HHS Public Access

Author manuscript

Spat Stat. Author manuscript; available in PMC 2024 June 01.

Published in final edited form as:

Spat Stat. 2023 June ; 55: . doi:10.1016/j.spasta.2023.100757.

A Hypothesis Test for Detecting Distance-Specific Clustering and Dispersion in Areal Data

Stella Self[1,2], **Anna Overby**[3], **Anja Zgodic**[2], **David White**[4], **Alexander McLain**[2,5], **Caitlin Dyckman**[3,5]

[2]Arnold School of Public Health, University of South Carolina, 921 Assembly Street, Columbia, SC 29208, USA

[3]College of Architecture, Arts and Humanities, Clemson University, Fernow Street, Clemson, SC 29634, USA

[4]College of Behavioral, Social and Health Sciences, Clemson University, Epsilon Zeta Dr, Clemson, SC 29634, USA

[5]Shared Last Author

Abstract

Spatial clustering detection has a variety of applications in diverse fields, including identifying infectious disease outbreaks, pinpointing crime hotspots, and identifying clusters of neurons in brain imaging applications. Ripley's K-function is a popular method for detecting clustering (or dispersion) in point process data at specific distances. Ripley's K-function measures the expected number of points within a given distance of any observed point. Clustering can be assessed by comparing the observed value of Ripley's K-function to the expected value under complete spatial randomness. While performing spatial clustering analysis on point process data is common, applications to areal data commonly arise and need to be accurately assessed. Inspired by Ripley's K-function, we develop the *positive area proportion function (PAPF)* and use it to develop a hypothesis testing procedure for the detection of spatial clustering and dispersion at specific distances in areal data. We compare the performance of the proposed PAPF hypothesis test to that

[1]Corresponding Author scwatson@mailbox.sc.edu (Stella Self).

Competing Interests
Declarations of interest: none

CReDiT Author Statement
Stella Self: conceptualization, methodology, software, writing- original draft, writing-review & editing visualization
Anna Overby: conceptualization, investigation, resources, data curation, writing- review & editing
Anja Zgodic: conceptualization, investigation, data curation, writing- review & editing
David White: resources, writing- review & editing, supervision
Alexander McLain: conceptualization, data curation, writing- review & editing, visualization, supervision, funding acquisition
Caitlin Dyckman: conceptualization, investigation, resources, data curation, writing- review & editing, supervision, funding acquisition

Supplementary Material
The Supplementary Material contains derivations of properties of the PAPF (Web Appendix A), additional simulation results (Web Appendix B) and additional data application results (Web Appendix C).

of the global Moran's I statistic, the Getis-Ord general G statistic, and the spatial scan statistic with extensive simulation studies. We then evaluate the real-world performance of our method by using it to detect spatial clustering in land parcels containing conservation easements and US counties with high pediatric overweight/obesity rates.

Keywords

cluster detection; clustering; areal data; Ripley's K-function

1. Introduction

The rapid rise in popularity of geographic information system (GIS) software over the past thirty years has led to an explosion of spatial data and associated analytical methods. Two of the most common types of spatial data are point process data and areal data. Point process data are associated with specific coordinate locations (such as a geocoded addresses), while areal data are associated with spatial regions (such as a counties or census tracts). The data may contain either location information only (e.g., the boundaries of a census tract that the United States Department of Agriculture has designated a food desert due to a paucity of stores selling healthy food) or location information combined with numerical attributes (e.g., the boundaries of a census tract with the number of healthy grocery stores). In this paper, we restrict our attention to areal data that contains only location information, that is, data consisting of a predefined set of spatial regions, some of which possess a characteristic of interest. We consider the set of spatial regions as fixed and only the possession of the characteristic as random. We consider the problem of assessing this type of data for spatial clustering and dispersion, loosely defined as an excess of regions with the characteristic of interest in part(s) of the study area (clustering) or semi-regular placement of such regions (dispersion). Census tracts designated as food deserts, tracts of land with development restrictions, or counties which required individuals to wear a mask in public during the COVID-19 pandemic are all examples of areal data that could be clustered. In this paper, we develop a method for detecting clustering and/or dispersion in areal data at specific distances.

Clustering can occur at different distances. For example, food desert census tracts might be clustered at close distances in metropolitan areas and at larger distances in more rural areas. Additionally, data may exhibit dispersion at one distance and clustering at another. For example, parks may exhibit small scale dispersion (e.g., city parks are unlikely to be within a quarter mile of another park), but large scale clustering (e.g., city parks are likely to be within 5 miles of a another park). In practice, the distance at which clustering and/or dispersion occur are generally informative about which processes may be causing the phenomena. Statistical methods that can detect clustering and/or dispersion at specific distances are desirable.

There are several existing methods to assess areal data for clustering, including the global Moran's I statistic [45], the Getis-Ord general G statistic [28], and the spatial scan statistic [35]. (These methods can also be used for point process data, with some modifications).

However, obtaining the distance at which clustering/dispersion occurs is not straightforward for any of these methods. Ripley's K-function is able to detect clustering/dispersion at specific distances, but it is only suitable for point process data [49, 50, 51]. Ripley's K-function and related variants have been widely used in ecology [30, 39, 37], epidemiology [19, 26, 10], and spatial economics [40, 22, 41]. Despite the suitability issue, Ripley's K-function is commonly (mis)applied to areal data by mapping each spatial region to its centroid. For example, many researchers have attempted to assess spatial patterns in land parcel data using Ripley's K-function [38, 52, 57, 48]. Other researchers have taken public health data associated with a geographical region (e.g., a city or health division) and computed Ripley's K-function using the centroids of the regions [55, 33, 53]. Ripley's K-function has also been used to assess areal data for clustering in a variety of ecological and geological applications [34, 16, 43].

Applying Ripley's K-function to the centroids of spatial regions is particularly problematic when the regions are vastly different sizes. For example, in one of our motivating data applications we wish to determine if land parcels with conservation easements (CEs) are clustered. Under the null hypothesis, all parcels are equally likely to have a CE. Sections of the study area with many small parcels (such as a metropolitan area) will have more parcels with CEs than portions of the study area with many large parcels, simply because there are more parcels per unit area. Put another way, centroids of smaller parcels will appear clustered relative to centroids of larger parcels simply because the size of the small parcels allows the centroids to be closer together (independent of the spatial pattern).

The (mis)application of Ripley's K-function to areal data is partially attributable to the lack of distance-specific cluster detection methods designed specifically for areal data. Further, it is enabled by popular spatial software packages like ArcGIS, which map areal data to their centroids in order to apply Ripley's K-function. Specifically, when performing hypothesis testing via Ripley's K-function with areal data in the Multidistance Spatial Cluster Analysis ArcGIS tool [23], the 'observed points' (i.e., in continuous space) are defined as centroids of the polygons with the characteristic of interest [23]. This is done by default and without a warning message. We show in our simulation studies that such (mis)applications of Ripley's K-function to areal data often result in a severely inflated type I error rate.

In this paper, we develop a method for detecting distance-specific clustering/dispersion in areal data. Our method is motivated by Ripley's K-function and has a similar interpretation. In Section 2, we introduce our method and explore some of its properties. Section 3 presents the results of an extensive simulation study in which we compare the ability of our method to detect clustering and dispersion to that of the global Moran's I statistic, the Getis-Ord general G statistic, the spatial scan statistic, and three point-process methods. In Section 4, we demonstrate the use of our method on two real datasets. First, we use it to determine if there is spatialy clustering in land parcels that contain CEs in Boulder County, Colorado. Next, we use our method to determine if US counties with high childhood overweight rates are spatially clustered. Section 5 provides concluding remarks and suggestions for future work.

2. Methodology

To motivate our methods, we begin with a brief review of Ripley's K-function. Suppose we have a two-dimensional spatial point process $\mathscr{P}$ defined on a Borel set $\mathscr{A} \subseteq \mathbb{R}^2$. For any Borel set $\mathscr{S} \subseteq \mathscr{A}$, let $N(\mathscr{S})$ count the number of points (events) in $\mathscr{S}$. The intensity of the point process at a point $\boldsymbol{\ell} \in \mathscr{A}$ is given by

$$\lambda(\boldsymbol{\ell}) = \lim_{d\mathscr{S} \to \boldsymbol{\ell}} \frac{E[N(\mathscr{S})]}{d\mathscr{S}}$$

where $\mathscr{S}$ is an arbitrary neighborhood surrounding $\boldsymbol{\ell}$. A point process is said to be *stationary* if the intensity $\lambda(\boldsymbol{\ell})$ is a constant function. Note that for a stationary process, the expected number of points in an area depends only on the size of the area and not on its location. See [15] or [3] for further information on point processes.

2.1. Ripley's K-function

For a stationary point process $\mathscr{P}$ with intensity $\lambda(\boldsymbol{\ell}) = \lambda$ and a distance $r > 0$, Ripley's K-function is defined as

$$K(r) = \lambda^{-1} E\{N[c(\boldsymbol{\ell}, r)] - 1\}.$$

where $\boldsymbol{\ell}$ is any point arising from $\mathscr{P}$ and $c(\boldsymbol{\ell}, r)$ is the circle centered at $\boldsymbol{\ell}$ with radius r [49]. Thus $K(r)$ is the expected number of additional points within a distance of r of any point in $\mathscr{P}$, re-scaled by the intensity of $\mathscr{P}$. For a homogeneous Poisson process (realizations of which exhibit complete spatial randomness) on an infinite study area, $K(r) = \pi r^2$. For a high-level overview of Ripley's K-function, see [20]; for a more in depth treatment of Ripley's K-function and related topics in point processes, see [8] or [12].

While the original formulation of Ripley's K-function assumes a stationary point process, the definition has been extended to handle certain types of nonstationary processes [4], and many related fucntions have been developed to further quantify the behavior of point processes. For instance, the L-function is a rescaling of the K-function, and the K_d-function is a kernel estimator of the probability density function of distances between points [22]. Variants of Ripley's K-function have also been developed for *marked point processes* (MPP), i.e., point processes for which each point is associated with a random value called a *mark*. The D-function is defined for MPP with binary marks, and consists of the difference between the K-function computed on points with one type of mark and the K-function computed on the remaining points [19, 2]. The M-function quantifies the pattern in points with a particular type of mark relative to other points by weighting the number of points of the type in question within a distance r of a given point by the total number of points within distance r [42]. The cross K-function gives the expected number of observed marks within a distance r of a given mark, where the observed mark types must be different than the given mark. The K_{mm}-function quantifies the correlation between marks of points separated by a

distance r [47]. For an effective overview of the practical use and interpretation of these functions, see [41].

Suppose that we have a realization of $\mathscr{P}$ consisting of n observations, $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_n$. We can estimate $K(r)$ with

$$\hat{K}(r) = \hat{\lambda}^{-1} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \frac{w_{ij}}{n}$$

where $\hat{\lambda} = n/A(\mathscr{A})$, $A(\mathscr{A})$ denotes the area of $\mathscr{A}$, and w_{ij} is a weight associated with points i and j. In the traditional approach, $w_{ij} = 1$ if the distance between points i and j is less than r and 0 otherwise [49, 50]. In practice, the traditional estimator generally exhibits some bias due to *edge effects*. This phenomena arises because $N[c(\boldsymbol{\ell}_i, r)]$ is often lower than expected for $\boldsymbol{\ell}_i$ near the boundary of $\mathscr{A}$, as some points which would otherwise contributed to $N[c(\boldsymbol{\ell}_i, r)]$ fall outside of $\mathscr{A}$. However, many adjusted estimators of Ripley's K-function exist which modify the w_{ij} to account for edge effects [49, 18, 27, 1].

