding 1s
(f) Large heterogeneous cluster of 1s and 2s
(g) Dispersion of 2s

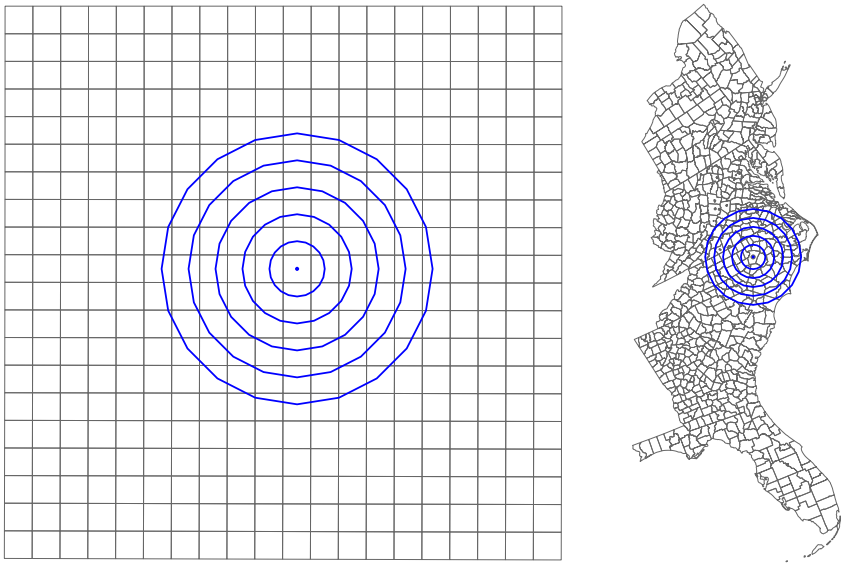


Fig. 2. The figure depicts study area $\mathscr{A}_1$ (left) and $\mathscr{A}_2$ (right) along with the circles having each radius in $\mathscr{R}_1$ and $\mathscr{R}_2$.

Examples of each data generation mechanism on $\mathscr{A}_1$ are given in Fig. 1. Web Figure 1 contains similar examples for study area $\mathscr{A}_2$. For additional details on how each type of data was generated, see Web Appendix B.

Categorical data commonly consists of an 'absence' category and several 'presence' categories. For example, in infectious disease epidemiology, individuals can be classified as non-infected (absence) or having one of several possible disease sub-types (presence) categories. In our motivating data application, land parcels are divided into those that are held as conservation easements (CEs) and those that are not. The CE parcels are further classified as agricultural, environmental, built, open or other, resulting in five presence categories and one absence category. Often, the absence category is the most common and we are not interested its clustering behavior. To mimic this scenario, we treat $k = 0$ as the absence category and assess the remaining categories $k = 1, 2, 3$ as the presence categories. For each study area, we assume the absence category comprises approximately two thirds of the observations, and take $\boldsymbol{n}_1 = (270, 10, 80, 40)'$ and $\boldsymbol{n}_2 = (500, 18, 147, 73)'$, which correspond to having approximately 2.5%, 20% and 10% of the observations in categories 1, 2, and 3, respectively.

For each study area $i = 1, 2$, we consider a set of five equally spaced radii $\mathscr{R}_i = \{r_1, \ldots, r_5\}$ shown in Fig. 2. Following Self et al. (2023), the smallest radius was chosen to capture a single unit and its first order neighbors and the largest radius was chosen to be approximately one fourth of the width of the study area. The null distributions of the CPAPF test statistics at all radii and the multinomial scan statistic were estimated with 500 Monte Carlo simulations with data generated under the random labeling hypothesis. The set of zones for the multinomial scan statistic was taken to be the set of all circles $c(\ell_i, r)$ where $i \in \{1, 2, \ldots, N\}$ and $r \in \mathscr{R}_i$. These zones were chosen to make the CPAPF and scan statistic tests as comparable as possible. To apply the global Moran's I and Getis–Ord general G tests to detect homogeneous (k, k) clustering, we first dichotomize the data so that units in one category k are 1 and other units are 0. We then apply the global Moran's I statistic and the global Getis–Ord general G statistic to the dichotomized data using adjacency-based weights. An upper tailed test with global Moran's I and Getis–Ord general G detects positive spatial autocorrelation and high-high clustering, respectively, so we expect these tests to detect homogeneous clustering. A lower tailed Moran's I test detects negative spatial autocorrelation and should be able to detect dispersion.

