Electronic Journal of Statistics
Vol. 13 (2019) 4367–4390
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1626

Distributional properties and estimation in spatial image clustering

Zijuan Chen[1] **and Suojin Wang**[2]*

[1] *Department of Statistics, Texas A&M University, College Station, TX 77843, USA* e-mail: zijuan@stat.tamu.edu

[2] *Department of Statistics, Texas A&M University, College Station, TX 77843, USA* e-mail: sjwang@stat.tamu.edu

Abstract: Clusters of different objects are of great interest in many fields, such as agriculture and ecology. One kind of clustering methods is very different from the traditional statistical clustering analysis, which is based on discrete data points. This method of clustering defines clusters as the connected areas where a well-defined spatial random field is above certain threshold. The statistical properties, especially the distributional properties, of the defined clusters are vital for the studies of related fields. However, the available statistical techniques for analyzing clustering models are not applicable to these problems. We study the distribution properties of the clusters by defining a distribution function of the clusters rigorously and providing methods to estimate the spatial distribution function. Our results are illustrated by numerical experiments and an application to a real world problem.

Keywords and phrases: Distributional properties, image processing, spatial statistics.

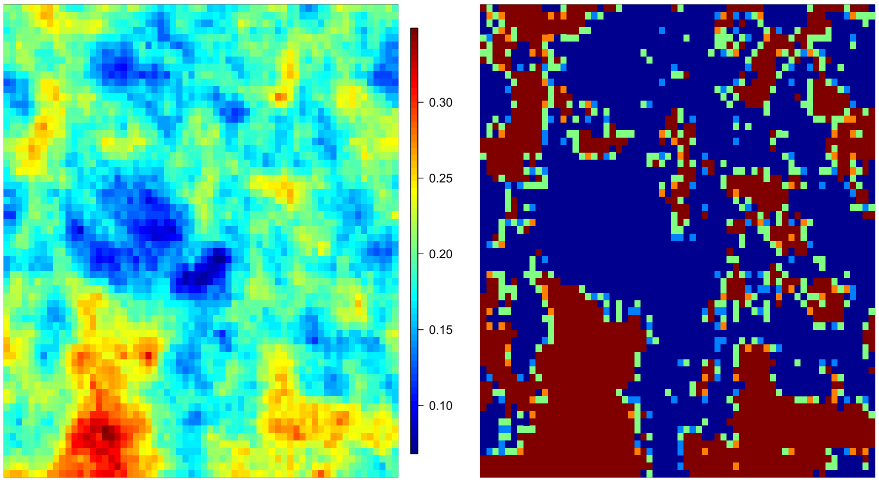
Received November 2018.

1. Introduction

Analyses of clusters of soil, water and species have been of great interest in agriculture, ecology and hydrology; see, for example, Asnera and Warner (2003); Wootton (2001); Martin and Goldenfeld (2006); Sole (2007). For instance, clusters of trees have been analyzed frequently since their properties are closely related to the environmental conditions. Many studies have been carried out based on spatial modeling (for instance, see Chen, Mohanty and Rodriguez-Iturbe (2017)) and simulations since data of clusters are often hard to collect. Real data have been used to verify whether models and simulations are capable of reproducing patterns observed in nature (see Scanlon et al. (2007)).

In ecology, the analyses of clusters often focus on the spatial properties, such as the size of an individual cluster and the locations of the centers of clusters, assuming the object of interest is modeled by a continuous random process $y(s, \omega)$, where s is the parameter of space and ω is some sample point. See Rodriguez-Iturbe et al. (2019) for modeling and simulations, and Staver et al.

*Corresponding Author.



(a) A spatial random field (b) Clusters of the random field

Fig 1. *(a) is a realization of the random field y on the squared spatial domain D. In (b), the clusters in red are the areas where $y > 0.2$.*

(2019) for real data analysis. For instance, $\{y(s,\omega) : s \in D, \omega \in \Omega\}$ represents the soil moisture in some area D, and $\{s \in D : y(s) > 0.2\}$ could be the set of interest where the number of trees might be relatively large. Figure 1 is an example for illustration. Figure 1a is a realization of the random field y, and the clusters in red in Figure 1b are the areas of interest (where $y > 0.2$). However, traditional statistical cluster analysis mostly studies methods of grouping a set of objects with similar properties based on discrete data points (see Azzalini and Torelli (2007); Cattelan and Varin (2018); Li (2006); McNicholas (2016); Menardi and Azzalini (2014); Steinwart (2015) for existing clustering methods). In contrast, we are interested in the distributional properties of the clusters based on a stochastic process defined on a continuous domain, assuming that the clusters can be easily identified. Currently there are no available statistical tools to study data about clusters in hydrology and ecology. The most common and fundamental property about this kind of cluster analysis is the size of an individual cluster, which contains much information about the environmental conditions. Data of sizes of clusters of different objects, such as canopies, have been collected and the "distribution" of the size of individual cluster has been studied by many researchers. However, the mathematical and statistical definition of the size of individual cluster has not been well defined and studied, though samples can be easily collected from images obtained from many different ways, such as remote sensors. Note that since samples of clusters are correlated, they can not be regarded as independent samples from an unknown distribution. In fact, given a well-defined spatial statistical model, it is generally difficult to define a random variable as the size of an individual cluster and study its distribution.

Without a well defined distribution function of the clusters, it is difficult for researchers to study the statistical properties of the data of image clustering and perform efficient statistical inferences. Much information in the data is not utilized, which is possible to result in inaccurate conclusions. Therefore, it is important to have a well defined distribution, and derive an efficient method to estimate the defined distribution function. Then we are able to get the distributional properties of the clusters, from which more accurate conclusions can be drawn.

The definition of the distribution function of image clusters in spatial random fields is crucial and cannot be done in a usual way, namely induced by a random variable. Thus in this paper we define the distribution of the size of an individual cluster in a special way, without defining a random variable to be the size of an individual cluster. The estimation of the defined distribution function will be introduced and the asymptotic properties of the estimators will be investigated, which enable us to make statistical inferences and hypothesis tests.

In the following sections, basic definitions are described in Section 2. Then main results are presented in Section 3, which include some statistical properties in Section 3.1, the definition of the distribution function and its estimation in Section 3.2, and applications to Gaussian random fields in Section 3.3. A simulation study and a data analysis are carried out in Sections 4 and 5, respectively. Some concluding remarks are given in Section 6.

2. Preliminaries

Let D be the spatial domain and $(\Omega, \mathscr{F}, P)$ be the probability space of interest where a random process $y(s, \omega)$, $s \in D$, $\omega \in \Omega$, is defined. Without loss of generality (WLOG), we assume $D = [0,1]^2$ for simplicity. To define the clusters of interest on D, we need the definitions of open sets and connected sets as follows, which are standard definitions from point set topology (Gemignani (1990), Chapter 9).

Definition 2.1. (Open in D) A set $S \subset D$ is called an open set in D if there exists an open set $\tilde{S} \subset \mathbb{R}^2$ such that $S = \tilde{S} \cap D$.

Definition 2.2. (Connected and disconnected in D) A set $S \subset D$ is called a disconnected set in D if it can be divided into two disjoint nonempty open sets in D, i.e., there exist A and B open in D such that $A \neq \emptyset$, $B \neq \emptyset$, $A \cap B = \emptyset$, and $S = A \cup B$. Otherwise, S is called a connected set in D.

All open sets and connected sets are meant to be open or connected in D if not specified.

Denote the area of interest by $A = A(\omega)$, which is a subset of D depending on y. For instance, A can be the area with soil moisture greater than some value c:

$$A = A(\omega) = \{s \in D : y(s, \omega) > c\}.$$

For simplicity, we assume $c = 0$ and

$$A = A(\omega) = \{s \in D : y(s, \omega) > 0\}$$

in the following sections. However, the conclusions still hold when A is more complicated. We assume further that $\forall \omega \in \Omega$, $y(s, \omega)$ is continuous in D and $\forall s \in D$, $P(y(s) > 0) > 0$. Then for any fixed ω, $A = A(\omega)$ is open in D (an open set with respect to the topology of D). By a basic property of $\mathbb{R}^2$, A can be represented as

$$A = \bigcup_{m=1}^{M} A_m, \tag{2.1}$$

where A_m's are mutually exclusive, open and connected subsets of D, also known as the connected components of A, and M is some positive integer or ∞, depending on ω. We say s' and s'' are connected if there exists m such that $s', s'' \in A_m$. These A_m's can be regarded as "islands" in the spatial domain D. When two islands are "very close", we consider them as a cluster since they would affect each other. This consideration is reasonable in applications: think of two tree canopies which are very close. They would probably be the same species and compete with each other for groundwater. Therefore, they should be considered as one cluster. More clearly, for some fixed positive number δ, if the distance between two islands is less than δ, they should be in the same cluster. We can now define clusters formally by introducing the following relation.