2.2. Using Ripley's K-function to Clustering and Dispersion

Ripley's K-function is often used to determine if an observed collection of points exhibits complete spatial randomness (CSR) (i.e., to determine if the points arise from a two-dimensional homogeneous Poisson process). The distribution of $\hat{K}(r)$ under CSR for a finite study area $\mathscr{A}$ can be approximated with Monte Carlo simulations, which are used to perform a hypothesis test with the null hypothesis being that the observed data arises from a homogeneous Poisson process with rate parameter $\hat{\lambda}$. In practice, these Monte Carlo simulations are often carried out conditional on a fixed number of observations (n), in which case the null hypothesis is that the data arise from a two dimensional continuous uniform distribution. Large values of $\hat{K}(r)$ indicate spatial clustering, that is, the number of points within a distance of r of any given point is larger than would be expected if the data exhibited CSR. Small values of $\hat{K}(r)$ indicate dispersion, that is, the number of points within a distance of r of any given point is smaller than would be expected under CSR. While Ripley's K-function is most commonly used to test a null hypothesis of CSR, it can be used to test more complicated null hypotheses, such as that the data arise from a Neyman-Scott process [18] or a Strauss process [14]. Ripley's K-function-based hypothesis tests are an attractive tool for spatial data analysis because of their flexibility and interpretability. Ripley's K-function can simultaneously detect different spatial patterns at different distances (e.g., small distance dispersion combined with large distance clustering), which is a highly desirable and somewhat rare property among cluster detection methods.

2.3. Areal Processes

In this work, we assume we have a set of spatial regions, referred to as areal units, whose boundaries are fixed and known (e.g. the census tracts in a particular state). We observe a binary response variable for each of these areal units (e.g. whether the census tract is a food desert) which we refer to as 'binary areal data' to distinguish it from areal data that

are associated with one or more non-binary numeric attributes. For brevity, we will refer to areal units for which this binary random variable is equal to 1 as *positive units*, as they are positive for the characteristic of interest. The location and boundaries of the areal units are considered fixed, with the binary random variable serving as a random 'mark'. To parallel the point process case, we loosely define clustering for binary areal data as an excess of positive units in a particular area and dispersion as positive units occurring a semi-regular intervals. The terms 'clustered data' or 'clustering' are sometimes used for data with positive spatial autocorrelation. For example, census tracts with a high rates of food insecurity might tend to be closer to other tracts with high rates. In binary areal data, there is no meaningful distinction between spatial clustering (an excess of positive units) and spatial autocorrelation (a higher concentration of similar values of the random variable).

We will define an areal process $\mathscr{A}$ on a Borel subset $\mathscr{A} \subseteq \mathbb{R}^2$ as a collection of N disjoint Borel sets $a_1, a_2, \ldots, a_N$ whose union covers $\mathscr{A}$ and a vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)'$ where Y_i is a binary random variable associated with a_i. We will refer to the a_i as *areal units*, and the set of a_i for which $Y_i = 1$ as *positive areal units*. For any Borel subset $\mathscr{S}$ of $\mathscr{A}$, define $\mathscr{N}(\mathscr{S}) = \sum_{i=1}^{N} A(\mathscr{S} \cap a_i) Y_i$. Thus $\mathscr{N}(\mathscr{S})$ is the amount of area in $\mathscr{S}$ which falls into positive areal units. Note that $\mathscr{N}(\cdot)$ maybe thought of as the areal analog of $N(\cdot)$ in the point process case: $N(\cdot)$ counts observed points, $\mathscr{N}(\cdot)$ counts observed area.

2.3.1. Extending the Concept of Stationarity—In the point process case, the intensity was a limit of the ratio of the number of points in a region divided by the area of the region. The analogous quantity for an areal process would be the ratio of the amount of positive area in a region divided by the area of that region. That is

$$\lambda(\boldsymbol{\ell}) = \lim_{d\mathscr{S} \to \boldsymbol{\ell}} \frac{E[\mathscr{N}(\mathscr{S})]}{d\mathscr{S}}$$

where $\mathscr{S}$ is an arbitrary neighborhood around $\boldsymbol{\ell}$. Thus for $\boldsymbol{\ell} \in a_i$, $\lambda(\boldsymbol{\ell}) = P(Y_i = 1)$. We define an areal process as 'stationary' if $P(Y_i = 1)$ is the same for all i. It can be shown that for a stationary areal process with no edges (for example, an areal process on $\mathbb{R}^2$), $E\left[\mathscr{N}(\mathscr{S})\right]$ depends only on the size of $\mathscr{S}$. For a stationary areal process with $P(Y_i = 1) = \lambda$ for all i, $E[\mathscr{N}(\mathscr{A})]/A(\mathscr{A}) = \lambda$. We also define an *independent* areal process as one for which the Y_i s are mutually independent.

2.4 . The Positive Area Proportion Function

Given an areal process $\mathscr{A}$, define the *positive area proportion function for an areal unit a_i at a distance $r > 0$* as

$$M_i(r) = E\left\{ \frac{\mathscr{N}[c(\boldsymbol{\ell}_i, r) \cap a_i^c] + A[c(\boldsymbol{\ell}_i, r) \cap a_i]}{A[c(\boldsymbol{\ell}_i, r) \cap \mathscr{A})]} \cdot \left(\frac{\mathscr{N}(\mathscr{A})}{A(\mathscr{A})} \right)^{-1} \mid n_Y > 0 \right\}.$$

where $\boldsymbol{\ell}_i$ is the centroid of a_i and n_Y is the number of positive units. Note that conditioning on n_Y is necessary for the expectation to be defined, as $\mathscr{N}(\mathscr{A}) = 0$ if $n_Y = 0$. The first term in the

expectation is the proportion of positive area within a distance of r of ℓ_i where the numerator captures area that either falls in a_i ($A[\cdot]$) or is positive ($\mathcal{N}[\cdot]$). The second term is the inverse of the total proportion of the study area which is positive.

Binary areal data is clustered if positive areas tend to occur near other positive areas. Here, $M_i(r)$ is used to quantify the expected amount of additional positive area near a_i. This quantity only meaningfully assesses clustering near a_i if a_i is positive. Just estimators of Ripley's K-function are only computed at observed points, estimators of $M_i(r)$ are only computed for positive units a_i. The inclusion of $A[c(\ell_i, r) \cap a_i]$ in the numerator of the first term ensures that sample based estimators, which are only computed if a_i is positive, are unbiased for $M_i(r)$. Just as Ripley's K-function quantifies the number of points within a certain distance of *an observed point*, the positive area proportion function quantifies the amount of positive area within a certain distance of *a positive unit's centroid*. See Figure 1 for an illustration of the quantities involved in $M_i(r)$.

Heuristically, $M_i(r)$ is loosely analogous to an edge-corrected Ripley's K-function, with positive area playing the role of observed points. However, the inherent areal structure makes $M_i(r)$ dependent on the choice of the positive unit a_i, while Ripley's K-function does not depend on the choice of the observed point. To remove this dependence on the choice of a_i, we can average the $M_i(r)$ values over the positive units. Define the *positive area proportion function at a distance r* by

$$M(r) = E\left[\frac{1}{n_Y}\sum_{i=1}^{N} M_i(r)Y_i \mid n_Y > 0\right]$$

where the expectation is taken over $\boldsymbol{Y}$. Thus, $M(r)$ is the expected value of the average of the positive area proportion function of the positive units. Allow $M_{0i}(r)$ and $M_0(r)$ to denote $M_i(r)$ and $M(r)$ for a stationary independent process, respectively. It can be shown that

$$M_0(r) = \frac{1}{N}\sum_{i=1}^{N} M_{0i}(r).$$

See Web Appendix A for more details.

For an areal process realization with $\boldsymbol{Y} = \boldsymbol{y}$ and $n_y > 0$, we can estimate $M_i(r)$ with

$$\widehat{M}_i(r, \boldsymbol{y}) = \frac{\mathcal{N}[c(\ell_i, r) \cap a_i^c] + A[c(\ell_i, r) \cap a_i]}{A[c(\ell_i, r) \cap \mathcal{A}]}\left(\frac{\mathcal{N}(\mathcal{A})}{A(\mathcal{A})}\right)^{-1}.$$

Note that by definition $E[\widehat{M}_i(r \mid \boldsymbol{Y}) \mid n_Y > 0] = M_i(r)$. Further, define

$$\widehat{M}(r, \boldsymbol{y}) = \frac{1}{n_y}\sum_{i:\, y_i = 1} \widehat{M}_i(r, \boldsymbol{y}).$$

as the sample mean of the $\widehat{M}_i(r, \boldsymbol{y})$ s for each positive unit. It can be shown that if $\mathscr{A}$ is a stationary, independent process, then $E[\widehat{M}(r, \boldsymbol{Y}) \mid n_{\boldsymbol{Y}} > 0] = M_0(r)$. This fact forms the crux of the hypothesis testing procedure presented in the next section. See Web Appendix A for additional details.

2.5. Hypothesis Testing using the Positive Area Proportion Function at Specific Distances

Suppose we have a realization $\boldsymbol{y}$ from an areal process $\mathscr{A}$, and we wish to test the null hypothesis that $\mathscr{A}$ is a stationary, independent areal process against the alternative hypothesis that $\mathscr{A}$ exhibits clustering and/or dispersion. To reduce the variability of the assumed null distribution, we will condition on the number of positive units in the realization. Conditional on $n_{\boldsymbol{Y}} = n$, the null hypothesis of a stationary, independent process implies that $P(\boldsymbol{Y} = \boldsymbol{y}) = \binom{N}{n}^{-1}$ for $\boldsymbol{y}$: $\sum_{i=1}^{N} y_i = n$ and $P(\boldsymbol{Y} = \boldsymbol{y}) = 0$ otherwise. The null hypothesis is thus equivalent to the so-called 'random labeling hypothesis', under which all configurations of n positive units are equally likely.

A clustered areal process might violate the stationarity assumption, or the independence assumption, or both. Formally, we consider an areal process to exhibit *excess-clustering* if there exists a subset of contiguous areal units indexed by $\mathscr{C}$ such that for all $i \in \mathscr{C}$, $P(Y_i = 1) > P(Y_j = 1)$ for $j \notin C$. That is, $\mathscr{A}$ exhibits excess clustering if the probability of any areal unit in $\mathscr{C}$ being positive is higher the probability of a unit out of $\mathscr{C}$ being positive. An areal process exhibiting excess-clustering is not stationary, though it could be independent.

We will consider an areal process to exhibit *autocorrelated-clustering* if there exists at least one areal unit a_i with at least one neighbor a_j such that $(Y_j = 1 \mid Y_i = 1) > P(Y_j = 1)$, that is $\boldsymbol{Y}$ exhibits positive spatial autocorrelation. Here the definition of neighbor can be taken to be any desired measure of proximity (shared border, centroids within a certain distance, etc.). Autocorrelated-clustering violates the independence assumption, but not necessarily the stationarity assumption. For the purposes of developing a hypothesis testing procedure, we will consider an areal process to be clustered if it exhibits excess-clustering or autocorrelated-clustering or both. Note that unless more than one realization of the same areal process is observed (which is rare), it will not be possible to distinguish between the two types of clustering.

Consider the following test statistic

$$T_n(r, \boldsymbol{y}) = \widehat{M}(r, \boldsymbol{y}) - M_0(r, n)$$

where $n = \sum_{i=1}^{N} y_i$, $M_0(r, n) = E[\frac{1}{n_{\boldsymbol{Y}}} \sum_{i=1}^{N} M_{i0}(r, n) Y_i \mid n_{\boldsymbol{Y}} = n]$ and $M_{i0}(r, n) = E\left\{ \frac{\mathscr{N}[c(\mathscr{C}_i, r) \cap a_i^c] + A[c(\mathscr{C}_i, r) \cap a_i]}{A[c(\mathscr{C}_i, r) \cap \mathscr{A})]} \cdot \left(\frac{\mathscr{N}(\mathscr{A})}{A(\mathscr{A})} \right)^{-1} \mid n_{\boldsymbol{Y}} = n \right\}$ for a stationary, independent process. Under the null hypothesis, $E[T_n(r, \boldsymbol{Y}) \mid n_{\boldsymbol{Y}} = n] = 0$. Positive values of $T_n(r, \boldsymbol{y})$ suggest

clustering at distance r. Data which exhibits excess-clustering will have a larger-than-expected number of positive units in $\mathscr{C}$, which will tend to inflate the values of $\widehat{M}_i(r, \boldsymbol{y})$ for $i \in \mathscr{C}$ and thus inflate $\widehat{M}(r, \boldsymbol{y})$. Data which exhibits autocorrelated-clustering will have a larger than expected number of positive units with other positive units nearby, which will also inflate $\widehat{M}(r, \boldsymbol{y})$. When the null hypothesis is rejected in favor of clustering at distance r, we conclude that there is more area within a distance of r of each positive unit centroid than expected under a stationary, independent process.