3.2. Simulation results

Table 1 and Web Table 1 display the simulation study results for study area $\mathscr{A}_1$; Table 2 and Web Table 2 display results for $\mathscr{A}_2$. For each $(k, j) \in \{1, 2, 3\} \times \{1, 2, 3\}$ (here $\times$ denotes the set product) we report the empirical rejection rate (ERR) using the test statistic $T_{kj}(r, \cdot)$ for all $r \in \mathscr{R}_i$ using a two-tailed test for data generated under DGM a, an upper-tailed test for DGMs b-f and a lower-tailed test for DGM g. All tests were at the $\alpha = 0.05$ significance level. For each pair (k, j), we also report empirical rejection rate for the global test over all radii (e.g., using $T_C\{\cdot, \mathscr{R}_i, \mathscr{K} = (j, k)\}$ or $T_D\{\cdot, \mathscr{R}_i, \mathscr{K} = (j, k)\}$ as the test statistic). Finally, we report the empirical rejection rate for the global test over all radii and all categories, (e.g., using $T_C\{\cdot, \mathscr{R}_i, \mathscr{K}\}$ or $T_D\{\cdot, \mathscr{R}_i, \mathscr{K}\}$ for $\mathscr{K} = \{0, 1, 2, 3\} \times \{0, 1, 2, 3\}$ as the test statistic). For DGM a, the ERR for all tests arises from conducting a lower tailed test on T_D and an upper tailed test on T_C and applying a Bonferroni correction for 2 tests. For DGMs b-f, an $\alpha = 0.05$ upper-tailed test was performed on T_C. For DGM g, an $\alpha = 0.05$ lower-tailed test was performed on T_D. The ERR of an $\alpha = 0.05$ upper-tailed test using the multinomial scan statistic is also reported for each DGM. For global Moran's I and Getis–Ord general G, we report the ERR for an $\alpha = 0.05$ two-tailed test for DGM a and an upper-tailed test for DGMs b-f. For DMG g, we apply a lower-tailed test for Moran's I statistic only.

Table 1
Simulation study results for study area for the regular grid for the CPAPF method, multinomial spatial scan statistic (SS), global Moran's I statistic (MI) and global Getis–Ord general G statistic (GG). All the empirical rejection rates (ERR) reported in this table should be interpreted as empirical power.

DGM	(k,j)	ERR									
		Global	r_1	r_2	r_3	r_4	r_5	Global	SS	MI	GG
b	1,1	12.0	8.0	15.0	26.0	39.0	48.0	99.0	74.6	32.2	29.0
	2,2	99.0	100.0	100.0	100.0	100.0	100.0			100.0	100.0
	3,3	86.0	0.95	98.0	100.0	100.0	100.0			99.4	98.6
c	2,2	99.4	100.0	100.0	99.8	92.2	80.0	90.8	15.2	100.0	100.0
d	1,1	5.2	4.8	5.6	4.6	4.8	4.6	100.0	32.4	5.2	4.2
	2,2	75.6	85.2	94.6	96.4	84.8	69.4			86.0	87.6
	1,2	100.0	100.0	100.0	100.0	98.2	91.2			–	–
	2,1	100.0	100.0	100.0	100.0	96.2	82.6			–	–
e	2,2	31.2	30.2	48.0	61.6	46.4	37.0	100.0	55.2	32.8	38.8
	3,3	21.0	32.8	44.8	56.0	42.4	38.2			36.4	36.4
	1,2	100.0	100.0	100.0	100.0	85.2	72.6			–	–
	2,1	100.0	100.0	100.0	99.2	77.0	58.0			–	–
	1,3	99.8	99.6	100.0	98.0	74.6	62.8			–	–
	3,1	99.8	99.6	100.0	97.6	75.2	58.0			–	–
	2,3	69.6	66.0	76.2	82.2	64.2	48.8			–	–
	3,2	72.0	64.6	76.0	85.2	72.2	57.2			–	–
f	1,1	16.0	19.2	35.0	49.4	58.0	65.2	97.8	56.6	35.0	30.8
	2,2	99.8	100	100	100	100	100			100.0	100.0
	1,2	87.8	91.0	98.4	99.6	99.6	99.6			–	–
	2,1	83.2	91.6	98.0	99.4	99.4	99.6			–	–
g	2,2	100.0	100.0	100.0	82.8	84.0	70.2	100.0	–	100.0	–

The ERR for a given spatial pattern should be interpreted as the empirical power of the method for DGMs that produce data exhibiting that spatial pattern. Otherwise the ERR should be interpreted as the empirical type I error rate. Tables 1 and 2 contain the results for the pairs (k, j) which are expected to exhibit a spatial pattern under each DGM; all ERRs in these tables are empirical powers. Web Tables 1 and 2 contain results for the remaining (k, j); these ERRs should be interpreted as empirical type I error rates. We note that under DGM a, we are generating data under the null hypothesis and a rejection of the null hypothesis for any type of spatial pattern is a type I error; results for this DGM therefore appear only in Web Tables 1 and 2. For a full description of which spatial patterns we expect to manifest under which DGMs, please see Web Appendix B.