Definition 2.3. Suppose $s', s'' \in A$ and $\delta > 0$. We say that s' and s'' belong to the same cluster, denoted by $s' \sim s''$, if there exist $0 \leqslant n < \infty$ and $s_1, s_2, ..., s_n \in A$, such that for each $i = 1, 2, ..., n+1$, $\|s_{i-1} - s_i\| < \delta$ ($s' = s_0$, $s'' = s_{n+1}$).

One may ask whether δ could tend to 0 when the resolution k tends to infinity. In general, in practice people would identify clusters of specific objects with $\delta = 0$. Introduce δ here is for the purpose of the theoretical derivations in probability. In applications, a fixed small δ would lead to a negligible difference compared to $\delta = 0$. For instance, when a study is focusing on the islands in oceans one can set δ equal to 1 cm, and when studying clusters of soil moisture one can set δ equal to 1 nm.

The relation $\sim$ groups together the points of A which are close to each other. Since $\sim$ is reflective, symmetric and transitive, it is an equivalence relation in A. Let $S/\sim$ denote the quotient space of a set S by an equivalence relation $\sim$. Now we can define clusters as follows.

Definition 2.4. (Clusters). The equivalence classes partitioned by the equivalence relation $\sim$ in A (elements of $A/\sim$) are called clusters and denoted by $\{C_\beta\}_{\beta \in \Delta}$.

The definition of clusters seem to be abstract and complicated at a first glance. In fact, each cluster C_β defined in Definition 2.4 is just a union of "islands" (connected components) of A, which is shown in the following Theorem. Let

$$d(s, S) = \inf_{s' \in S} \|s' - s\|, \quad d(S', S'') = \inf_{s' \in S', s'' \in S''} \|s' - s''\|$$

denote the distance between a point and a set and the distance between two sets in $\mathbb{R}^2$, respectively.

Theorem 2.1. *Suppose that $\{C_\beta\}_{\beta\in\Delta}$ are the clusters defined in Definition 2.4 and $\{A_m\}_{m=1}^{\infty}$ are the connected components of A. Then for any cluster $C = C_\beta$ of A, we have*

$$C = \bigcup_{i: A_i \subset C} A_i.$$

Proof. See Appendix A.1. □

Let $\lambda(\cdot)$ denote the Lebesgue measure in $\mathbb{R}^2$ and $\overline{\mathbb{R}}_+ = [0, \infty)$. Then we have

Definition 2.5. (Number of clusters). For $x \in \overline{\mathbb{R}}_+$, define

$$N_x = card\left(\{\beta \in \Delta : \lambda(C_\beta) > x\}\right),$$

which is the number of clusters with Lebesgue measure greater than x.

When $x = 0$, $N_x = N_0$ is the total number of clusters. Let

$$B(s, r) = \{s' : \|s' - s\| < r\}$$

denote the open balls in $\mathbb{R}^2$. For each $\beta \in \Delta$, define

$$\tilde{C}_\beta = \{s \in D : d(s, C_\beta) < \delta/2\}.$$

Then $\tilde{C}_\beta$'s are open and mutually exclusive since by Definition 2.3 and 2.4, $d(C_\beta, C_{\beta'}) \geqslant \delta$ if $\beta \neq \beta'$. Furthermore, each $\tilde{C}_\beta$ contains an open set $B(s_\beta, \delta) \cap D$, where $s_\beta \in C_\beta$. Therefore, $\lambda(\tilde{C}_\beta) \geqslant \pi\delta^2/16$ (since $\lambda(B(s_\beta, \delta) \cap D) \geqslant \pi(\delta/2)^2/4$ for sufficient small δ) and

$$0 \leqslant N_x \leqslant N_0 \leqslant \frac{\lambda(D)}{\pi}\frac{16}{\delta^2} = \frac{16}{\pi\delta^2} < \infty. \tag{2.2}$$

In applications, collected data are often transferred into images with certain resolutions. Therefore, the information we have is based on pixels or grid points. In this paper, we assume that

$$G_k = \left\{\frac{1}{2^k}, \frac{2}{2^k}, ..., \frac{2^k - 1}{2^k}, 1\right\}^2 \subset \mathbb{R}^2, \quad k = 1, 2, ...,$$

are the sets of grid points, and

$$G = \bigcup_{k=1}^{\infty} G_k = \lim_{k\to\infty} G_k$$

is the set of all grid points when the resolution goes to infinity.

Now we have similar definitions for the grid points.

Definition 2.6. Suppose $s', s'' \in A \cap G_k$ for some k and $\delta > 0$. We say that s' and s'' belong to the same cluster of G_k, denoted by $s' \overset{G_k}{\sim} s''$, if there exist $0 \leqslant n < \infty$ and $s_1, s_2, ..., s_n \in A \cap G_k$, such that for each $i = 1, 2, ..., n+1$, at least one of the following two conditions is satisfied ($s_0 = s', s_{n+1} = s'', s_i = (y_i, z_i), i = 0, 1, ..., n+1$):

(1) $|y_{i-1} - y_i| + |z_{i-1} - z_i| = 2^{-k}$ (s_i is in the Von Neumann Neighborhood of s_{i-1});

(2) $\|s_{i-1} - s_i\| < \delta$ (s_{i-1} and s_i are very close).

We say $s' \overset{G}{\sim} s''$ if there exists $K > 0$ such that for all $k > K$, $s' \overset{G_k}{\sim} s''$.

Definition 2.6 has one more "neighborhood" condition than Definition 2.3. The reason is that when identifying clusters in the continuous domain D, we only need to consider the true distance between points. However, for finite resolution k, the true image, $A(\omega) = \{s \in D : y(s, \omega) > 0\}$ is approximated by the pixels $y(G_k, \omega)$. When $2^{-k} < \delta$, no points of G_k would satisfy condition (2) of Definition 2.6, and each cluster can only contain one pixel. In practice, it is reasonable to group these points in $A \cap G_k$ that are neighbors when studying the cluster properties with finite resolution k. Therefore, condition (1) is added to Definition 2.6, although it is not necessary for theoretical derivations for $\delta > 0$ and $k \to \infty$.

Definition 2.7. (Clusters of grids). The equivalence classes partitioned by the equivalence relation $\overset{G_k}{\sim}$ in G_k (elements of $A \cap G_k / \overset{G_k}{\sim}$) are called clusters of G_k and denoted by $\{C_{\beta,k}\}_{\beta \in \Delta_k}$.

Definition 2.8. (Number of clusters on grids). For $x \in \overline{\mathbb{R}}_+$, define

$$N_{x,k} = card\left(\{\beta \in \Delta_k : \lambda_k(C_{\beta,k}) > x\}\right),$$

where λ_k is the counting measure defined in G_k (each grid point with mass 4^{-k}).

Similarly to Equation (2.2), define

$$\tilde{C}_{\beta,k} = \{s \in D : d(s, C_{\beta,k}) < \delta/2\},$$

and $\tilde{C}_{\beta,k}$'s are open and mutually exclusive. Thus we have

$$0 \leqslant N_{x,k} \leqslant N_{0,k} \leqslant \frac{\lambda(D)}{\pi(\delta/2)^2/4} = \frac{16}{\pi\delta^2} < \infty. \tag{2.3}$$

3. Main results

3.1. Some statistical properties

To study the statistical properties of the clusters from real data, we need to ensure that when the resolution gets higher, the plot with pixels obtained from the data becomes closer to the true spatial random field. In other words, the clusters of G_k's should be almost the same as the true clusters. Since the definition of clusters is based on connectivity, we need the following theorem, which shows the relationship between the connectivity of girds and the connectivity in D.

Theorem 3.1. *Suppose $s', s'' \in A \cap G$. Then $s' \overset{G}{\sim} s''$ if and only if $s' \sim s''$.*

Proof. See Appendix A.2. □

Before we define the distribution function through $N_{x,k}$, we should make sure that $N_{x,k}$ is well defined, i.e., we should make sure that $N_{x,k}$ is a random variable.

Theorem 3.2. *For any $x \in \overline{\mathbb{R}}_+$ and $k \in \mathbb{N}_+$, $N_{x,k} : \Omega \mapsto \mathbb{N}$ is a random variable.*

Proof. See Appendix A.3. □

Let $\emptyset$ be the empty set and

$$\partial S = \{s \in D : \forall \epsilon > 0, B(s, \epsilon) \cap S \neq \emptyset, B(s, \epsilon) \cap S^c \neq \emptyset\}$$

denote the boundary of $S \subset \mathbb{R}^2$. The next lemma and theorem show the behavior of $N_{x,k}$ when the resolution k goes to infinity.