The null distribution of $T_n(r, \cdot)$ can be estimated using a Monte Carlo procedure in which data is generated under the random labeling hypothesis, that is, the $\boldsymbol{Y}$ values are generated by selecting n positive areal units with equal probability. Note that $M_0(r, n)$ can be computed exactly, but doing so requires evaluating $N\binom{N}{n}$-dimensional sums. Additionally, obtaining the terms in these sum requires computing the area of polygon intersections, which is generally computationally intense. As an alternative, we propose approximating $M_0(r, n)$ by averaging the $\widehat{M}(r, \boldsymbol{y})$ values from the same Monte Carlo procedure used to estimate the null distribution of $T_n(r, \cdot)$. An α-level test for clustering at distance r can then be conducted by rejecting the null hypothesis if $T_n(r, \boldsymbol{y})$ exceeds $1 - \alpha$ quantile estimated from the simulated null distribution.

The same test statistic $T_n(r, \cdot)$ can be used to detect dispersion. Dispersion is a difficult phenomena to precisely quantify. Intuitively, an areal process is dispersed if positive areal units tend to be located further away from other positive units than would be expected for a stationary, independent process. To produce a working definition of dispersion, suppose we have defined a neighbor structure on the areal units. We will consider an areal process to exhibit *buffered-dispersion* if there exists a set of non-neighboring unit(s) indexed by $\mathscr{D}$ such that for all $i \in \mathscr{D}$ and all neighbors a_j of a_i, $P(Y_i = 1) > P(Y_j = 1)$. Here, the units in $\mathscr{D}$ are surrounded by 'buffer units' with a lower probability of being positive. An areal process which exhibits buffer-dispersion is not stationary, but it could be independent. We will consider an areal process to exhibit *autocorrelated-dispersion* if there exists at least one areal unit a_i with at least one neighbor a_j such that $P(Y_j = 1 \mid Y_i = 1) < P(Y_j = 1)$. Autocorrelated dispersed areal processes are not independent, though they may be stationary. For the purposes of hypothesis testing, we will consider an areal process dispersed if it is either buffer-dispersed or autocorrelated-dispersed. While this definition likely does not cover all processes which could give rise to dispersed areal data, it does cover two most common cases of areal unit(s) with a high probability of being positive surrounded by areal units with lower probability of being positive and the case of negative spatial autocorrelation.

If an areal process is dispersed, then at least some of the positive units have fewer positive units nearby than we would expect for a stationary, independent process. These units will tend have to have lower than expected $\widehat{M}_i(r, \boldsymbol{y})$ values, decreasing $\widehat{M}(r, \boldsymbol{y})$. We can perform an α-level test of the null hypothesis that $\mathscr{A}$ is a stationary, independent process against the alternative hypothesis that $\mathscr{A}$ exhibits dispersion by rejecting the null hypothesis if $T_n(r, \boldsymbol{y})$ falls below the α-quantile of the null distribution of $T_n(r, \cdot)$, which can be estimated using the Monte Carlo procedure described previously. If we reject the null hypothesis in favor

of dispersion, we conclude that the amount of positive areal within a distance r of positive unit centroids is less than expected for stationary independent process. Finally, an α-level two-tailed test for either clustering or dispersion can be conducted by rejecting the null hypothesis if $T_n(r, \boldsymbol{y})$ is less than the estimated $\alpha/2$ quantile of the null distribution or if $T_n(r, \boldsymbol{y})$ is greater than the $1 - \alpha/2$ quantile.

2.6. A Global Hypothesis Test for Type I Error Rate Control

In many applications, the ideal distance at which to test for spatial patterns may not be known. Computing the PAPF test statistic at a variety of radii induces a multiple testing problem and the potential for an inflated type I error rate. To control the overall type I error rate associated with such a procedure, we propose the following global test for clustering over a range of distances $r \in \mathscr{R} = \{r_1, \ldots, r_R\}$. Define the global test statistic $T_{nC}(\boldsymbol{y}) = \max_{r \in \mathscr{R}} \{T_n(r, \boldsymbol{y})/S_{nr}\}$ where $S_{nr} = \text{var}[T_n(r, \cdot)]^{1/2}$. We can estimate the null distribution of $T_{nC}(\cdot)$ within the Monte Carlo procedure discussed above. That is, $T_n(r, \boldsymbol{y})$ is computed for each $r \in \mathscr{R}$ and $T_{nC}(\cdot)$ is computed for each dataset. Note that we can estimate $\text{var}[T_n(r, \cdot)]$ from the same Monte Carlo procedure. An upper tailed test is indicative of clustering. To test for dispersion, define the test statistic $T_{nD}(\boldsymbol{Y}) = \min_{r \in \mathscr{R}} \{T_n(r, \boldsymbol{y})/S_{nr}\}$. The null distribution of $T_{nD}(\cdot)$ can be estimated analogously to that of $T_{Cn}(\cdot)$ and a lower tailed test is indicative of dispersion. To simultaneously test for either clustering or dispersion, both tests can be performed and a Bonferroni correction for two tests applied. Similar methods have been proposed to derive global tests from Ripley's K-function and Ripley's D-function [19].

3. Simulation Study

3.1. Simulation Specifications

In this section, we perform an extensive simulation study to compare the performance our proposed PAPF hypothesis testing method to the performance of the global Moran's I statistic, the Getis-Ord general G statistic, and the spatial scan statistic. We also consider the performance of the misapplication of three point process methods to the areal unit centroids: a edge-corrected Ripley's K-function test, a related test based on Ripley's D-function [19], and the average nearest neighbor method [13]. These misapplications are included to highlight the importance of analyzing areal data only with methods designed for areal data. After describing our data generation procedures, we provide details on the implementation of each method.

We consider the performance of our proposed hypothesis testing procedure using two study areas which are shown in Figure 2:

$\mathscr{A}_1$ A 20 by 20 regular grid of $N_1 = 400$ cells

$\mathscr{A}_2$ The $N_2 = 3,108$ counties (and county-equivalents) in the contiguous US.

We define three distributions that will be used to sample the observed units. For each, areal unit a_i is positive when $Y_i = 1$. First, $\boldsymbol{Y} \sim \text{SWoR}(N, k, \boldsymbol{p})$ indicates that the random variable

$\boldsymbol{Y} = (Y_1, \ldots, Y_N)$ arises by selecting k elements from $\{1, 2, \ldots, N\}$ via sampling without replacement (SWoR) where $\boldsymbol{p} = (p_1, \ldots, p_N)'$ gives the probability of selecting each element, and $Y_i = 1$ if i was selected and 0 otherwise. Second, $\boldsymbol{Y} \sim C(k, m, q)$ denotes that $\boldsymbol{Y}$ is generated using the following two-step process where $Y_i = 1$ if element i was selected in either step:

- **i.** m elements of $\{1, 2, \ldots, N\}$ are randomly selected via $\text{SWoR}(N, m, \boldsymbol{p}_1)$ where $p_{1i} = N^{-1}$ for all i.
- **ii.** $k - m$ elements of $\{1, 2, \ldots, N\}$ are selected via $\text{SWoR}(N, k - m, \boldsymbol{p}_2)$, where $p_{2i} = 0$ if i was selected in step (i), $p_{2i} = q/D$ for i such that a_i shares a border with at least one unit selected in step (i), and $p_{2i} = 1/D$ otherwise where D is such that $\sum_i p_{2i} = 1$.

Third we define $\boldsymbol{Y} \sim M(n, m, q, k)$ if $\boldsymbol{Y}$ is generated as follows:

- **i.** Divide the study area $\mathscr{A}$ into three regions: a clustered region R_c, a dispersed region R_d and a random scatter region R_r.
- **ii.** Select m units from R_c as follows:
  - **a.** Select 1 unit via $\text{SWoR}(N_c, 1, \boldsymbol{p}_c)$, where N_c is the number of units in R_c and $\boldsymbol{p}_c = \left(N_c^{-1}, \ldots, N_c^{-1}\right)'$.
  - **b.** Set $p_{ci} = q/D$ if a_i is adjacent to a previously selected unit, $p_{ci} = 0$ if a_i is a previously selected unit and $p_{ci} = 1/D$ otherwise; D is chosen so that $\sum_{i=1}^{N_c} p_{ci} = 1$.
  - **c.** Select one unit via $\text{SWoR}(N_c, 1, \boldsymbol{p}_c)$
  - **d.** Return to (b) until m units have been selected.
- **iii.** Select km units from R_d via an analogous process to (ii) with p_i now taken to be $1/(qD)$ in step (b) if unit a_i is adjacent to a previously selected unit.
- **iv.** Select $n - k(m + 1)$ units from R_r via $\text{SWoR}(N_r, n - k(m + 1), \boldsymbol{p}_r)$, where N_r is the number of units in R_r and $\boldsymbol{p}_r = \left(N_r^{-1}, \ldots, N_r^{-1}\right)'$.

For the first distribution, $\boldsymbol{p}$ can be used to to generate data under the null hypothesis or data with clustering in certain locations where the p_i's are larger. For the second distribution, q controls the degree of clustering or dispersion where larger $q > 1$ lead to more clustering while smaller $q < 1$ lead to more dispersion. The third distribution produces an area of small-scale clustering in R_c and an area of small-scale dispersion in R_d. When k is sufficiently large relatively to N_c, this process will also produce clustering at larger distances in R_c.

For each study area $\mathscr{A}_j$, $j = 1, 2$, we generate data under 21 different scenarios. For brevity, we remove the subscript j when describing these scenarios (all depend on N_j). First we consider the null hypothesis of a stationary independent areal process (e.g. no spatial pattern in the positive units) via the following three scenarios:

$$I_1: \boldsymbol{Y} \sim \text{SWoR}(N, \lceil N/10 \rceil, \boldsymbol{p})$$

$$I_2: \boldsymbol{Y} \sim \text{SWoR}(N, \lceil N/4 \rceil, \boldsymbol{p})$$

$$I_3: \boldsymbol{Y} \sim \text{SWoR}(N, \lceil N/2 \rceil, \boldsymbol{p})$$

where $\boldsymbol{p} = (p_1, \ldots, p_N)' = (1/N, \ldots, 1/N)'$.

We also consider 12 scenarios in which the locations of observed units are clustered. The following 6 scenarios assess the ability to excess-clustering:

$$C_1: \boldsymbol{Y} \sim \text{SWoR}(N, \lceil N/10 \rceil, \boldsymbol{p}_1)$$

$$C_2: \boldsymbol{Y} \sim \text{SWoR}(N, \lceil N/10 \rceil, \boldsymbol{p}_2)$$

$$C_3: \boldsymbol{Y} \sim \text{SWoR}(N, \lceil N/4 \rceil, \boldsymbol{p}_3)$$

$$C_4: \boldsymbol{Y} \sim \text{SWoR}\left(N, \left\lceil \frac{N}{4} \right\rceil, \boldsymbol{p}_4\right)$$

$$C_5: \boldsymbol{Y} \sim \text{SWoR}(N, \lceil N/2 \rceil, \boldsymbol{p}_5)$$

$$C_6: \boldsymbol{Y} \sim \text{SWoR}(N, \lceil N/2 \rceil, \boldsymbol{p}_6).$$

The N-dimensional vectors $\boldsymbol{p}_l = (p_{l1}, \ldots, p_{lN})'$ for $l = 1, \ldots, 6$ are defined as follows: an entry of p_l is equal to q/D if the unit is shown in blue in Figure 3 and equal to $1/D$ otherwise where D is such that $\sum_i p_{li} = 1$. For $l \in \{1, 3, 5\}$ we take $q = 5$ and for $l \in \{2, 4, 6\}$ we take $q = 10$. Thus the blue units are 5 times more likely to be observed than the white units under C_1, C_3 and C_5, and 10 times more likely to be observed under C_2, C_4 and C_6.