Results for DGM a show that the CPAPF tests maintain their nominal type I error rates. While the empirical type I error rate varies slightly around the nominal 5% level of the test (e.g. 1%–9%, with most ERRs between 3%–7%), this is likely due to Monte Carlo error in the estimation of the null distribution. In practice, concerns about Monte Carlo error can be alleviated by increasing the number of Monte Carlo simulations used to estimate the null distribution. The global CPAPF outperformed the multinomial spatial scan statistic across all scenarios. The method has high power to detect a variety of different types of clustering. For the large clustering scenarios (2 and 6), the radii-specific tests generally have more power when the radius is large, with this effect being more pronounced for the US counties. Similarly, for the small clustering scenarios (3, 4, and 5), the power of the radii-specific tests is generally higher when the radius is small. Finally, we note that the method appears to be relatively robust against falsely detecting the wrong type of clustering, as the rejection rates for pairs (k, j) not involved in pattern creation reported in Web Tables 1 and 2 are all relatively low.

The global Moran's I and Getis–Ord general G approaches applied to dichotomized data tended to have slightly higher empirical power (5%–10%) to detect homogeneous clustering than the global CPAPF test for the corresponding type of clustering. However these approaches would require a Bonferroni correction (or similar approach) to control for multiple testing if we wish to use them for a general test for homogeneous clustering, which would likely decrease their power beyond that reported here. The CPAPF method can be used to perform a general test for any type of homogeneous clustering simply by applying a global test with $\mathscr{K} = \{(k, k) : k = 0, 1, \ldots, K\}$. The global Moran's I and Getis–Ord general G approaches also cannot be used to detect heterogeneous clustering and do not provide information on the distance at which the clustering is occurring. Finally, it is worth nothing that the CPAPF method and the multinomial spatial scan statistic appear to have different strengths. The CPAPF method has higher power and is able to distinguish between different types of clustering and dispersion. The scan statistic cannot differentiate between different types of clustering and cannot detect dispersion. However, the scan statistic can identify the location of any detected cluster(s), which the CPAPF method cannot do.

4. Data application

Our motivating dataset consists of the 112,819 parcels in Boulder County Colorado in 2008; 117 of these parcels were held as conservation easements (CEs). These data are displayed in Fig. 3. CEs serve as voluntary legal mechanisms to designate land for diverse objectives, encompassing biological and ecological preservation, recreational access, historic preservation, and more Federico Cheever and McLaughlin (0000). A CE is executed through a formal agreement between a landowner and either

Table 2
Simulation study results for the US counties for the CPAPF method, multinomial spatial scan statistic (SS), global Moran's I statistic (MI) and global Getis–Ord general G statistic (GG). All the empirical rejection rates (ERR) reported in this table should be interpreted as empirical power.

DGM	(k,j)	ERR									
		Global	r_1	r_2	r_3	r_4	r_5	Global	SS	MI	GG
b	1,1	24.8	60.8	67.0	77.4	80.0	80.0	100.0	15.0	29.8	36.2
	2,2	100.0	100.0	100.0	100.0	100.0	100.0			100.0	100.0
	3,3	100.0	99.8	100.0	100.0	100.0	100.0			97.8	98.2
c	2,2	94.2	99.8	98.6	86.4	68.0	50.4	62.0	3.2	100.0	100.0
d	1,1	4.6	4.2	3.6	5.2	4.8	6.6	100.0	3.2	9.4	9.6
	2,2	55.8	70.6	80.2	61.6	47.0	34.0			66.6	88.4
	1,2	100.0	100.0	100.0	99.8	93.8	82.0			–	–
	2,1	99.2	100.0	99.8	92.0	65.8	42.6			–	–
e	2,2	16.4	17.2	25.6	18.8	16.0	11.6	100.0	2.8	14.0	29.4
	3,3	17.6	31.2	40.2	32.6	23.8	18.8			23.6	34.4
	1,2	99.8	100.0	99.2	90.2	70.4	48.2			–	–
	2,1	89.6	99.2	91.2	62.4	38.6	22.2			–	–
	1,3	99.6	100.0	100.0	90.8	69.4	57.0			–	–
	3,1	95.4	100.0	97.2	74.6	48.6	37.6			–	–
	2,3	27.2	30.2	50.8	42.2	30.8	26.0			–	–
	3,2	29.6	35.4	52.2	45.0	32.0	21.4			–	–
f	1,1	23.6	11.6	21.8	38.4	52.0	64.2	100.0	1.8	29.8	36.2
	2,2	100.0	100.0	100.0	100.0	100.0	100.0			100.0	100.0
	1,2	98.6	95.0	99.4	99.6	99.8	100.0			–	–
	2,1	81.8	78.4	94.6	96.0	96.8	97.2			–	–
g	2,2	100.0	100.0	97.2	70.2	47.8	31.2	100.0	–	100.0	–