Lemma 3.1. *Suppose that the random process y satisfies*

$$\lambda\left(\partial A(\omega)\right) = \lambda\left(\partial\left\{s \in D : y(s, \omega) > 0\right\}\right) = 0, \quad \forall \omega \in \Omega. \tag{3.1}$$

Then for any $x \in \overline{\mathbb{R}}_+$, N_x is a random variable. Define

$$U_x = \{\omega \in \Omega : \lambda(C_i(\omega)) \neq x, \quad i = 1, 2, ..., N_0(\omega)\}, \tag{3.2}$$

where $C_1, C_2, ..., C_{N_0}$ denote the clusters of interest. Then $N_{x,k} \to N_x$ as $k \to \infty$ for all $\omega \in U_x$.

Proof. See Appendix A.4. □

Theorem 3.3. *Suppose that the random process y satisfies the condition in Lemma 3.1. Then $\forall x \in \overline{\mathbb{R}}_+$, $U_x \subset \Omega$ is measurable with respect to $\mathscr{F}$. Define*

$$V = \{x \in \overline{\mathbb{R}}_+ : P(U_x) = 1\}. \tag{3.3}$$

Then $\overline{\mathbb{R}}_+ \backslash V$ is at most countable. In other words,

$$N_{x,k} \overset{a.s.}{\to} N_x$$

except for an at most countable set in $\overline{\mathbb{R}}_+$.

Proof. According to the proof of Lemma 3.1, Equation (A.1) holds for all $\omega \in \Omega$. Thus for any $x \in \overline{\mathbb{R}}_+$, we can rewrite U_x as

$$U_x = \bigcup_{h=1}^{\infty} \bigcup_{j=1}^{\infty} \bigcap_{k=j}^{\infty} \left\{\omega \in \Omega : N_{x-1/h,k}(\omega) = N_{x+1/h,k}(\omega)\right\},$$

which indicates that U_x is measurable.

Now suppose $\overline{\mathbb{R}}_+ \backslash V$ is uncountable. Since

$$\overline{\mathbb{R}}_+ \backslash V = \{x \in \overline{\mathbb{R}}_+ : P(U_x) < 1\} = \bigcup_{m=1}^{\infty} \left\{x \in \overline{\mathbb{R}}_+ : P(U_x) \leqslant 1 - \frac{1}{m}\right\},$$

there exists $m_0 \in \mathbb{N}_+$ such that $\left\{x \in \overline{\mathbb{R}}_+ : P(U_x) \leqslant 1 - \frac{1}{m_0}\right\}$ is uncountable. Then we can choose a sequence

$$\{x_i\}_{i=1}^{\infty} \subset \left\{x \in \overline{\mathbb{R}}_+ : P(U_x) \leqslant 1 - \frac{1}{m_0}\right\}$$

such that $x_i \neq x_j$ if $i \neq j$. Let $U_x^c = \Omega \backslash U_x$. Then U_x^c consists of ω's such that $y(s,\omega)$ has at least one cluster with size x. Since, by Equation (2.2), the number of clusters is finite, $\forall \omega \in \Omega$, it cannot belong to infinite many sets in $\{U_{x_i}^c\}_{i=1}^{\infty}$. In other words, $\limsup_{i\to\infty} U_{x_i}^c = \emptyset$. Thus by Fatou's Lemma,

$$0 = P(\emptyset) = P(\limsup_{i\to\infty} U_{x_i}^c) \geqslant \limsup_{i\to\infty} P(U_{x_i}^c) \geqslant \limsup_{i\to\infty} \frac{1}{m_0} = \frac{1}{m_0},$$

which is a contradiction.

□

By Equation (2.2) we know that N_0 is bounded and $E(N_0) < \infty$. Besides, $P(N_0 > 0) \geqslant P(y((0,0)) > 0) > 0$ and $E(N_0) > 0$. Define

$$\tilde{F}(x) = 1 - \frac{E(N_x)}{E(N_0)}, \quad x \in V. \tag{3.4}$$

Then we have the following corollary:

Corollary 3.1. *Suppose that y satisfies the conditions in Lemma 3.1 and $x_0 \in V$. Then*

$$\lim_{\substack{x\to x_0 \\ x\in V}} N_x = N_{x_0}, \quad a.s., \tag{3.5}$$

and $\tilde{F}(x)$ is continuous in V.

Proof. Fix $\omega \in U_{x_0}$. Let $C_1, C_2, \ldots, C_{N_0}$ be the clusters and

$$\gamma = \min_{1\leqslant i\leqslant N_0} |\lambda(C_i) - x_0|.$$

Then $\gamma > 0$ and when $|x - x_0| < \gamma$ and $x \in V$, $N_x = N_{x_0}$. Since $P(U_{x_0}) = 1$, we have

$$\lim_{\substack{x\to x_0 \\ x\in V}} N_x = N_{x_0}, \quad a.s.$$

Finally, by the Dominated Convergence Theorem, we have

$$\lim_{\substack{x\to x_0 \\ x\in V}} \tilde{F}(x) = 1 - \frac{1}{E(N_0)} \lim_{\substack{x\to x_0 \\ x\in V}} E(N_x) = 1 - \frac{E(N_{x_0})}{E(N_0)} = \tilde{F}(x_0).$$

Hence $\tilde{F}(x)$ is continuous in V. □

Remark 1. Now it is obvious that $\tilde{F}(x)$ has all the properties of distribution functions, but it is only defined in V. Note that V is dense in $\overline{\mathbb{R}}_+$, we can let the right limit of $\tilde{F}(x)$ be the well-defined distribution function. It may look strange that the distribution function of clusters has the form (3.4). One can think of $E(N_0)$ as the expected number of clusters, and $E(N_x)$ as the expected number of clusters with size greater than x. Then $E(N_x)/E(N_0)$ can be regarded as the proportion of clusters with size greater than x, which should be $1 - \tilde{F}(x)$ by definition. This suggests that the right limit of $\tilde{F}(x)$ should be the cumulative distribution function of the size of clusters. This idea comes from practical problems in various fields, such as agriculture, ecology and hydrology (Pascual et al. (2002); Scanlon et al. (2007); Staver et al. (2019)). More examples will be given after the Empirical Distribution Function (EDF) is defined in the next section.

3.2. The distribution function and its estimation

Now we are ready to formally obtain the distribution function of cluster size.

Theorem 3.4. *Suppose y satisfies the conditions in Lemma 3.1 and $\tilde{F}(x)$ is defined as in Corollary 3.1. Define*

$$F(x) = \inf_{z \in V \cap [x,\infty)} \tilde{F}(z), \quad x \in \overline{\mathbb{R}}_+, \tag{3.6}$$

and $F(x) = 0$ when $x < 0$. Then $F(x)$ has the following properties:
(1) $F(x)$ is non-decreasing;
(2) $F(-\infty) = 0$, $F(+\infty) = 1$;
(3) $F(x)$ is right continuous.
Therefore, $F(x)$ is a valid distribution function defined in $\mathbb{R}$.

Proof. (1) follows immediately by the definition of $F(x)$. Note that $F(x) = \tilde{F}(x) = 1$ when $x > \lambda(D) = 1$, we have $F(-\infty) = 0$, $F(+\infty) = 1$ and (2) holds. Besides, $\forall x_0 \in \overline{\mathbb{R}}_+$ and $\forall \epsilon > 0$, by the definition of $F(x_0)$, there exists $x' \in V$ such that $0 \leqslant \tilde{F}(x') - F(x_0) < \epsilon/2$. Note that $\tilde{F}(x)$ is monotone and continuous in V and that $\overline{\mathbb{R}}_+ \backslash V$ is at most countable, there exists $x'' > x'$ such that $0 \leqslant \tilde{F}(x'') - \tilde{F}(x') < \epsilon/2$. Let $\gamma = x'' - x_0$. When $0 < x - x_0 < \gamma$, since $F(x) = \tilde{F}(x)$ when $x \in V$, we have

$$\begin{aligned} 0 \leqslant F(x) - F(x_0) &\leqslant F(x'') - F(x_0) \\ &= \tilde{F}(x'') - F(x_0) \\ &= \tilde{F}(x'') - \tilde{F}(x') + \tilde{F}(x') - F(x_0) \\ &< \epsilon/2 + \epsilon/2 = \epsilon, \end{aligned}$$

which completes the proof. □

For any real function h defined in $\mathbb{R}$, let $\mathscr{C}(h) = \{x \in \mathbb{R} : h \text{ is continuous at } x\}$. Define

$$F_k(x) = 1 - \frac{E(N_{x,k})}{E(N_{0,k})}, \quad x \in \mathbb{R}. \tag{3.7}$$

The following theorem shows explicitly what V in (3.3) is and how we can approximate the true distribution function F at resolution k. Though the distribution functions are not induced by specific random variables, we will show that F_k converges to F in distribution, which means $F_k(x)$ converges to $F(x)$ for all $x \in \mathscr{C}(F)$.