Next, we assess the ability of our hypothesis test to autocorrelated-clustering by considering the following 6 scenarios:

$$C_7: \boldsymbol{Y} \sim C(\lceil N/10 \rceil, \lceil N/100 \rceil, 5)$$

$$C_8: \boldsymbol{Y} \sim C(\lceil N/10 \rceil,\ \lceil N/100 \rceil,\ 10)$$

$$C_9: \boldsymbol{Y} \sim C(\lceil N/4 \rceil,\ \lceil N/40 \rceil,\ 5)$$

$$C_{10}: \boldsymbol{Y} \sim C(\lceil N/4 \rceil,\ \lceil N/40 \rceil,\ 10)$$

$$C_{11}: \boldsymbol{Y} \sim C(\lceil N/2 \rceil,\ \lceil N/20 \rceil,\ 5)$$

$$C_{12}: \boldsymbol{Y} \sim C(\lceil N/2 \rceil,\ \lceil N/20 \rceil,\ 10)$$

Note that C_7, C_9 and C_{11} correspond to 'weaker' clustering, in the sense that they tend to select fewer adjacent units than mechanisms C_8, C_{10} and C_{12}

Next, we assess the ability of our hypothesis testing procedure to detect spatial dispersion under the following 4 scenarios.

$$D_1: \boldsymbol{Y} \sim C\left(\lceil N/10 \rceil,\ \lceil N/100 \rceil,\ \frac{1}{10}\right)$$

$$D_2: \boldsymbol{Y} \sim C(\lceil N/10 \rceil,\ \lceil N/100 \rceil,\ 0)$$

$$D_3: \boldsymbol{Y} \sim C\left(\lceil N/6 \rceil,\ \lceil N/100 \rceil,\ \frac{1}{10}\right)$$

$$D_4: \boldsymbol{Y} \sim C(\lceil N/6 \rceil,\ \lceil N/100 \rceil,\ 0)$$

Here, in D_1 and D_3 adjacent units are one tenth as likely to be observed as non-adjacent units, creating a mild dispersion effect. Under D_2 and D_4 adjacent units cannot be selected at all, creating a stronger dispersion effect. Finally, we consider only two samples sizes when assessing dispersion ($N/10$ and $N/6$) because it becomes increasingly difficult or impossible to select only non-adjacent units as number of selected units increases.

Finally, we assess the ability of the PAPF method to detect the simultaneous presence of spatial clustering and dispersion at different distances using the following two scenarios:

$$M_1: \boldsymbol{Y} \sim M(\lceil N/10 \rceil, m_{j1}, 10,\ 3)$$

$$M_2: \boldsymbol{Y} \sim M(\lceil N/6 \rceil, m_{j2}, 10,\ 3)$$

We take $m_{11} = 9,\ m_{21} = 13,\ m_{12} = 60$ and $m_{22} = 100$. The regions $R_c,\ R_d$ and R_r defined for our two study areas are shown in Figure 3. Examples of data generated each scenario for $\mathscr{A}_1$ are shown in Figure 4. Web Figure 1 provides similar examples for $\mathscr{A}_2$.

3.2. Method Specifications

All hypothesis tests were performed with at a level of $\alpha = 0.05$. For each study area, both global and radii-specific PAPF tests, Ripley's K tests and Ripley's D tests were performed. The radii-specifc PAPF tests were performed for 10 radii, $r \in \mathscr{R}_j = \{r_{j1}, \ldots, r_{j10}\}$. The set of radii used for each study area are shown in Figure 2. The smallest radius was equal the smallest distance between any areal unit centroids, and the radii were increased incrementally, with the largest radii being approximately one fourth of the width of the study area.

3.2.1. PAPF Method—To perform the PAPF hypothesis test, Monte Carlo simulations were used. For each study area $\mathscr{A}_j$, and observation size $n \in \{\lceil N_j/10 \rceil,\ \lceil N_j/6 \rceil,\ \lceil N_j/4 \rceil,\ \lceil N_j/2 \rceil\},\ G = 200$ datasets were simulated under the null hypothesis of an stationary and independent areal process, conditional on n positive units being observed. Specifically, for a given $\mathscr{A}_j$, n, and $\boldsymbol{p}_j = (N_j^{-1}, \ldots, N_j^{-1})',\ \boldsymbol{y}_g \sim \text{SWoR}(N_j, n, \boldsymbol{p}_j)$ was sampled for $g = 1, \ldots, G$, and

$$\widehat{M}_0(r, n) = \frac{1}{G} \sum_{g=1}^{G} \widehat{M}(r, \boldsymbol{y}_g)$$

was calculated for each $r \in \mathscr{R}_j$. Then to approximate the distribution of the test statistic $T_n(r, \cdot)$,

$$T_n(r, \boldsymbol{y}_g) = \widehat{M}(r, \boldsymbol{y}_g) - \widehat{M}_0(r, n)$$

was calculated for $g = 1, \ldots, G$ and the quantiles of $\{T_n(r, \boldsymbol{y}_1,), \ldots, T_n(r, \boldsymbol{y}_G)\}$ were used as approximate critical values for the hypothesis test. These Monte Carlo simulatons were also used to estimate the null distributions of $T_{nC}(\cdot)$ and $T_{nD}(\cdot)$.

After approximating each null distribution, 500 instances of $\boldsymbol{y}$ were generated under each of the 21 DGMs. For each generated $\boldsymbol{y},\ T_{nC}(\boldsymbol{y}), T_{nD}(\boldsymbol{y})$ and $T_n(r, \boldsymbol{y})$ for $r \in \mathscr{R}_j$ were computed and compared to critical values from the corresponding null distribution.

3.2.2. Comparison Methods—Table 1 provides the assumed null, right-tailed and left-tailed alternative hypotheses for the PAPF, the global Moran's I statistic, the Getis-Ord general G statistic, the spatial scan statistic, and the misapplications of Ripley's K-function, Ripley's D-function, and the average nearest neighbor method. If the method requires a

Monte Carlo procedure for estimating the null distribution of the test statistic, details of the data generation mechanism are also provided.

The global Moran's I statistic was computed for each dataset using the `Moran.I` function in the R package `ape` [46] using an adjacency-based spatial weights matrix. The Getis-Ord general G statistic was computed for each dataset using the `globalGtest` function in the R package `spdep` [9] using the same adjacency-based spatial weights matrix. The spatial scan statistic was computed using the `scan.test` function in the `spatstat` R package [5]. A binomial likelihood was assumed, with each areal unit having 1 trial. A set of circular zones with radii $r \in \mathscr{R}_j$ were considered. The Monte Carlo procedure for the spatial scan statistic is consistent with the null hypothesis of the statistic after conditioning on the number of observations n.

Ripley's K-function was computed with the `Kest` function in the R `spatstat` package [5], using the centroids of positive units as the set of observation locations. The `envelope` function (also in the `spatstat` package) with nsims = 200 was used to perform hypothesis testing for the radii-specific tests. This corresponds to generating 200 simulated datasets consisting of n observations generated from a two dimensional continuous uniform distribution on $\mathscr{A}$. Note that neither the assumed null hypothesis nor the Monte Carlo procedure for Ripley's K-function is consistent with the random labelling hypothesis. In fact, the Monte Carlo procedure does not even produce 'centroids' which are consistent with the assumed areal structure, that is, the points generated by the Monte Carlo procedure are not a subset of the areal unit centroids. Nevertheless, this is the approach used by ArcGIS to apply Ripley's K-function to areal data [23] and appears to be widely used in practice (see Section 1 for examples). As the centroids of the areal units are fully dependent on the assumed areal structure, it is highly unlikely that the centroids of positive units will be consistent with the assumed null hypothesis/Monte Carlo procedure of homogeneous point process (or for that matter, any stationary point process) even in the absence of clustering. We also applied a global Ripley's K test inspired by [19]. We used the following test statistic to detect clustering $\hat{K}_{nC}(\boldsymbol{y}) = \max\{\hat{K}(r)/\sqrt{\text{var}[\hat{K}(r)]} : r \in \mathscr{R}_j\}$ and defined $\hat{K}_{nD}(\boldsymbol{y})$ analogously to detect dispersion. The null distributions of both test statistics were estimated with Monte Carlo simulations for which data was generated using the `evelope` function as described above.

Diggle and Chetwynd propose a method for adapting Ripley's K-function for the detection of spatial clustering relative to a non-homogeneous null distribution, referred to as Ripley's D-function [19]. Observations are assumed to belong to one of two types (cases or controls). Ripley's D-function $D(r)$ is defined as the difference between Ripley's K-function computed using only the cases and Ripley's K-function computed using only the controls. To apply this method to our datasets, the centroids of positive areal units were treated as cases and the centroids of the other areal units were treated as the controls. The null hypothesis of random labeling is thus the same as an independent stationary areal process after conditioning on the number of positive areal units. The `Kest` function was used to compute Ripley's K-function for the cases and controls separately, and the test statistic was taken to be difference in Ripley's K-function between the two groups. An upper-tailed test was used to detect

clustering, and a lower tailed test was used to detect dispersion. Monte Carlo simulations were used to estimate the null distribution of the test statistic. Again following [19], we also applied a global Ripley's D test. The test statistics D_{nC} and D_{nD} were defined analogously to the global PAPF and global Ripley's K-function test statistics, with the estimated Ripley's D-function playing the role of $\widehat{M}(r, \boldsymbol{y})$ or $\hat{K}(r, \boldsymbol{y})$. The null distributions of both test statistics were estimated with Monte Carlo simulations as described above.

Clark and Evans develop the average nearest neighbor method for detecting clustering in point process data [13]. The test statistic compares the average distance between each point and its nearest neighbor to the expected distance under a null hypothesis of complete spatial randomness. To apply this method to our data, the centroids of the positive areal units were treated as the observation locations. We note that the assumed null hypothesis of this method is not consistent with the random labeling hypothesis. The average nearest neighbor test statistic was computed using the `nni` function in the `spatialEco` R package [24].

For each method, a two-tailed test was performed in scenarios $I_1 - I_3$. For the spatial scan statistic and the Getis-Ord general G, these two-tailed test detect only clustering; for all other methods, the two-tailed test detects both clustering and dispersion. The two-tailed test for the global PAPF, global Ripley's K and global Ripley's D methods were conducting by performing separate tests for clustering and dispersion using the corresponding test statistics and applying a Bonferroni correction for 2 tests. As the empirical type I error rate for Ripley's K-function and the ANN method were severely inflated, they were not applied to the other scenarios. For scenarios $C_1 - C_{12}$, an α-level test for clustering was used for all the remaining methods. We note that a test for clustering is a single tailed test for all methods except the spatial scan statistic, for which it is a two-tailed test (see Table 1). For scenarios $D_1 - D_3$, an α-level test for dispersion was performed for the PAPF, Moran's I and Ripley's D methods. As neither the Getis-Ord general G statistic nor the spatial scan statistic can detect dispersion, these methods were not applied to these scenarios. Finally, for scenarios $M_1 - M_2$, separate α-level tests for clustering and dispersion were performed using the PAPF, Moran's I, and Ripley's D-function to assess the presence of both clustering and dispersion. An α-level test for clustering was also performed using the Getis-Ord general G and the spatial scan statistic.