a land trust or a governmental entity. This agreement enforces deed restrictions on the property, restricting its development and utilization, as mutually consented to by the landowner. In return, the landowner can become eligible for federal, state, or local tax benefits if they did not sell the CE. The specific purposes and associated tax benefits are contingent upon jurisdictional statutes and the objectives of the involved land trust or organization. The Internal Revenue Service (IRS) has established guidelines and requirements that must be met for a CE to qualify for tax benefits. A pivotal federal regulation addressing conservation easements is Internal Revenue Code Section 170(h), introduced as a component of the Tax Reform Act of 1976 (Hollingshead, 1996).

Perpetual CEs are unique land use cases in that they ensure the protection of their conservation purposes (agricultural, open-space, biological, or historical) in perpetuity. More awareness of where and what types of CE clusters exist in a county could inform local land use planning processes to aid in the creation of critical masses of agricultural protected lands or biologically protected lands. For example, it could highlight areas to consider for future industry-supportive infrastructure development and areas for more robust farmland preservation policies, areas to consider for additional protection from residential, commercial, and industrial development, or areas to consider for additional investment in community access to nature.

Documentation describing all CEs filed (or amended) between 1997 and 2008 was obtained from the Boulder County Clerk and Recorder Division. We developed a CE coding instrument with sections related to CE disposition, holder typology, public access, and other nominal variables for analysis. The most relevant data for our analysis is the set of CE reasons/purposes. There were 31 reason codes, the first three of which relate to direct quotation of the federal tax code definitions for conservation purposes (26 U.S.C. 170(h)(4)(A)(i) – (iii)). The remainder describes other uses such as open space preservation, agricultural viability/livestock, etc.

Using this instrument, we manually coded the CE documents, which range in length between 5–70 pages apiece. The research team relied on 4 to 5 coders (including the lead researcher), and we used inter-coder reliability checks, as well as weekly sub-team meetings, to verify the coding accuracy and to address emergent anomalies and interpretation questions.

For this analysis, we classified CEs into 5 broad categories based on the 31 CE reason codes described above: biological, agricultural, open, built and other. For more details on the classification system, see Web Appendix C. For this analysis, we took $\mathscr{R} = \{1\text{ km}, 3.25\text{ km}, 5.5\text{ km}, 7.75\text{ km}, 10\text{ km}\}$ and $\mathscr{K}$ as the set of all ordered reason pairs whose components are taken from the categories biological, agricultural, open and built. All hypothesis tests were conducted with a significance level of $\alpha = 0.05$, and the null distribution of all test statistics was estimated with 500 Monte Carlo replications.

A global two-sided test was performed using $T_C(\boldsymbol{y}, \mathscr{R}, \mathscr{K})$ and $T_D(\boldsymbol{y}, \mathscr{R}, \mathscr{K})$ while applying a Bonferroni correction for two tests. The null hypothesis was rejected in favor of clustering. To determine which specific types of CEs exhibit spatial patterns, two-sided tests were performed using $T_C(\boldsymbol{y}, \mathscr{R}, \{(k,j)\})$ and $T_D(\boldsymbol{y}, \mathscr{R}, \{(k,j)\})$ for all $(k,j) \in \mathscr{K}$. The following reason pairs exhibited significant clustering: biological with biological, agricultural with agricultural, open with open, built with built, built surrounded by agricultural, and built surrounded by open, while agricultural surrounded by biological and open surrounded by biological exhibited dispersion.