Theorem 3.5. *Under the same conditions as in Theorem 3.3, we have* $V = \mathscr{C}(F) \cap \overline{\mathbb{R}}_+$. *Moreover,* $F_k(x)$ *is well-defined and*

$$F_k(x) \overset{\mathscr{D}}{\to} F(x) \quad as \quad k \to \infty.$$

Proof. Firstly note that $0 \in V \cap \mathscr{C}(F)$. $\forall x_0 \in V \backslash \{0\}$ and $\epsilon > 0$, since $\tilde{F}$ is continuous at x_0, there exists $x', x'' \in V$, such that $x' < x_0 < x''$ and $\tilde{F}(x'') - \epsilon < \tilde{F}(x_0) < \tilde{F}(x') + \epsilon$. Let $\gamma = \min(x'' - x_0, x_0 - x')$. Then for all x satisfying $|x - x_0| < \gamma$, we have

$$F(x) - \epsilon \leqslant F(x'') - \epsilon = \tilde{F}(x'') - \epsilon < \tilde{F}(x_0) < \tilde{F}(x') + \epsilon = F(x') + \epsilon \leqslant F(x) + \epsilon.$$

Hence $x_0 \in \mathscr{C}(F) \cap \overline{\mathbb{R}}_+$ and $V \subset \mathscr{C}(F) \cap \overline{\mathbb{R}}_+$.

Now suppose $x_0 \in \overline{\mathbb{R}}_+ \backslash V$. Then by definition, $P(U^c_{x_0}) > 0$. $\forall \gamma > 0$, since $\overline{\mathbb{R}}_+ \backslash V$ is at most countable, there exist $x' \in (x_0 - \gamma, x_0) \cap V$ and $x'' \in (x_0, x_0 + \gamma) \cap V$. Note that $\omega \in U^c_{x_0}$ implies there is at least one cluster with size x_0 and $N_{x'}(\omega) > N_{x''}(\omega)$, we have

$$\begin{aligned} F(x'') - F(x') = \tilde{F}(x'') - \tilde{F}(x') &= \frac{E\left(N_{x'} - N_{x''}\right)}{E(N_0)} \\ &\geqslant \frac{E\left(\left(N_{x'} - N_{x''}\right) 1_{\{N_{x'} - N_{x''} > 0\}}\right)}{E(N_0)} \\ &\geqslant \frac{E\left(1_{\{N_{x'} - N_{x''} > 0\}}\right)}{E(N_0)} = \frac{P(N_{x'} > N_{x''})}{E(N_0)} \\ &\geqslant \frac{P(U^c_{x_0})}{E(N_0)} > 0. \end{aligned}$$

The last term above is a constant only depending on x_0, which indicates that $\lim_{x \to x_0} F(x) \neq F(x_0)$ and $x_0 \in \overline{\mathbb{R}}_+ \backslash \mathscr{C}(F)$. Therefore, $\overline{\mathbb{R}}_+ \backslash V \subset \overline{\mathbb{R}}_+ \backslash \mathscr{C}(F)$ and $\mathscr{C}(F) \cap \overline{\mathbb{R}}_+ \subset V$. Hence $V = \mathscr{C}(F) \cap \overline{\mathbb{R}}_+$.

Now since $P(N_{0,k} > 0) \geqslant P(y((1,1)) > 0) > 0$, by Equation (2.3) we have $0 < E(N_{0,k}) < \infty$ and thus $F_k(x)$ is well defined in $\mathbb{R}$. To show $F_k(x) \overset{\mathscr{D}}{\to} F(x)$, it suffices to show that $\forall x \in V$, $F_k(x) \to F(x)$. Since $N_{0,k}$'s and $N_{x,k}$'s are bounded, we have, by Theorem 3.3 and the Dominated Convergence Theorem,

$$\lim_{k \to \infty} F_k(x) = 1 - \frac{\lim_{k \to \infty} E(N_{x,k})}{\lim_{k \to \infty} E(N_{0,k})} = 1 - \frac{E(N_x)}{E(N_0)} = F(x), \quad x \in V,$$

which completes the proof. □

We now address the problem of estimating the distribution function $F_k(x)$ for some specific k. In practice, data are often obtained as images with some fixed resolution. For instance (see Chen, Mohanty and Rodriguez-Iturbe (2017)), a 800 m by 800 m square field D is divided into 64 pixels, and each pixel is 100 m by 100 m. Then a remote sensor detects the soil moisture of the center point of each pixel, and produces a image of soil moisture with resolution $k = 3$. Now let y denote the process of soil moisture and $c = 0.2$ be the threshold. After each measurement, we obtain a realization of y at 64 locations and a series of sizes of clusters $x_1, x_2, ..., x_{N_0}$. Suppose n different fields $D_1, D_2, ..., D_n$ with same size and similar soil properties are chosen, and they are far away from each other so that the soil moisture of each field is considered to be independent of each other. Then we have $y_1, y_2, ..., y_n$ that are independent and identically distributed random processes defined in $D_1, D_2, ..., D_n$, respectively. After measuring the soil moisture of field D_i, $i = 1, 2, ..., n$, a series of sizes of clusters $x_{i,1}, x_{i,2}, ..., x_{i,N_{0,k,i}}$ is obtained, where $N_{x,k,i}$ is the corresponding number of clusters with size greater than x in D_i. Then in practice, a commonly used EDF of clusters $\hat{F}_{k,n}(x)$ is defined as the sample EDF of the whole data set

$$x_{1,1}, ..., x_{1,N_{0,k,1}}, x_{2,1}, ..., x_{2,N_{0,k,2}}, ..., x_{n,1}, ..., x_{n,N_{0,k,n}}.$$

It is obvious that it is equivalent to define $\hat{F}_{k,n}(x)$ as

$$\hat{F}_{k,n}(x) = 1 - \frac{\sum_{i=1}^{n} N_{x,k,i}}{\sum_{i=1}^{n} N_{0,k,i}}. \tag{3.8}$$

We now show an asymptotic property of this EDF in the following theorem.

Theorem 3.6. *Suppose that $y_1, y_2, ..., y_n$ are independent and identically distributed random processes that are defined in D and satisfy the condition in Lemma 3.1. Define*

$$T_n = \sup_{x \in \mathbb{R}} \left| \hat{F}_{k,n}(x) - F_k(x) \right|.$$

Then we have

$$T_n \overset{a.s.}{\to} 0, \tag{3.9}$$

i.e., $\hat{F}_{k,n}$ converges to F_k almost surely uniformly.

Proof. Let $x_j = j/2^k$, $j = 0, 1, 2, ..., 2^k$, and

$$I_0 = (-\infty, x_1), \quad I_j = [x_j, x_{j+1}), \quad j = 1, 2, ..., 2^k - 1, \quad I_{2^k} = [1, \infty).$$

Then by the definition of $N_{x,k}$ (note that $\lambda_k(\cdot)$ only takes finite values), we have

$$N_{x,k} = N_{x_j,k}, \quad \forall x \in I_j$$

and

$$F_k(x) = F_k(x_j), \quad \hat{F}_{k,n}(x) = \hat{F}_{k,n}(x_j), \quad \forall x \in I_j.$$

Therefore,

$$T_n = \sup_{x\in\mathbb{R}} \left|\hat{F}_{k,n}(x) - F_k(x)\right| = \max_{0\leqslant j\leqslant 2^k} \left|\hat{F}_{k,n}(x_j) - F_k(x_j)\right| \leqslant \sum_{j=0}^{2^k} \left|\hat{F}_{k,n}(x_j) - F_k(x_j)\right|.$$

Note that by the strong law of large numbers (SLLN) we have

$$\left|\hat{F}_{k,n}(x_j) - F_k(x_j)\right| = \left|\frac{\frac{1}{n}\sum_{i=1}^n N_{x_j,k,i}}{\frac{1}{n}\sum_{i=1}^n N_{0,k,i}} - \frac{E(N_{x_j,k})}{E(N_{0,k})}\right| \overset{a.s.}{\to} 0, \quad j = 0, 1, 2, ..., 2^k.$$

Since there are only finite many j's, we conclude that

$$T_n \leqslant \sum_{j=0}^{2^k} \left|\hat{F}_{k,n}(x_j) - F_k(x_j)\right| \overset{a.s.}{\to} 0,$$

as desired. □

3.3. Applications to Gaussian random fields

All results of the previous section are based on the condition (3.1) in Lemma 3.1. Note that the area of interest $A = \{y > 0\} = \{y \geqslant 0\}\backslash\{y = 0\}$ is the difference between an excursion set and a level set, and many results of properties of level sets and excursion sets have already been obtained; see, for example, Flores and Leon (2010); Worsley (1995, 1997). The following theorem is based on Rice's Formula (see Chapter 11 of Adler and Taylor (2007) and Ulrich (1984)), which makes the condition (3.1) easy to check when y is a Gaussian random field. Let ∇y denote the almost surely gradient of y: $(\partial y/\partial s_1, \partial y/\partial s_2)$, and $\nabla^2 y$ denote the almost surely Hessian matrix of y with entries $\partial^2 y/\partial s_i \partial s_j$. The joint distribution of $(y, \nabla y, \nabla^2 y)$ is defined as the joint distribution of y, ∇y and the $2(2+1)/2 = 3$ dimensional vector $\text{vech}(\nabla^2 y)$. First of all, we introduce the following lemma, which is necessary for the proof of the next theorem.