3.3. Simulation Results

Tables 2 and 3 summarize selected results from study area $\mathscr{A}_1$ (the regular grid) and study area $\mathscr{A}_2$ (the US counties), respectively. The remaining results can be found in Web Tables 1 and 2. Each table reports the empirical rate of rejection for the null hypothesis. For scenarios I_1, I_2 and I_3, this quantity is the empirical type I error rate; for the other scenarios, this quantity is the empirical power. As the PAPF, Ripley's K and Ripley's D tests were performed at 10 different radii, the rejection rate for each radius is reported separately.

Under the null scenarios (I_1, I_2 and I_3), the empirical type I error rates of the global and radii-specific PAPF methods, the global Moran's I statistic, the Getis-Ord general G statistic, the spatial scan statistic and the global and radii-specific Ripley's D methods are within the Monte Carlo margin of error of their nominal levels. However the empirical type I error rate

of global and radii-specific Ripley's K methods and the average nearest neighbor method were highly inflated for most scenarios. This is presumably due to the incongruities between the assumed null hypothesis of these tests and the null hypothesis under which data was generated.

Under the excess clustering scenarios ($C_1 - C_6$), the global PAPF test had 100% empirical power to detect clustering, with the exception of scenario C_1 under the regular grid. Additionally, the power of the radii-specific PAPF tests is fairly consistent across radii. For these scenarios, the performance of the global PAPF test is comparable to or better than the performance of the Moran's I statistic, the Getis-Ord general G statistic and the spatial scan statistic. Interestingly, while the performance of the global and radii-specific Ripley's D tests is comparable to or better than that of the global and radii-specific PAPF test for DGMs $C_1 - C_4$, the Ripley's D test performs quite poorly for the larger sample sizes (DGMs $C_5 - C_6$) on the regular grid, while the PAPF test continues to perform well.

Under the autocorrelated-clustering DGMs ($C_7 - C_{12}$) the empirical power of the global PAPF test is generally high and the performance of the PAPF is comparable to that of Moran's I, Getis-Ord general G, and the scan statistic, with the exception scenarios C_7 on the regular grid and C_{11} on the US counties, for which Moran's I and Getis-Ord are the top performers. The performance of the PAPF test and the Ripley's D test are generally comparable for scenarios $C_7 - C_{10}$ on the regular grid, but the performance of the Ripley's D test deteriorates drastically for scenarios $C_{11} - C_{12}$, while the PAPF test continues to exhibit good performance. Both the global and local PAPF tests tends to outperform their Ripley's D counterparts for the US county scenarios.

Under the dispersion DGMs ($D_1 - D_4$) the performance of all methods was roughly comparable, except for the smaller sample sizes for the regular grid, for which the performance of the global PAPF statistic was noticeably worse than the others.

Under the mixture of clustering and dispersion scenarios ($M_1 - M_2$), the ability of the global PAPF test and global Ripley's D tests to detect clustering is quite good for the regular grid, but performance deteriorates noticeably for the US counties. The opposite is true for the spatial scan statistic, which exhibits poor performance (0% power) for the regular grid but excellent performance (100% power) for the US counties. The performance of the global Moran's I and Getis-Ord general G statistics to detect clustering is poor for all scenarios. The ability of the PAPF test, the Ripley's D test, and the global Moran's I statistic to detect dispersion is poor for the regular grid. However the ability of the PAPF test to detect dispersion on the US counties improves dramatically, while the performance of the other methods remain poor.

In summary, of the nine methods considered, only seven (the global and radii-specific PAPF tests, the global Moran's I statistic, the Getis-Ord general G statistic, the spatial scan statistic and the global and radii-specific Ripley's D tests) maintained their nominal type I error rates. The empirical power of the PAPF test was generally comparable to or better than all other methods, though there were a few exceptions to this rule.

Only Ripley's K-function, Ripley's D-function and the PAPF method allow one to detect distance-specific clustering or dispersion. Thus, if one wishes to test for spatial patterns at a single, specific distance, the radii-specific Ripley's K, Ripley's D and PAPF methods are the only potential options. However, the Ripley's K methods exhibited a severely inflated type I error rate and are thus unreliable. Notably, the performance of the global PAPF test is comparable to or better than that of the global Ripley's D test for all scenarios. In the some scenarios, the performance of the global PAPF test was 4–6 times better than that of the global Ripley's D test. The Ripley's D test is designed for point process data and is here being misapplied to areal data, which may explain some of the strange behavior in it's results (e.g. worsening performance as the number of positive units increases). The Ripley's D test compares Ripley's K-function on the positive units to Ripley's K-function on the negative units. In scenarios C_5 and C_6 on the regular grid, the global Ripley's D test performed very poorly. In these scenarios, almost all of the 200 negative units occur outside the region of excess clustering (shown in blue in Figure 3). As there are 300 total units outside this region, this implies that probability that a unit in this region is negative is roughly 0.66, as opposed to 0.5 under the random labeling hypothesis. This implies that the negative units are *also clustering*- that is, they are more likely to be in the non-blue region than expected under the null hypothesis. Since Ripley's D-function compares the degree of clustering in the positive units to the degree of clustering in the negative units, the presence of clustering in both types of units may explain the lack of statistical significance in the global Ripley's D test. While this clustering of negative units also occurs in scenarios $C_1 - C_4$, it is less extreme. For example, in scenario C_2, the probability of a unit outside the blue region being negative is approximately 0.97, versus 0.9 under the null hypothesis. The generally superior performance of the global PAPF method combined with the fact that Ripley's D-function is a point process method being misapplied to areal data compel use to recommend the use of the PAPF rather than Ripley's D-function to detect distance-specific patterns.

3.4. Supplementary Simulation Results

To assess the performance of the PAPF method under extremely small sample sizes, a simulation study was conducted using a 4×5 regular grid. In general, the global Ripley's D test had the highest power to detect clustering and the ability of all global methods to detect dispersion was roughly the same. However power of the radii-specific PAPF tests to detect both clustering and dispersion was generally higher than that of the radii-specific Ripley's D tests. The full simulation results can be found in Web Table 3.

Another simulation study was conducted to assess the sensitivity of the PAPF method to the dependence on areal unit centroids. In brief, the role of each areal unit's centroid in the definition of the PAPF was replaced with a randomly choosen point from each areal unit. For full details, see Web Appendix B of the Supplementary Material. The results of this simulation study are found in Web Tables 4 and 5 in the Supplementary Material. The performance of this alternative PAPF method was comparable to that of the standard method.

4. Data Application

In this section, we consider the performance of our method on real world applications from two different fields. First, we use the method to determine if land parcels with CEs are clustered in Boulder County, Colorado, using the 112,819 distinct land parcels in Boulder County as the areal structure. Next, we apply our method to determine if US counties with high childhood overweight rates are spatially clustered, using the 3,108 county and county-equivalents in the contiguous US as the areal structure.

4.1. Application to Conservation Easements

CEs are a private and generally perpetual form of land conservation that legally severs aspects of private landownership (e.g., development rights, resource extraction, etc.) from a parcel of land [44]. Although a landowner makes an individual decision to place a CE, prior research has indicated spatial clustering of CEs over time, throughout the US [36]. Cumulative and clustered CE use may impact regional ecosystem character by altering the degree of CE parcels' isolation or connectivity with other ecologically valuable parcels and may change ecologic quality on the CE parcel itself [29]. The greater the mass of clustering and ecological systems integrity, the more impact there may be on the land conversion rates at the county level, and on the decision to leave a parcel in open space (or not), potentially affecting placement of other socially valuable land uses. Recognizing if and where CEs are clustered and linking the social, political, biological, and geographical characteristics to the clustered areas may help elucidate the factors driving CE placement [7].

The Boulder County data consists of 112,819 land parcels in place in 2008. Of these land parcels, 817 were held as CEs. A parcel was considered to be part of a CE if any part of the parcel was part of an easement. Figure 5 depicts the land parcels; CE parcels are shown in blue. Global PAPF tests for clustering and dispersion were applied to the dataset using a Bonferroni correction to maintain an overall significance level of α, along with two-tailed radii-specific tests at 10 different radii, depicted in Figure 5.

In order to apply our method, the distribution of the global and radii-specific PAPF test statistic under the null hypothesis was estimated using 200 Monte Carlo simulations. In each Monte Carlo simulation, 817 parcels were selected via simple random sampling without replacement. The observed global PAPF test statistic for clustering ($T_{817C}(\boldsymbol{y})$) was larger than the 97.5th quantile of it's estimated null distribution. The radii-specific test for radius r_2 (approximately 1.1 miles) was also larger than the 97.5th quantile of it's null distribution. No other radii-specific tests were significant. These results indicate that parcels which contain CEs are significantly clustered at small distances. The exact test statistics and estimated quantiles can be found in Web Table 6 in the Supplementary Material.

In the context of CEs with a purpose of biological conservation, clustering easements close to one another is one reserve design principle to improve landscape connectivity and combat the adverse effects of habitat fragmentation from human land conversion [17, 31]. Larger and higher quality habitats (particularly on CEs) increase the size and stability of source populations and subsequently increase species dispersal capabilities [32]. Clustering and structural connectivity between conservation areas are not always positive, however,

as clustering may also leave these areas vulnerable to spatially autocorrelated extinction pressures, such as diseases, invasive species, stochastic environmental events, or negative effects from localized urban growth [21]. Given that the PAPF method indicated the spatial clustering of CEs at short distances in Boulder County, more detailed landscape connectivity studies focused on functional connectivity may be warranted [6, 54].

4.2. Application to Counties with High Childhood Overweight/Obesity Rates

Next, we use the PAPF method to determine if US counties with high childhood overweight/obesity rates are spatially clustered. County-level childhood overweight rates were estimated from data collected in the 2016 National Survey of Children's Health using a multilevel small area estimation approach as described in [56]. A county was considered to have a high overweight rate if its estimated rate exceeded the 75th percentile of all county overweight rates. There are 3,108 counties and county-equivalents in the contiguous US, and 786 of these counties were found to have a high rate of childhood overweight. These counties are shown in blue in Figure 5, along with the radii at which PAPF was applied. As for the CEs data application, an $\alpha = 0.05$ global test for clustering or dispersion was conducted using $T_{786C}(\boldsymbol{y})$ and $T_{786D}(\boldsymbol{y})$ and applying a Bonferroni correction. A two-tailed $\alpha = .05$ level test was also conducted for each radii.

The distribution of the global and radii-specific PAPF test statistics under the null hypothesis was estimated using 200 Monte Carlo simulations. In each Monte Carlo simulation, 786 counties were selected via simple random sampling without replacement. The observed global PAPF test statistic for clustering ($T_{786C}(\boldsymbol{y})$) was larger than the 97.5th quantile of it's estimated null distribution. All 10 radii-specific test statistics were also greater than the 97.5th quantiles of their null distributions, indicating that counties with high rates of childhood overweight are significantly clustered at small and large distances. As Southeastern and Midwest states tend to have higher overweight and obesity rates than the rest of the country [25, 11], these results are not surprising. The exact test statistics and estimated quantiles can be found in Web Table 7 in the Supplementary Material.

5. Conclusion

The problem of assessing binary areal data for distance-specific spatial clustering or dispersion has been the subject of relatively little attention. Existing methods such as the global Moran's I statistic, the Getis-Ord general G statistic, and the spatial scan statistic are global tests for clustering which are not directly amenable to determining the distance at which clustering is occurring. The global Moran's I and Getis-Ord general G statistics can be sometimes be tuned to pick up patterns at a specific distance by choosing an appropriate spatial weight matrix. However, the interpretation of the spatial scale induced by the weight matrix is often very complex, particularly for inverse-distance based weights (for which all observations are related to some degree) or for adjacency based weights (which ignore the size of the underlying units). One can attempt to deduce the distance at which clustering occurs from the spatial scan statistic by examining the size of the most likely cluster, but as the spatial scan statistic cannot detect dispersion, one cannot detect the simultaneous presence of clustering and dispersion operating at different distances. Further, the spatial

scan statistic has some challenges with areal data since large neighboring units that can have centroids which are relatively far from each other. The PAPF method improves on the 0/1 indicator used by the spatial scan statistic by allowing observed units to be partially inside a zone of interest.