To determine the distances at which these patterns were occurring, we performed a one-sided test for clustering using $T_C(\boldsymbol{y}, \{r_i\}, \{(k,j)\})$ for all $r_i \in \mathscr{R}$ and all reason pairs for which the corresponding global test was rejected in favor of clustering. Significant clustering of agricultural with agricultural, open with open, built surrounded by agricultural and built surrounded by

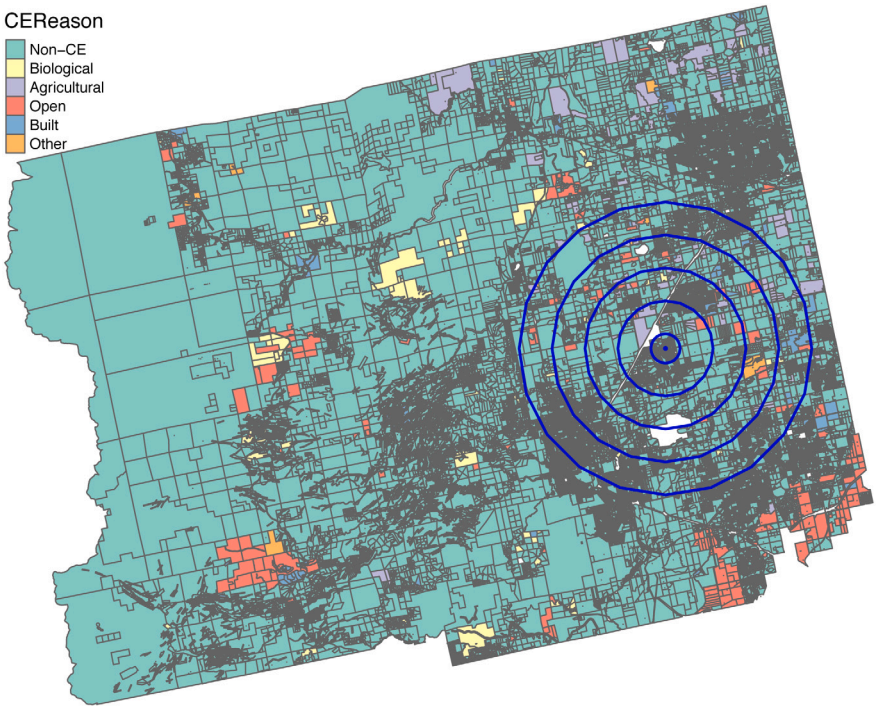


Fig. 3. The figure depicts the 112,819 parcels in Boulder County, Colorado in 2008. The radii at which the CPAPF test statistics were computed are shown in blue.

open was detected at all radii in $\mathscr{R}$. Significant clustering of biological with biological was detected only for the smallest radii and significant clustering of built with built was detected only for the smallest three radii. We performed a one-sided test for dispersion using $T_D(\boldsymbol{y}, \{r_i\}, \{(i,j)\})$ for all $r_i \in \mathscr{R}$ and all reason pairs for which the corresponding global test was rejected in favor of dispersion. Agricultural surrounded by biological and open surrounded by biological both exhibited dispersion at all radii except the smallest. Full numerical results for all hypothesis tests, including critical values and test statistics, are provided in the Supplementary Table in Web Tables 3–5.

5. Conclusion

In this work, we developed a hypothesis test to detect spatial patterns at various distances in categorical areal data. Our hypothesis testing procedure is based on the categorical positive area proportion function, an extension of the positive area proportion function developed by Self et al. (2023). Our method performs well on simulated and real data and can detect a wide variety of spatial patterns with higher power than the comparable multinomial spatial scan statistic. Our method can be used to perform a global test to detect the presence of any spatial patterns at any set of distances, or can be used to test for specific patterns at specific distances. To facilitate the widespread use of the CPAPF hypothesis test, code that implements our method is publicly available https://github.com/scwatson812/Cat_PAPF.

Our method presents many areas for potential future work. The Monte Carlo procedure for estimating the null distribution of our test statistic can be computationally expensive when the number of radii or areal units considered is large. Developing a more computationally efficient means of estimating the null distribution is a worthwhile area for future work. Extending our method to other types of areal data, such as ordinal or numeric data, may also be worthwhile. As the volume of spatial grows, so will the need for statistical methods to detect nuanced spatial patterns.