Lemma 3.2. *Suppose that f is a deterministic function defined in D and $f \in \mathscr{C}^1(D)$. Let*

$$B_0 = \partial\left\{s \in D : f(s) > 0\right\}, \quad R_0 = \left\{s \in D : f(s) = \nabla f(s) = 0\right\}.$$

Then $\lambda(B_0) = 0$ if $\lambda(R_0) = 0$.

Proof. See Appendix A.5. □

Now suppose that y is a centered Gaussian random field (GRF) defined in D. Furthermore, assume that y is twice continuously differentiable almost surely, i.e., $y \in \mathscr{C}^2(D)$ a.s., and the joint distributions of $(y, \nabla y, \nabla^2 y)$ are non-degenerate. Let $C(s,t)$ denote the covariance function of y and $C_{ij}(s,t)$ denote the covariance function of $\partial^2 y/\partial s_i \partial s_j$, namely for $s, t \in D$,

$$C(s,t) = E\left(y(s)y(t)\right), \quad C_{ij}(s,t) = E\left(\frac{\partial^2 y}{\partial s_i \partial s_j}(s)\frac{\partial^2 y}{\partial t_i \partial t_j}(t)\right). \tag{3.10}$$

Then we have the main theorem of this subsection:

Theorem 3.7. *Suppose that, for some finite $K > 0$, $\alpha > 0$ and small enough $|t-s|$, C_{ij}'s satisfy*

$$\max_{i,j=1,2} |C_{ij}(t,t) + C_{ij}(s,s) - 2C_{ij}(s,t)| \leqslant K \left|\ln|t-s|\right|^{-(1+\alpha)} \tag{3.11}$$

and

$$|C(t,t) + C(s,s) - 2C(s,t)| \leqslant K \left|\ln|t-s|\right|^{-(1+\alpha)}. \tag{3.12}$$

Then condition (3.1) in Lemma 3.1 holds.

Proof. By Lemma 3.2, condition (3.1) is satisfied if with probability one,

$$R_0(\omega) = \{s \in D : y(s) = \nabla y(s) = 0\}$$

has zero Lebesgue measure. In Theorem 11.2.1, Corollary 11.2.2 and Lemma 11.2.12 of Adler and Taylor (2007), let $T = D$, $B = (0, \infty)$, $f = \nabla y$, $g = y$. Then Lemma 11.2.12 indicates that when conditions (3.11) and (3.12) are satisfied, $R_0 = \emptyset$ with probability one. This implies $\lambda(R_0) = 0$ a.s., as desired. □

Remark 2. If y is stationary, we can let $C(t) = C(t_1, t_2)$ be the covariance function of y. By the property of GRF, C is fourth differentiable and (3.11) becomes (see Section 5.5 of Adler and Taylor (2007))

$$\max_{i,j=1,2} \left| \frac{\partial^4 C}{\partial^2 t_i \partial^2 t_j}(0) - \frac{\partial^4 C}{\partial^2 t_i \partial^2 t_j}(t) \right| \leqslant K \left|\ln|t|\right|^{-(1+\alpha)} \tag{3.13}$$

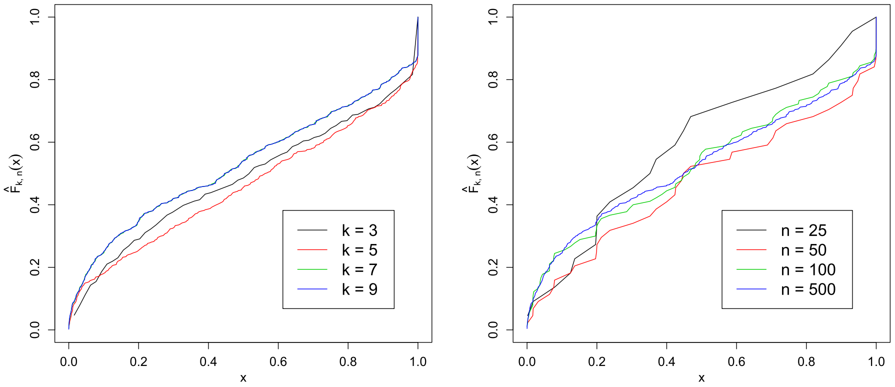
when t is small enough. (3.12) is not needed anymore since the differentiability of C implies that the left hand side of (3.12) is $O(|t-s|)$. Condition (3.13) is satisfied, for example, when y is isotropic and C is the Matern covariance function with $\nu > 2$.

4. Simulation study

Firstly, we simulated y as a Gaussian process in D with mean 0 and isotropic Gaussian covariance function $K_1(r) = e^{-r^2}$. We used $\delta = 0.001$ and y was simulated $n = 500$ times with $k = 3$, $k = 5$, $k = 7$ and $k = 9$ respectively. Then we calculated the EDFs and plotted them against x as in Figure 2a. After that, instead of fixing the sample size, we fixed $k = 7$ and used $n = 25$, $n = 50$, $n = 100$ and $n = 500$ respectively. The EDFs with different sample sizes are shown in Figure 2b.

Figure 2a shows the convergence rate of the distribution function. When $n = 500$ is fixed, $\hat{F}_{7,n}(x)$ and $\hat{F}_{9,n}(x)$ are almost identical. This indicates that the convergence of the distribution function in resolution is quite fast. Regarding the sample sizes, Figure 2b shows that when $k = 7$ is fixed, $\hat{F}_{7,n}(x)$ is close to $\hat{F}_7(x)$ when $n \geqslant 100$.

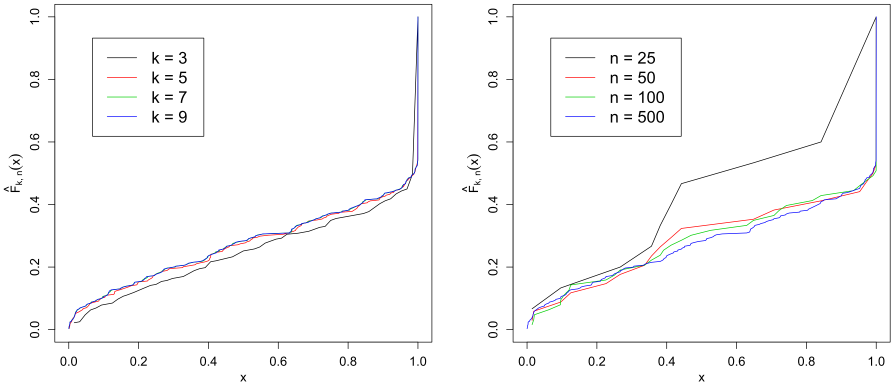
We also considered the distribution function with covariance function $K_2(r) = \sin(r)/r$, which is only valid in $\mathbb{R}^d$, $d \leqslant 3$. We used the same value of δ and



(a) EDFs with different resolutions

(b) EDFs with different sample sizes

Fig 2. *The EDFs of the areas of clusters with covariance function* $K_1(r) = e^{-r^2}$. *(a): The EDFs of* $n = 500$ *samples with different resolutions and* $\delta = 0.001$. *(b): The EDFs of resolution* $k = 7$ *with different sample sizes and* $\delta = 0.001$.



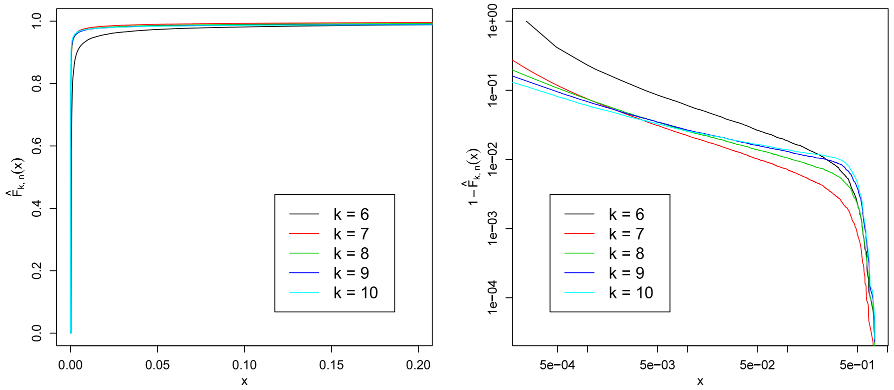
(a) EDFs with different resolutions

(b) EDFs with different sample sizes

Fig 3. *The EDFs of the areas of clusters with covariance function* $K_2(r) = \sin(r)/r$. *(a): The EDFs of* $n = 500$ *samples with different resolutions and* $\delta = 0.001$. *(b): The EDFs of resolution* $k = 7$ *with different sample sizes and* $\delta = 0.001$.

plotted the EDF curves for different resolutions and sample sizes, as described above. The results are shown in Figure 3. Though the shape of the EDF curves of K_2 are different from the curves obtained using K_1, both of them essentially converged at $k = 7$.