While the average nearest neighbor method and the traditional Ripley's K method are often used to assess areal data for clustering by mapping each areal unit to its centroid, these methods were not designed for areal data. Our simulation study shows that applying these methods in this manner results in a highly inflated type I error rate. In fact, in many settings these methods had a 100% type I error rate. Since such an approach is the default method used by ArcGIS software, these results are concerning.

To provide a means of testing binary areal data for clustering and dispersion at specific distances, we developed the positive area proportion function (PAPF). The PAPF is motivated by Ripley's K-function, and has a similar interpretation. The PAPF method quantifies the average proportion of positive area within a specified distance of each positive unit centroid. The PAPF can be used to perform a hypothesis test for the presence of spatial clustering or dispersion by comparing the observed PAPF test statistic to the distribution of the PAPF test statistic under the null hypothesis of a stationary and independent areal process, which is equivalent to the well-studied random labelling hypothesis after conditioning on the number of positive units. Simulation studies demonstrated that PAPF hypothesis testing procedure maintains its nominal type I error rate and has high power to detect a variety of spatial patterns, including clustering and dispersion at a variety of distances. The PAPF generally displayed comparable or higher power to other methods.

To our knowledge, the PAPF method is the only method for detecting distance-specific patterns in binary areal data. The ability to detect patterns at specific distances is of critical importance in many fields that use and evaluate spatial relationships including urban planning, regional science, conservation biology, public health, epidemiology and many others. For example, the ecological implications of small-distance clusters of conserved land are quite different than the implications of clustering across large distances; addressing many small clusters of food desert census tracts requires a different policy approach than addressing a single expansive swath of food desert tracts. Our method allows researchers to identify the spatial scale at which any clustering or dispersion is operating, facilitating a deeper understanding of the processes at work. Our method also provides a reliable alternative to the misapplication of Ripley's K-function to areal data.

To facilitate the use of our method, R code which implements the PAPF method and performs the necessary Monte Carlo simulations has been made available online at https://github.com/scwatson812/PAPF. The computational expense of the method increases with the number of positive units. When the number of observed areal units is large, the Monte Carlo simulations may be run in parallel to reduce computation time. The development of faster methods for approximating the null distribution is an excellent area for future work. Future work could also consider the extension of the PAPF to continuous data or categorical data with more than two categories. While this work considered the areal units as fixed and only the designation of being 'positive' as random, it is possible to treat both the

locations/boundaries of the areal units and the designation of being positive as random. Such an application might arise when partitioning US states into congressional districts, with the 'positive' districts being those dominated by a particular political party. In such a case, one could treat the centiods of the areal units as a marked point process, with the $\widehat{M}_i(r, \boldsymbol{y})$ s as the associated marks. One could then analyze the resulting marked point process via the associated second order marked random measure and the marked K_{mm}-function.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

SS and AM were supported in part by the Research Center for Child Well-Being [NIGMS P20GM130420]. SS, AZ, and AM were supported in part by the Centers for Disease Control [5 U19 DD 001218]. SS, AO, DW, and CD were supported in part by the National Science Foundation [CNH-L 1518455]. The funding sources played no role in study design, data collection, data analysis, or manuscript publication.

References

[1]. Andersen M (1992). Spatial analysis of two-species interaction. Oecologia, 91:134–140. [PubMed: 28313385]

[2]. Arbia G, Espa G, and Quah D (2009). A class of spatial econometric methods in the empirical analysis of clusters of firms in the space, pages 81–103. Physica-Verlag HD, Heidelberg.

[3]. Baddeley A, Gregori P, Mateu J, Stoica R, and Stoyan D (2005). Case Studies in Spatial Point Process Modeling. Springer, New York.

[4]. Baddeley A, Møller J, and Waagepetersen R (2000). Non- and semiparametric estimation of interaction in inhomogeneous point patterns. Statistica Neerlandica, 54(3):329–350.

[5]. Baddeley A, Rubak E, and Turner R (2015). Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press, London.

[6]. Balbi M, Petit EJ, Croci S, Nabucet J, Georges R, Madec L, and Ernoult A (2019). Ecological relevance of least cost path analysis: An easy implementation method for landscape urban planning. Journal of Environmental Management, 244:61–68. [PubMed: 31108311]

[7]. Baldwin R and Leonard P (2015). Interacting social and environmental predictors for the spatial distribution of conservation lands. PLOS ONE, 10(10).

[8]. Bene C and Rataj J (2004). Stochastic Geometry: Selected Topics. Kluwer Academic Publishers.

[9]. Bivand RS, Pebesma E, and Gomez-Rubio V (2013). Applied spatial data analysis with R, Second edition. Springer, NY.

[10]. Caprarelli G and Fletcher S (2014). A brief review of spatial analysis concepts and tools used for mapping, containment and risk modelling of infectious diseases and other illnesses. Parasitology, 141(5):581–601. [PubMed: 24476672]

[11]. Centers for Disease Control and Prevention (2021). Trends and maps, https://nccd.cdc.gov/dnpao_dtm/rdpage.aspx?rdreport=dnpao_dtm.explorebytopic&islclass=ows&isltopic=&go=go.

[12]. Chiu SN, Stoyan D, Kendall W, and Mecke J (2013). Stochastic Geometry and Its Applications. John Wiley & Sons.

[13]. Clark PJ and Evans FC (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology, 35(4):445–453.

[14]. Cuzick J and Edwards R (1990). Spatial clustering for inhomogeneous populations (with discussion). Journal of teh Royal Statistical Society, Series B, 52.

[15]. Daley DJ and Vere-Jones D (1998). An Introduction to the Theory of Point Processes. Springer, New York.

[16]. Davarpanah A, Babaie HA, and Dai D (2018). Spatial autocorrelation of neogene-quaternary lava along the Snake River Plain, Idaho, USA. Earth Science Informatics, 11(1):59–75.

[17]. Diamond J (1975). The island dilemma: lessons of modern biogeographic studies for the design of natural reserves. Biological Conservation, 7:129–146.

[18]. Diggle P (1983). Statistical analysis of spatial point patterns. Academic Press, London.

[19]. Diggle P and Chetwynd A (1991). Second-order analysis of spatial clustering for inhomogeneous populations. Biometrics, 47(3):1155–63. [PubMed: 1742435]

[20]. Dixon PM (2014). Ripley's K Function. John Wiley & Sons, Ltd.

[21]. Donaldson L, Wilson R, and Maclean I (2016). Old concepts, new challenges: adapting landscape-scale conservation to the twenty-first century. Biodiversity and Conservation, 26(3):527–552. [PubMed: 32269427]

[22]. Duranton G and Overman HG (2005). Testing for localization using micro-geographic data. The Review of Economic Studies, 72(4):1077–1106.

[23]. ESRI (2021). Arcmap 10.5.1: Multi-distance spatial cluster analysis (Ripley's K function) (spatial statistics).

[24]. Evans JS (2021). spatialEco. R package version 1.3–6.

[25]. Gartner DR, Taber DR, Hirsch JA, and Robinson WR (2016). The spatial distribution of gender differences in obesity prevalence differs from overall obesity prevalence among us adults. Annals of Epidemiology, 26(4):293–298 [PubMed: 27039046]

[26]. Gatrell AC, Bailey TC, Diggle PJ, and Rowlingson BS (1996). Spatial point pattern analysis and its application in geographical epidemiology. Transactions of the Institute of British Geographers, 21(1):256–274.

[27]. Getis A and Franklin J (1987). Second-order neighborhood analysis of mapped point patterns. Ecology, 68(3):473–477.

[28]. Getis A and Ord JK (1992). The analysis of spatial association by use of distance statistics. Geographical Analysis, 24(3):189–206.

[29]. Graves R, Williamson M, Belote T, and Brandt J (2019). Quantifying the contribution of conservation easements to large-landscape conservation. Biological Conservation, 232:83–96.

[30]. Haase P (1995). Spatial pattern analysis in ecology based on ripley's k-function: Introduction and methods of edge correction. Journal of Vegetation Science, 6(4):575–582.

[31]. Haddad NM, Brudvig LA, Clobert J, Davies KF, Gonzalez A, Holt RD, Lovejoy TE, Sexton JO, Austin MP, Collins CD, Cook WM, Damschen EI, Ewers RM, Foster BL, Jenkins CN, King AJ, Laurance WF, Levey DJ, Margules CR, Melbourne BA, Nicholls AO, Orrock JL, Song D-X, and Townshend JR (2015). Habitat fragmentation and its lasting impact on earth's ecosystems. Science Advances, 1(2):e1500052. [PubMed: 26601154]

[32]. Hodgson JA, Thomas CD, Wintle BA, and Moilanen A (2009). Climate change, connectivity and conservation decision making: back to basics. Journal of Applied Ecology, 46(5):964–969.

[33]. Karunaweera ND, Ginige S, Senanayake S, Silva H, Manamperi N, Samaranayake N, Siriwardana Y, Gamage D, Senerath U, and Zhou G (2020). Spatial epidemiologic trends and hotspots of leishmaniasis, sri lanka, 2001–2018. Emerging infectious diseases, 26.

[34]. Kretser H, Sullivan P, and Knuth B (2008). Housing density as an indicator of spatial patterns of reported human-wildlife interactions in northern new york. Landscape and Urban Planning, 84:282–292.

[35]. Kulldorff M (1997). A spatial scan statistic. Communications in Statistics - Theory and Methods, 26(6):1481–1496.

[36]. Lamichhane S, Sun C, Gordon J, Grado S, and Poudel K (2021). Spatial dependence and determinants of conservation easement adoptions in the united states. Journal of Environmental Management, 296.

[37]. Law R, Illian J, Burslem DFRP, Gratzer G, Gunatilleke CVS, and Gunatilleke IAUN (2009). Ecological information from spatial patterns of plants: Insights from point process theory. Journal of Ecology, 97(4):616–628

[38]. Lee S-K and Lee B (2013). Assessing the appropriateness of the spatial distribution of standard lots using the l-index. Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, 31(6.2):601–609.

[39]. Loosmore NB and Ford ED (2006). Statistical inference using the g or k point pattern spatial statistics. Ecology, 87(8):1925–1931. [PubMed: 16937629]

[40]. Marcon E and Puech F (2003). Evaluating the geographic concentration of industries using distance-based methods. Journal of Economic Geography, 3(4):409–428.

[41]. Marcon E and Puech F (2017). A typology of distance-based measures of spatial concentration. Regional Science and Urban Economics, 62:56–67.

[42]. Marcon E, Puech F, and Traissac S (2012). Characterizing the relative spatial structure of point patterns. International Journal of Ecology, 2012.

[43]. Marj T and Abellan A (2013). Rockfall detection from terrestrial LiDAR point clouds: A clustering approach using R. Journal of Spatial Information Science, 8.

[44]. McLaughlin N and Weeks W (2009). In defense of conservation easements: A response to the end of perpetuity. Wyoming Law Review, 9:1–96.

[45]. Moran PAP (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1/2):17–23. [PubMed: 15420245]

[46]. Paradis E and Schliep K (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35:526–528. [PubMed: 30016406]

[47]. Penttinen A, Stoyan D, and Henttonen HM (1992). Marked Point Processes in Forest Statistics. Forest Science, 38(4):806–824.

[48]. Qiao L, Huang H, and Tian Y (2019). The identification and use efficiency evaluation of urban industrial land based on multi-source data. Sustainability, 11(21).

[49]. Ripley B (1976). The second-order analysis of stationary point processes. Journal of Applied Probability, 13(2):255–266.

[50]. Ripley B (1977). Modelling spatial patterns. Journal of the Royal Statistical Society. Series B (Methodological), 39(2):172–212.

[51]. Ripley B (1981). Spatial Statistics. Wiley, New York, NY.