Funding

SS and AM were supported in part by the Research Center for Child Well-Being, funded by the National Institutes of Health [NIGMS P20GM130420]. SS, AO, DW, and CD were supported in part by the National Science Foundation, United States [CNH-L 1518455]. XZ was supported by the Companion Animal Parasite Council. The funding sources played no role in study design, data collection, data analysis, or manuscript publication.

CRediT authorship contribution statement

Stella Self: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Visualization. **Xingpei Zhao:** Methodology, Data curation, Visualization, Writing – review & editing. **Anja Zgodic:** Conceptualization, Investigation, Data curation, Writing – review & editing. **Anna Overby:** Conceptualization, Investigation, Resources, Data curation, Writing – review & editing. **David White:** Resources, Writing – review & editing, Supervision. **Alexander C. McLain:** Conceptualization, Data curation, Writing – review & editing, Visualization, Supervision, Funding acquisition. **Caitlin Dyckman:** Conceptualization, Investigation, Resources, Data curation, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

The Supplementary Material contains proofs, simulation study results, and the CE reason classification system.
Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spasta.2024.100839.

References

Anselin, Luc, 1995. Local indicators of spatial association—LISA. Geogr. Anal. 27 (2), 93–115. http://dx.doi.org/10.1111/j.1538-4632.1995.tb00338.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1995.tb00338.x.

Cançado, André L.F., Fernandes, Lucas B., da Silva, Cibele Q., 2017. A Bayesian spatial scan statistic for zero-inflated count data. Spat. Statist. 20, 57–75. http://dx.doi.org/10.1016/j.spasta.2017.01.005, URL https://www.sciencedirect.com/science/article/pii/S2211675317300374.

Federico Cheever, McLaughlin, Nancy A., An introduction to conservation easements in the United States: A simple concept and a complicated mosaic of law. J. Law Prop. Soc..

Getis, Arthur, Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. Geogr. Anal. 24 (3), 189–206. http://dx.doi.org/10.1111/j.1538-4632.1992.tb00261.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1992.tb00261.x.

Hollingshead, J., 1996. Conservation easements: A flexible tool for land preservation. Environ. Lawyer 3, 319–361.

Huang, Lan, Kulldorff, Martin, Gregorio, David, 2007. A spatial scan statistic for survival data. Biometrics 63 (1), 109–118. http://dx.doi.org/10.1111/j.1541-0420.2006.00661.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2006.00661.x.

Jung, Inkyung, Kulldorff, Martin, Klassen, Ann C., 2007. A spatial scan statistic for ordinal data. Stat. Med. 26 (7), 1594–1607. http://dx.doi.org/10.1002/sim.2607, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2607.

Jung, Inkyung, Kulldorff, Martin, Richard, Otukei John, 2010. A spatial scan statistic for multinomial data. Stat. Med. 29 (18), 1910–1918. http://dx.doi.org/10.1002/sim.3951, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3951.

Kulldorff, Martin, 1997. A spatial scan statistic. Comm. Statist. Theory Methods 26 (6), 1481–1496. http://dx.doi.org/10.1080/03610929708831995.

Kulldorff, Martin, Huang, Lan, Konty, Kevin, 2009. Scan statistic for continuous data based on the normal probability model. Int. J. Health Geogr. 8, 58. http://dx.doi.org/10.1186/1476-072X-8-58.

Majumder, Suman, Coull, Brent A., Welch, Jessica L. Mark, Riviere, Patrick J. La, Dewhirst, Floyd E., Starr, Jacqueline R., Lee, Kyu Ha, 2022. arXiv URL https://doi.org/10.48550/arXiv.2202.04198.

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37 (1/2), 17–23, URL http://www.jstor.org/stable/2332142.

Neill, Daniel B., Moore, Andrew W., Cooper, Gregory F., 2005. A Bayesian spatial scan statistic. In: Proceedings of the 18th International Conference on Neural Information Processing Systems. NIPS '05, MIT Press, Cambridge, MA, USA, pp. 1003–1010.

Neyman, Jerzy, Scott, Elizabeth L., 1958. Statistical approach to problems of cosmology. J. R. Stat. Soc. Ser. B Stat. Methodol. 20 (1), 1–43, URL http://www.jstor.org/stable/2983905.

Self, Stella, Overby, Anna, Zgodic, Anja, White, David, McLain, Alexander, Dyckman, Caitlin, 2023. A hypothesis test for detecting distance-specific clustering and dispersion in areal data. Spat. Statist. 55, 100757. http://dx.doi.org/10.1016/j.spasta.2023.100757, URL https://www.sciencedirect.com/science/article/pii/S2211675323000325.