Finally we changed the covariance function to the exponential covariance function $K_3(r) = e^{-10r}$ and used $\delta = 0.01$. Note that K_3 is not differentiable at $r = 0$ and that a centered Gaussian process with covariance function $K(r) = e^{-10r}$ does not satisfy the conditions of Theorem 3.7 ($K_3(r) = e^{-10r}$ is not differentiable at $r = 0$ and y is not differentiable in D). We again simulated y



(a) EDFs in original scale

(b) complimentary EDFs in log-log scale

FIG 4. *The EDFs of $n = 500$ samples for different resolutions and $\delta = 0.01$ with covariance function $K_3(r) = e^{-10r}$. (a): The EDFs versus x in original scale. (b): The complimentary EDFs versus x in log-log scale.*

500 times with k from 6 to 10. The corresponding EDFs $\hat{F}_{k,n}(x)$ against x are shown in Figure 4a. We also plot the complimentary EDFs $1 - \hat{F}(x)$ against x in log-log scale in Figure 4b.

Figures 4a and 4b suggest that the EDF still converges as k gets large, though $K_3(r) = e^{-10r}$ does not satisfy the condition of Theorem 3.7. However, it converges much slower compared to the EDF with covariance $K_1(r) = e^{-r^2}$ and the EDF with $K_2(r) = \sin(r)/r$.

5. Data analysis

In this section we perform an analysis of the tree clusters data introduced in Staver et al. (2019). The tree clusters data were collected across $n = 10$ landscapes in April 2012 in Kruger National Park, South Africa, with each pixel $= 56$ cm on a side. Let $y(s)$ denote the height of the tree at location s and $y(s) = 0$ imply that there is no tree presenting at s. According to Staver et al. (2019), the area of interest is the set of locations where there are trees with height > 3.5 meters presenting, namely

$$A = \{s \in D : y(s) > 3.5\},$$

and the tree clusters are the connected components in A. The tree clusters in the data set are identified by Moore Neighborhood, which are equivalent to the tree clusters identified by Von Neumann Neighborhood with δ satisfying $\sqrt{2}l < \delta < 2l$, where $l = 56$ cm is the side length of each pixel.

The data set we analyzed is Dataset S1 in the Supporting Information of Staver et al. (2019), which contains the areas and the perimeters of $M =$

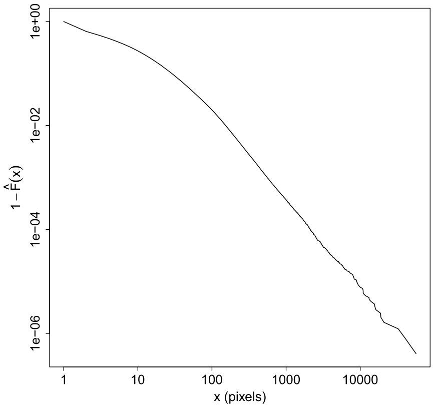


Fig 5. *The complimentary EDF of the area of the tree clusters in the log-log scale. The horizontal axis, x, is in unit of number of pixels (56 cm $\times$ 56 cm each).*

2,450,127 identified tree clusters. This data set is available on the PNAS website and more detailed descriptions and analyses of this data set can be found in Staver et al. (2019).

The complimentary EDF $1-\hat{F}(x)$ versus x is plotted in the log-log scale in Figure 5. We observe in Figure 5 that the size of tree clusters, x, may have a log-log linear relationship with its complimentary distribution function $1-F(x)$. Let x_i, $i = 1, 2, ..., M$, be the samples of sizes of tree clusters that we obtained, and $y_i = 1 - \hat{F}(x_i)$, $i = 1, 2, ..., M$, be the value of the complimentary EDF at x_i. A simple linear regression of $\log(y_i)$ versus $\log(x_i)$ resulted in a slope of -1.26 and $R^2 = 0.98$. This implies that for the tree clusters,

$$1 - F(x) \propto x^{-\beta},$$

and the size of tree clusters has a power-law distribution. Interestingly, a recently developed soil moisture space-time model has shown that the soil moisture clusters also have a similar power-law distributional property, indicating that the distributional properties of tree clusters result from the space-time probabilistic structure of soil moisture fields (Rodriguez-Iturbe et al. (2019)).

This type of power-law clustering distribution is of great interest in various fields. For instance, the power-law cluster size distribution for the nonwetting phase in sandstones reveals the existence of ganglia of all sizes presenting a large surface area for dissolution and reaction in waterflooded oil reservoirs or CO2 storage sites (Iglauer et al. (2010)). In addition, the power-law distribution of forest fires indicates the relationship between fire probability and population density, which can be used in forest-fire danger rating method and system (Song et al. (2006)). Moreover, the change in the power-law distribution of vegetation patterns should be regarded as early warning signals of ecological transitions (Kefi et al. (2014)).

6. Concluding remarks

In this paper, we considered the problem of a particular and practically useful image clustering focusing on the distributional properties of clusters of spatial random fields. It is different from the traditional statistical clustering models. A formal definition of a well-defined distribution function of the clusters $F(x)$ is given in Theorem 3.4. The definition of distribution function at specific resolution $F_k(x)$ and the EDF $\hat{F}_{k,n}(x)$ are also defined, respectively. The asymptotic properties of these two functions under general conditions are shown in Theorems 3.5 and 3.6. This provides an efficient way to estimate $F(x)$ in applications. However, the regularity condition (3.1) in Lemma 3.1 is not easy to verify. Theorem 3.7 shows that under the Gaussian assumption, instead of verifying condition (3.1), one can verify the smoothness of the covariance function of the Gaussian random field to ensure the asymptotic results hold.

The simulation studies demonstrated the convergence of $\hat{F}_{k,n}(x)$ with different covariance functions and different values of k and n. The results imply that the convergence rate depends highly on the smoothness of the covariance function $K(\cdot)$. When the isotropic exponential covariance function $K(r) = e^{-10r}$ is used, $\hat{F}_{k,n}(x)$ appears to converge, though $K(r) = e^{-10r}$ does not satisfy the regularity conditions of Theorem 3.7 ($K(r) = e^{-10r}$ is not differentiable at $r = 0$ and y is not differentiable in D).

Section 5 presented an analysis of a data set of tree clusters that became publicly available recently. Based on this data set, we obtained the empirical distribution function of the size of tree clusters. The result indicates that the tree clusters have a power-law distributional property, which is widely observed in many studies (Scanlon et al. (2007); Staver et al. (2019); Rodriguez-Iturbe et al. (2019)). A recently developed space-time model of soil moisture has indicated that the soil moisture clusters also have similar distributional properties. We conjecture that this power-law distributional property may result from some specific covariance functions, such as the isotropic exponential covariance function $K(r) = e^{-cr}$.

As a future research problem, it would be interesting to study the relationship between the model of the random field, y, and F, the distribution of the size of clusters. Assuming that y is a Gaussian Markov random field, F mainly depends on C, the covariance function of y. Studying the relationship between C and F is of great importance in many related fields since one can get the information of the random field through the image data using the relationship. In addition, in Section 3.2, we assume that D_1, D_2, ..., D_n are far from each other so that y_1, y_2, ..., y_n are independent. However, in applications, D_1, D_2, ..., D_n might be close to each other and y_1, y_2, ..., y_n would be correlated. In this case, how to obtain the distributional properties of the size of clusters using correlated data still remains a nontrivial question. Furthermore, in this paper we have obtained the main results when the domain D is bounded. When the area of D tends to infinity, stronger regularity conditions might be needed, and the asymptotic limiting distribution of cluster size could be very different from the case when D is bounded.

Appendix A: Proofs of Lemmas and Theorems

This appendix includes some proofs of the results in Sections 2 and 3.

A.1. Proof of Theorem 2.1

Proof. For any $s \in C$, there exists m such that $s \in A_m$. By Definitions 2.2 and 2.3, $\forall \tilde{s} \in A_m$, $\tilde{s} \sim s$. Thus we have $\tilde{s} \in C$ by Definition 2.4. Hence $A_m \subset C$ and $s \in A_m \subset \bigcup_{i:A_i \subset C} A_i$. Since s is arbitrary, we conclude that $C \subset \bigcup_{i:A_i \subset C} A_i$ and thus $C = \bigcup_{i:A_i \subset C} A_i$. □

A.2. Proof of Theorem 3.1

Proof. The proof consists of two parts.