[52]. Siordia C (2013). Benefits of small area measurements: a spatial clustering analysis on medicare beneficiaries in the usa. Human Geographies - Journal of Studies and Research in Human Geography, 7(‘):53–59.

[53]. Skog L, Linde A, Palmgren H, Hauska H, and Elgh F (2014). Spatiotemporal characteristics of pandemic influenza. BMC Infectious Diseases, 14.

[54]. Tischendorf L and Fahrig L (2000). On the usage and measurement of landscape connectivity. Oikos, 90(1):7–19.

[55]. Wade BJ (2014). Spatial analysis of global prevalence of multiple sclerosis suggests need for an updated prevalence scale. Multiple Sclerosis International, 2014.

[56]. Zgodic A, Eberth JM, Breneman CB, Wende ME, Kaczynski AT, Liese AD, and McLain AC (2021). Estimates of Childhood Overweight and Obesity at the Region, State, and County Levels: A Multilevel Small-Area Estimation Approach. American Journal of Epidemiology. kwab176.

[57]. Zipp KY, Lewis DJ, and Provencher B (2017). Does the conservation of land reduce development? an econometric-based landscape simulation with land market feedbacks. Journal of Environmental Economics and Management, 81:19–37.

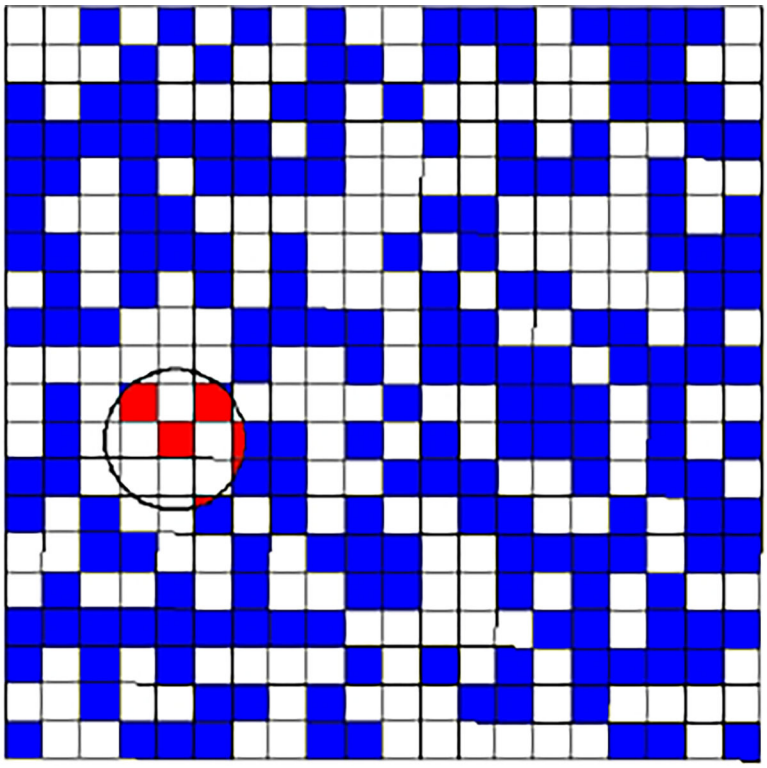

Figure 1:
An illustration of $\mathcal{N}[c(\ell_i, r) \cap a_i^c] + A[c(\ell_i, r) \cap a_i]$ for a realization of an areal process on a 20×20 regular grid. Positive areal units (i.e. units for which $y_i = 1$ are shown in color. The area of the shaded red region is equal to $\mathcal{N}[c(\ell_i, r) \cap a_i^c] + A[c(\ell_i, r) \cap a_i]$, where a_i is the grid cell in the 12th row and 5 th column.

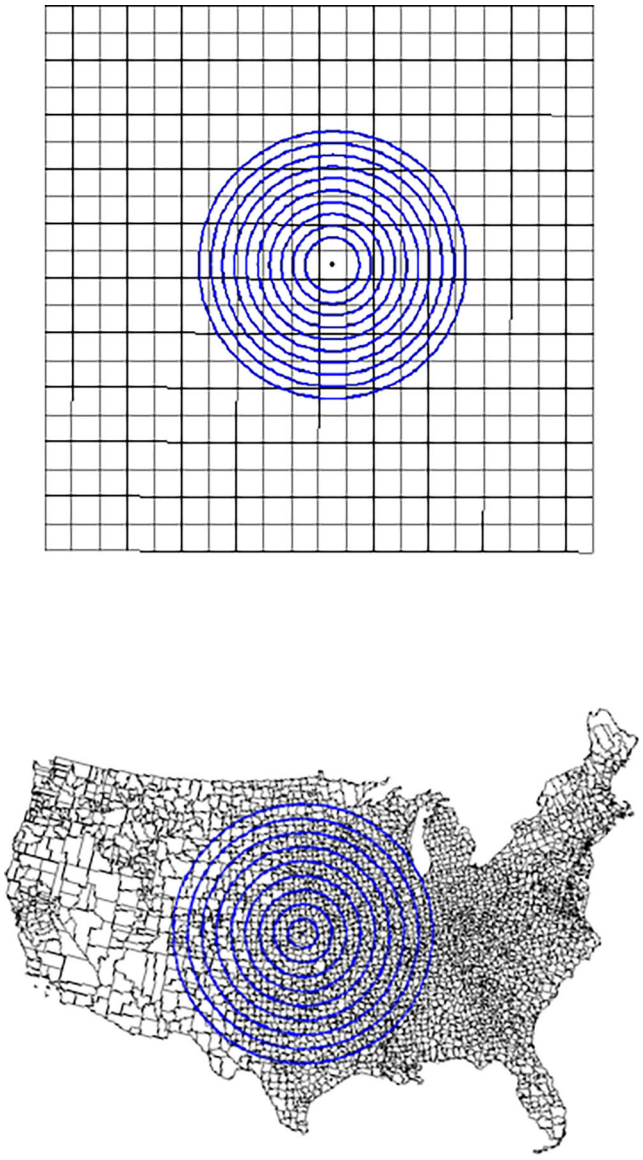

Figure 2:
The 2 study areas considered in the simulation study. The 10 radii at which the positive area proportion function, Ripley's K-function and Ripley's D-function are evaluated are shown for a single location in blue.

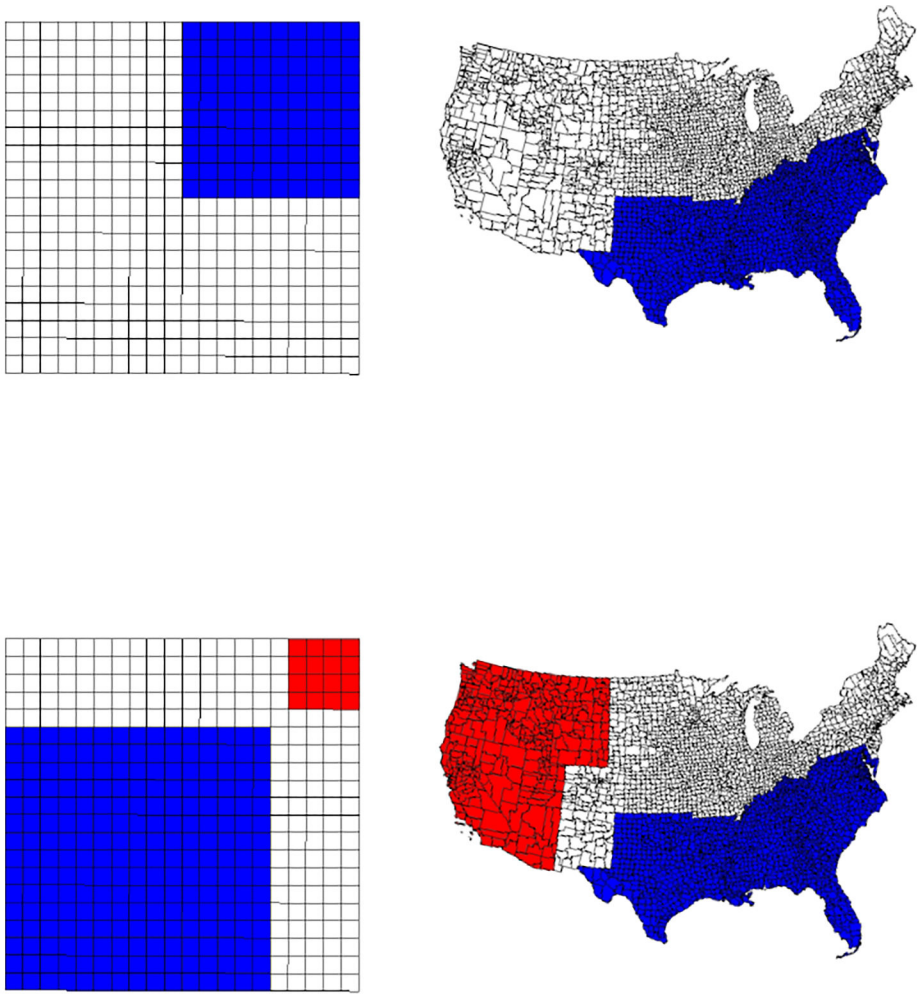

Figure 3:

The top row illustrates the clustered regions for DGMs $C_1 - C_6$ for study areas $\mathscr{A}_1$ (top left), and $\mathscr{A}_2$ (top right). Blue units are q times more likely to be selected than white units under spatial dependence configurations $C_1 - C_6$. The bottom row illustrates the clustered (R_{jc}), dispersed (R_{jd}) and random scatter regions (R_{jr}) for DGMs M_1 and M_2 for $\mathscr{A}_1$ (bottom left) and $\mathscr{A}_2$, (bottom right). Red denotes R_{jc}, blue denotes R_{jd} and white denotes R_{jr}.

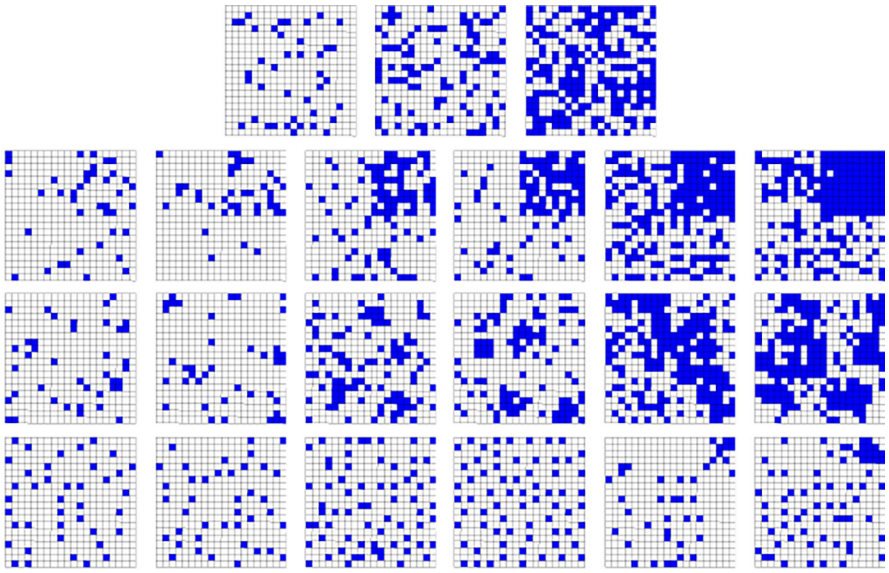

Figure 4:

Examples of observed units generated under each scenario for study area $\mathscr{A}_1$. The first row displays examples of data generated under the null hypothesis of equal probability sampling without replacement (left to right: I_1, I_2, I_3). The second row displays examples of data generated with excess clustering (left to right C_1, C_2, C_3, C_4, C_5, C_6). The third row displays examples of data generated with autocorrelated clustering (left to right C_7, C_8, C_9, C_{10}, C_{11}, C_{12}). The fourth row displays examples of data generated with dispersion or a mixture of clustering and dispersion (left to right D_1, D_2, D_3, D_4, M_1, M_2).