(Necessity) Assume that $s' \overset{G}{\sim} s''$. Then there exists $K > 0$ such that $s' \overset{G_k}{\sim} s''$ when $k > K$. Fix $k > \max\{K, -\log_2 \delta\}$. Then since $2^{-k} < \delta$, Definition 2.6 implies that there exist $s_1, s_2, ..., s_n \in A \cap G_k$ such that $\|s_{i-1} - s_i\| < \delta$, $i = 1, 2, ..., n+1$, where $s_0 = s'$ and $s_{n+1} = s''$. Thus the condition of Definition 2.3 is satisfied for all $i = 1, 2, ..., n+1$, which means $s' \sim s''$.

(Sufficiency) Assume that $s' \sim s''$. Then there exist $0 \leqslant n < \infty$ and $s_1, s_2, ..., s_n \in A$, such that for each $i = 1, 2, ..., n+1$, $\|s_{i-1} - s_i\| < \delta$ ($s' = s_0$, $s'' = s_{n+1}$). Note that $s_1, s_2, ..., s_n$ might not belong to G. Let

$$\epsilon = \frac{1}{2}\left(\delta - \max_{i=1,2,\dots,n+1} \|s_{i-1} - s_i\|\right) > 0.$$

Since G is dense in D and A is open, for each $i = 1, ..., n$, there exists $\tilde{s}_i$ such that $\tilde{s}_i \in G \cap A \cap B(s_i, \epsilon)$. Let $\tilde{s}_0 = s'$, $\tilde{s}_{n+1} = s''$. Then we have

$$\begin{aligned}\|\tilde{s}_{i-1} - \tilde{s}_i\| &\leqslant \|\tilde{s}_{i-1} - s_{i-1}\| + \|s_{i-1} - s_i\| + \|s_i - \tilde{s}_i\| \\ &< \epsilon + \max_{i=1,2,\dots,n+1} \|s_{i-1} - s_i\| + \epsilon = \delta.\end{aligned}$$

Besides, since for all $i = 0, 1, ..., n+1$, $\tilde{s}_i \in A \cap G$, there exists $K_i > 0$, such that $\tilde{s}_i \in A \cap G_{K_i}$ for all $k > K_i$. Let $K = \max\{K_0, K_1, ..., K_{n+1}\}$, we have, when $k > K$, $\tilde{s}_i \in A \cap G_k$ for all $i = 0, 1, ..., n+1$. This satisfies the condition (2) in Definition 2.6, which implies $s' \overset{G_k}{\sim} s''$ for all $k > K$. Hence $s' \overset{G}{\sim} s''$, as desired. □

A.3. Proof of Theorem 3.2

Proof. Let $\sigma(X)$ denote the σ-algebra generated by a random variable X. Fix $x \in \mathbb{R}$ and $k \in \mathbb{N}_+$. It suffices to show that $N_{x,k}$ is a composition of two measurable functions. Let E_k be the set of $2^k \times 2^k$ matrices whose entries are

0 or 1, and the collection of all subsets of E_k is defined as the σ-algebra in E_k. Define $f_k : \Omega \mapsto E_k$ such that

$$f_k(\omega)_{(m,n)} = 1_{\{y(m/2^k,n/2^k)>0\}}, \quad m = 1,2,...,2^k, \quad n = 1,2,...,2^k,$$

i.e., the $(m,n)^{th}$ entry of $f_k(\omega)$ is equal to 1 if $y\left(m/2^k, n/2^k\right) > 0$ and is equal to 0 otherwise. Then f_k is measurable since $\forall e \in E_k$ we can express $f_k^{-1}(e)$ as

$$f_k^{-1}(e) = \bigcap_{n=1}^{2^k} \bigcap_{m=1}^{2^k} F_{m,n},$$

where $F_{m,n} = \{y\left(m/2^k, n/2^k\right) > 0\}$ or $F_{m,n} = \{y\left(m/2^k, n/2^k\right) \leqslant 0\}$, depending on $e(m,n)$, the $(m,n)^{th}$ entry of e. Now for $e \in E_k$, let

$$Q_e = \left\{(m,n) : e(m,n) = 1, m = 1,2,...,2^k, n = 1,2,...,2^k\right\}.$$

We define an equivalence class $\overset{e}{\sim}$ in Q_e (similar to Definition 2.6) as follows: if $q' = (m',n')$ and $q'' = (m'',n'')$, then $q' \overset{e}{\sim} q''$ if there exist $0 \leqslant l < \infty$ and $q_1, q_2, ..., q_l \in Q_e$, such that for each $i = 1,2,...,l+1$, at least one of the following two conditions is satisfied ($q_0 = s', q_{l+1} = q'', q_i = (m_i,n_i), i = 0,1,...,l+1$):

(1) $|m_{i-1} - m_i| + |n_{i-1} - n_i| = 1$;

(2) $|m_{i-1} - m_i|^2 + |n_{i-1} - n_i|^2 < (\delta \cdot 2^k)^2$.

Now we define $g_{x,k}$ as the number of "clusters" with size larger than x on a given matrix in E_k:

$$g_{x,k} : E_k \mapsto \mathbb{N}, \quad g_{x,k}(e) = card\left(\left\{U \in Q_e/\overset{e}{\sim}: card(U) \cdot 4^{-k} > x\right\}\right).$$

Then automatically $g_{x,k}$ is measurable since any subset of E_k belongs to the σ-algebra defined in E_k. Finally, it is clear that $N_{x,k} = g_{x,k} \circ f_k$, which completes the proof. □

A.4. Proof of Lemma 3.1

Proof. We first prove the case when $x = 0$ (note that $U_0 = \Omega$). Fix $\omega \in \Omega$. Note that from Theorem 2.1,

$$C_i = \bigcup_{j:A_j \subset C_i} A_j, \quad i = 1,2,...,N_0,$$

where C_i's are open, nonempty and mutually exclusive. Since G is dense in D, there exists $K_1 \in \mathbb{N}_+$ such that for any $k \geqslant K_1$, $C_i \cap G_k \neq \emptyset$, $i = 1,2,...,N_0$. Besides, from the first part of the proof of Theorem 3.1, there exists $K_2 \in \mathbb{N}_+$ such that for any $k \geqslant K_2$, $s' \overset{G_k}{\sim} s''$ implies $s' \sim s''$. Let $K = \max\{K_1, K_2\}$. Then we can choose s_i such that $s_i \in C_i \cap G_K$, $i = 1,2,...,N_0$. For $k > K$, if $i \neq j$ then s_i and s_j must belong to different clusters in G_k since $s_i \overset{G_k}{\sim} s_j$ would imply $s_i \sim s_j$. Therefore, there should be at least N_0 different clusters in G_k and thus $N_{0,k} \geqslant N_0$ when $k > K$. Therefore, we have

$$\liminf_{k\to\infty} N_{0,k} \geqslant N_0.$$

Now it suffices to show

$$\limsup_{k\to\infty} N_{0,k} \leqslant N_0.$$

Suppose $\limsup_{k\to\infty} N_{0,k} > N_0$. Then there exists an increasing sequence of integer $\{k_n\}_{n=1}^{\infty}$ such that $k_1 > K$ and $N_{0,k_n} > N_0$ for all n. Let $C_{1,k_n}, C_{2,k_n}, \ldots, C_{N_{0,k_n},k_n}$ be the clusters of G_{k_n} and $s_j \in C_{j,k_n} \subset C_j$, $j = 1, 2, \ldots, N_0$, as chosen above. Let $\tilde{s}_n$ be a point in C_{N_0+1,k_n}, $n = 1, 2, \ldots$. Since D is compact in $\mathbb{R}^2$, there exists a subsequence of $\tilde{s}_n$ that converges to a limit $\tilde{s} \in D$. WLOG, we can assume $\tilde{s}_n \to \tilde{s}$ as $n \to \infty$. Then there exists $N \in \mathbb{N}_+$ such that $\|\tilde{s}_n - \tilde{s}_N\| < \delta$ for all $n \geqslant N$. Again, WLOG, assume $\tilde{s}_N \in C_1$. Then $\tilde{s}_N \sim s_1$, which implies $\tilde{s}_N \overset{G}{\sim} s_1$ by Theorem 3.1. Hence there exists $n_0 > N$ such that $\tilde{s}_N \overset{G_{k_{n_0}}}{\sim} s_1$. Since $\|\tilde{s}_{n_0} - \tilde{s}_N\| < \delta$, we have

$$s_1 \overset{G_{k_{n_0}}}{\sim} \tilde{s}_N \overset{G_{k_{n_0}}}{\sim} \tilde{s}_{n_0}.$$

However, this is a contradiction since $s_1 \in C_{1,k_{n_0}}$, $\tilde{s}_{n_0} \in C_{N_0+1,k_{n_0}}$, s_1 and $\tilde{s}_{n_0}$ would not belong to a same cluster of $G_{k_{n_0}}$. Hence we conclude that $N_{0,k} \to N_0$ for all $\omega \in \Omega = U_0$, which also implies that N_0 is a random variable.