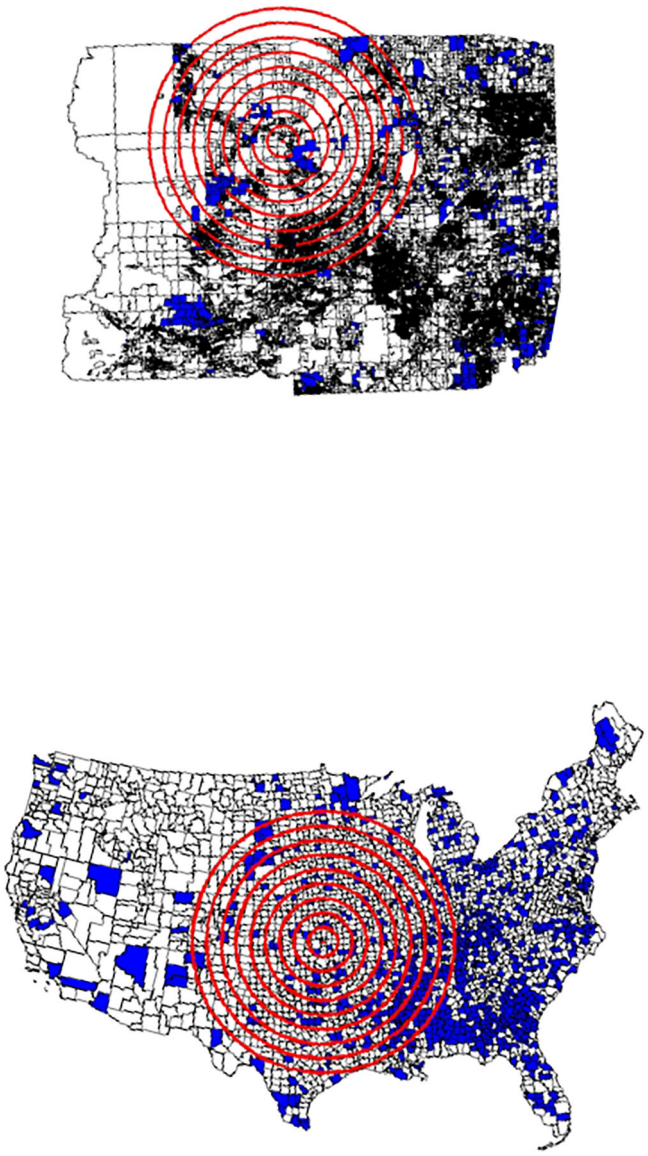

Figure 5:

The top pane displays the 112,819 land parcels in Boulder County, Colorado in 2008. Parcels held as a CE are shown in blue. The radii at which the PAPF method was evaluated are shown in red. Note that areas with many small parcels appear black. The bottom pane displays the 3,108 counties in the contiguous US. Counties with a high rate of childhood overweight/obesity are shown in blue. The radii at which the PAPF method was evaluated are shown in red.

Table 1:

The table provides an overview of the spatial clustering assessment methods considered in the simulation study: the positive area proportion function (PAPF), the global Moran's I statistic (MI), the Getis-Ord general G statistic (GG), the spatial scan statistic (SSS), Ripley's K-function (RK), Ripley's D-function (RD) and the average nearest neighbor method (ANN). The null and alternative hypotheses (both left- and right-tailed) are specified for each method, along with the Monte Carlo procedure used to estimate the null distribution, if applicable.

Method	H_0	H_1 **(Left-tailed)**	H_1 **(Right-tailed)**	**Monte Carlo Procedure**
PAPF	Random labeling (conditional on n)	Dispersion	Clustering	$\mathbf{Y}\sim\text{SWoR}(N, n, \boldsymbol{p})$, $\boldsymbol{P} = \left(N^{-1}, \ldots, N^{-1}\right)$
MI	Random labeling	Negative spatial autocorrelation	Positive spatial autocorrelation	NA
GG	Random labeling	Low-low clustering	High-high clustering	NA
SSS	y_i's are iid Bernoulli(p)	There exists a set $\mathcal{J}$ indexing a (contiguous) cluster of a_is such that y_i's are iid Bernoulli(p_{in}) for $i \in \mathcal{J}$ and iid Bernoulli(p_{out}) for $i \notin \mathcal{J}^*$		$\boldsymbol{Y}\sim\text{SWoR}(N, n, \boldsymbol{p})$, $\boldsymbol{p} = \left(N^{-1}, \ldots, N^{-1}\right)$
RK	centroids arise from HPP†	centroids dispersed	centroids clustered	Generate n points from a continuous uniform distribution on $\mathcal{A}$
RD	Random labeling	Controls more clustered than cases	Cases more clustered than controls	Select cases via $\text{SWoR}(N, \boldsymbol{n}, \boldsymbol{p})$, $\boldsymbol{p} = \left(N^{-1}, \ldots, N^{-1}\right)$
ANN	centroids arise from HPP†	centroids clustered	centroids dispersed	NA

*For left-tailed test $p_{out} > p_{in}$ for right-tailed test, $p_{in} > p_{out}$

†HPP: homogeneous Poisson process

Table 2:

Simulation study results for study area $\mathscr{A}_1$ Journal Pre-proof A1 (the regular grid). Results displayed include the empirical rejection rate (ERR) of the positive area proportion function (PAPF), the global Moran's I statistic (MI), the Getis-Ord general G statistic (GG), the spatial scan statistic method (SSS), Ripley's K-function (RK), Ripley's D-function (RD) and the average nearest neighbor method (ANN). For DGMs $M_1 - M_2$, single-tailed test indicative of clustering are denoted with a C, while dispersion tests are denoted with a D. All tests were conducted at a level of $\alpha = 0.05$.

DGM	Method	ERR	Method	ERR										
				Global	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
I_1	ANN	97.5	RK	100.0	25.8	4.8	18.6	11.2	13.8	2.0	7.4	4.0	1.8	18.0
	SSS	3.0	RD	6.0	6.4	3.6	3.6	7.6	7.6	7.0	7.6	6.0	6.6	10.6
	MI	5.8	PAPF	3.6	4.2	3.6	5.8	4.4	3.6	3.2	3.4	3.2	3.0	3.6
	GG	4.8												
I_3	ANN	100.0	RK	100.0	100.0	100.0	100.0	100.0	100.0	0.2	100.0	75.6	18.6	99.8
	SSS	2.8	RD	4.4	10.0	7.0	7.0	2.0	2.0	2.8	4.4	6.0	5.6	3.4
	MI	5.0	PAPF	7.6	5.8	5.4	8.6	5.4	5.0	6.0	5.0	6.4	4.6	4.8
	GG	6.4												
C_2	SSS	99.8	RD	100.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	MI	97.0	PAPF	100.0	91.6	95.2	99.0	99.6	100.0	100.0	100.0	100.0	100.0	100.0
	GG	96.4												
C_6	SSS	100.0	RD	0.4	0.0	1.4	1.4	2.0	2.0	0.0	2.2	1.0	0.4	0.4
	MI	100.0	PAPF	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	GG	100.0												
C_8	SSS	78.0	RD	91.6	37.2	93.4	93.4	91.6	91.6	79.9	72.2	60.2	54.4	53.2
	MI	97.6	PAPF	93.6	91.8	95.8	97.6	96.0	92.2	87.2	75.0	65.8	60.8	56.6
	GG	97.2												
C_{12}	SSS	93.6	RD	73.6	20.6	83.2	83.2	67.2	67.2	39.0	53.6	41.8	38.4	34.2
	MI	100.0	PAPF	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6	96.6	90.0	86.0
	GG	100.0												
D_2	MI	100.0	RD	100.0	2.6	100.0	100.0	77.4	77.4	56.6	54.2	25.4	24.2	28.4
			PAPF	100.0	100.0	100.0	100.0	98.0	69.8	52.8	41.2	30.0	32.6	24.4

DGM	Method	ERR	Method	Global	ERR									
					r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
D_4	MI	100.0	RD	100.0	100.0	100.0	100.0	82.2	82.2	53.6	82.4	61.6	57.8	30.8
			PAPF	100.0	100.0	100.0	100.0	100.0	98.4	100.0	98.6	97.4	92.0	91.8
M_1	MI D	0.0	RD D	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MI C	30.6	RD C	98.8	0.0	42.6	42.6	99.4	99.4	99.2	99.0	98.0	86.8	73.4
	SSS c	0.0	PAPF D	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	GG C	15.6	PAPF C	100.0	28.8	33.8	72.8	97.2	99.8	100.0	100.0	100.0	99.4	97.0

Table 3:

Simulation study results for study area $\mathscr{A}_2$ (the US counties). Results displayed include the empirical rejection rate of the positive are proportion function (PAPF), the global Moran's I statistic (MI), the Getis-Ord general G statistic (GG), the spatial scan statistic method (SSS) Ripley's K-function (RK), Ripley's D-function (RD) and the average nearest neighbor method (ANN). For DGMs $M_1 - M_2$, single-tailed test indicative of clustering are denoted with a C, while dispersion tests are denoted with a D. All tests were conducted at a level of $\alpha = 0.05$.

DGM	Method	ERR	Method	ERR										
				Global	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
I_1	ANN	26.8	RK	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	SSS	3.6	RD	5.2	0.8	5.0	5.8	5.0	5.6	6.4	6.8	7.0	6.0	5.4
	MI	3.8	PAPF	7.6	5.2	4.8	7.0	7.2	8.6	6.8	6.8	7.0	6.2	6.4
	GG	4.2												
I_3	ANN	100.0	RK	100.0	76.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	SSS	2.8	RD	2.8	0.0	4.2	3.6	2.2	4.6	3.4	2.6	2.8	2.6	2.8
	MI	7.2	PAPF	6.6	5.0	5.8	5.2	7.0	6.2	5.6	5.4	6.0	6.0	6.4
	GG	7.0												
C_2	SSS	100.0	RD	100.0	3.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	MI	100.0	PAPF	100.0	99.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	GG	100.0												
C_6	SSS	100.0	RD	100.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	MI	100.0	PAPF	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	GG	100.0												
C_8	SSS	87.8	RD	98.2	1.8	99.6	67.2	31.2	24.8	22	16.4	15.4	16.2	16.4
	MI	100.0	PAPF	100.0	11.4	99.0	78	42.0	32.0	24.0	22.2	20.6	18.2	16.2
	GG	100.0												
C_{12}	SSS	98.6	RD	23.0	0.0	16.6	19.8	17.2	17.6	19.6	20.4	18.6	18.6	19.0
	MI	100.0	PAPF	97.2	8.0	98.2	57.2	36.2	28.0	25.4	24.0	22.6	21.6	21.0
	GG	100.0												
D_2	MI	100.0	RD	93.8	0.0	98.0	39.8	15.6	5.4	4.2	3.8	3.6	3.6	3.6
			PAPF	98.8	1.2	100.0	40.8	17.2	7.0	6.8	5.4	3.8	2.6	2.4

DGM	Method	ERR	Method	Global	ERR									
					r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
D_4	MI	100.0	RD	99.2	0.0	99.2	55	14.4	8.4	8.0	5.2	4.8	4.6	4.4
			PAPF	100.0	0.0	100.0	64.0	22.0	11.6	6.0	3.4	2.0	2.6	1.2
M_2	MI D	30.0	RD D	20.4	0.0	0.4	0.0	0.0	0.2	0.8	1.8	4.2	14.6	40.2
	MI C	0.0	RD C	4.8	2.0	3.2	31.4	9.8	2	0.4	0.0	0.0	0.0	0.0
	SSS	100.0	PAPF D	87.0	90.8	14.8	0.2	0.6	2.0	4.6	8.4	14.4	25.6	36.2
	GG	0.0	PAPF C	24.8	0.0	9.0	31.2	23.8	16.6	10.4	6.4	4.0	1.8	0.0