Now we prove the case when $x > 0$. From the proof above, there exists $K \in \mathbb{N}_+$ such that $N_{0,k} = N_0$ and $s' \overset{G_k}{\sim} s''$ implies $s' \sim s''$ for all $k \geqslant K$. Let $C_{1,k}, C_{2,k}, \ldots, C_{N_0,k}$ be the clusters of G_k. Then for any i and j, either $C_{j,k} \subset C_i$ or $C_{j,k} \cap C_i = \emptyset$. Assume that $C_{j,k} \subset C_j$, $j = 1, 2, \ldots, N_0$. Then we have

$$\lambda_k(C_{j,k}) = \left(\frac{1}{2^k}\right)^2 \sum_{s\in G_k} 1_{C_{j,k}}(s) = \left(\frac{1}{2^k}\right)^2 \sum_{s\in G_k} 1_{C_j}(s), \quad j = 1, 2, \ldots, N_0,$$

which is a Riemann sum. To show that it converges to the Lebesgue integral

$$\int_D 1_{C_j}(s)ds = \lambda(C_j), \quad j = 1, 2, \ldots, N_0,$$

it suffices to show that the set of discontinuous points of 1_{C_j} has zero Lebesgue measure, i.e., $\lambda(\partial C_j) = 0$ for $j = 1, 2, \ldots, N_0$. For $j = 1$, suppose $s \in \partial C_1$. Then for any $0 < \epsilon < \delta/2$, there exist s' and s'' such that $s' \in B(s, \epsilon) \cap C_1^c$ and $s'' \in B(s, \epsilon) \cap C_1$, $y(s'') > 0$. Since $\|s' - s''\| < \delta$, if $y(s') > 0$, we would have $s' \sim s''$ and $s' \in C_1$, which contradicts the fact that $s' \in C_1^c$. Thus $y(s') \leqslant 0$. Therefore, for any $0 < \epsilon < \delta/2$, there exist s' and s'' in $B(s, \epsilon)$ such that $y(s') \leqslant 0$, $y(s'') > 0$. By the definition of $\partial A(\omega)$, we have $s \in \partial A(\omega)$. Therefore, we have shown that

$$\partial C_1 \subset \partial A(\omega) = \partial \{s \in D : y(s, \omega) > 0\},$$

and thus

$$\lambda(\partial C_1) = \lambda(\partial A(\omega)) = 0.$$

Similarly, this equation holds when $j = 2, 3, ..., N_0$. Therefore, $\forall \omega \in \Omega$,

$$\lambda_k(C_{j,k}) \to \lambda(C_j), \quad j = 1, 2, ..., N_0(\omega). \tag{A.1}$$

Now to show N_x is a random variable, it suffices to show that $\forall l \in \mathbb{N}_+$,

$$\{\omega \in \Omega : N_x(\omega) = l\} = \bigcup_{p=1}^{\infty} \bigcap_{q=p}^{\infty} \bigcup_{r=1}^{\infty} \bigcap_{k=r}^{\infty} \{\omega \in \Omega : N_{x+1/q,k}(\omega) = l\}.$$

WLOG, assume $\lambda(C_1) \geqslant \lambda(C_2) \geqslant \cdots \geqslant \lambda(C_{N_0})$. Suppose $N_x(\omega) = l$. Let $p = 1$ if $l = 0$. Otherwise, choose $p > 0$ large enough such that $x+1/p < \lambda(C_{N_x})$. Then we have $\lambda(C_i) > x + 1/p$, $i = 1, 2, ..., N_x$ and $\lambda(C_i) < x + 1/p$, $i = N_x + 1, N_x + 2, ..., N_0$. By (A.1), there exists $r > 0$, such that for all $k \geqslant r$, $\lambda_k(C_{j,k}) > x+1/p$, $j = 1, 2, ..., N_x$ and $\lambda_k(C_{j,k}) < x = 1/p$, $j = N_x + 1, N_x + 2, ..., N_0$. Therefore, for all $q \geqslant p$ and $k \geqslant r$, $N_{x+1/q,k}(\omega) = l$, which implies that ω is in the right hand side (RHS). Thus the left hand side (LHS) $\subset$ RHS.

Now suppose that ω is in RHS. Then there exists $q_0 > 0$ such that

$$\omega \in \bigcup_{r=1}^{\infty} \bigcap_{k=r}^{\infty} \{N_{x+1/q_0,k}(\omega) = l\},$$

which indicates that $N_x(\omega) \geqslant l$ by (A.1). Suppose $N_x(\omega) > l$. Then for all $q > 1/(\lambda(C_{N_x}) - x)$, we have $x + 1/q < \lambda(C_{N_x})$. Similarly, as the proof above, when k is large enough, we would have $N_{x+1/q,k}(\omega) > l$, which is a contradiction since ω is in RHS. Hence $N_x(\omega) = l$ and ω is in LHS. This implies LHS = RHS.

Finally, assume $\omega \in U_x$. Again, WLOG, we can assume $\lambda(C_i) > x$, $i = 1, 2, ..., N_x$ and $\lambda(C_i) < x$, $i = N_x + 1, N_x + 2, ..., N_0$. Then we have, for sufficiently large k, $\lambda_k(C_{j,k}) > x$, $j = 1, 2, ..., N_x$ and $\lambda_k(C_{j,k}) < x$, $j = N_x + 1, N_x + 2, ..., N_0$. Thus $N_{x,k} = N_x$ for sufficiently large k, which implies $\forall \omega \in U_x$, $N_{x,k} \to N_x$ as $k \to \infty$, as desired. □

A.5. Proof of Lemma 3.2

Proof. Firstly, note that for any $x \in B_0$ and $\epsilon > 0$, there exist x' and x'' such that $x', x'' \in B(x, \epsilon)$ and $f(x') > 0$, $f(x'') \leqslant 0$. Since f is continuous, we have $f(x) = 0$ and $B_0 \subset \{f(x) = 0\}$. Since

$$\begin{aligned}
\lambda(B_0) &= \lambda(B_0 \cap \{\nabla f(x) = 0\}) + \lambda(B_0 \cap \{\nabla f(x) \neq 0\}) \\
&\leqslant \lambda(\{f(x) = 0\} \cap \{\nabla f(x) = 0\}) + \lambda(\{f(x) = 0\} \cap \{\nabla f(x) \neq 0\}) \\
&= \lambda(R_0) + \lambda(\{f(x) = 0\} \cap \{\nabla f(x) \neq 0\}),
\end{aligned}$$

it suffices to show $\lambda(\{f(x) = 0\} \cap \{\nabla f(x) \neq 0\}) = 0$. Since $\lambda(\partial D) = 0$, it suffices to show

$$\lambda(\{f(x) = 0\} \cap \{\nabla f(x) \neq 0\} \cap D^{\mathrm{o}}) = 0,$$

where $D^{\rm o} = (0,1)^2$ is the interior of D. Let $S_0 = \{f(x) = 0\} \cap \{\nabla f(x) \neq 0\} \cap D^{\rm o}$. Now suppose $x = (x_1, x_2) \in S_0$. WLOG, we can assume that $\partial f(x)/\partial x_2 \neq 0$. By the Implicit Function Theorem, there exist $\gamma_x^1, \gamma_x^2 > 0$ and a continuously differentiable function $g : I_x \mapsto J_x$ such that

$$\begin{aligned}\{(y_1, g(y_1)) : y_1 \in I_x\} &= \{(y_1, y_2) \in I_x \times J_x : f(y_1, y_2) = 0\} \\ &= (I_x \times J_x) \cap \{f(y) = 0\},\end{aligned}$$

where $I_x = (x_1 - \gamma_x^1, x_1 + \gamma_x^1)$ and $J_x = (x_2 - \gamma_x^2, x_2 + \gamma_x^2)$ are the neighborhoods of x_1 and x_2 respectively. Now we show that the above set has zero Lebesgue measure. For any $0 < \gamma < \gamma_x^1$ and $\epsilon > 0$, since g is uniformly continuous in $[x_1 - \gamma, x_1 + \gamma]$, there exists $\xi > 0$ such that when $z', z'' \in [x_1 - \gamma, x_1 + \gamma]$ and $|z_1 - z_2| < \xi$, $|g(z_1) - g(z_2)| < \epsilon$. Choose $K \in \mathbb{N}$ such that $2\gamma/K < \xi$ and let

$$z_k = x_1 - \gamma + \frac{2\gamma k}{K}, \quad k = 0, 1, 2, ..., K.$$

Then we have

$$\{(z, g(z)) : z \in [x_1 - \gamma, x_1 + \gamma]\} \subset \\ \bigcup_{k=1}^{K} [z_{k-1}, z_k] \times \left[g\left(\frac{z_{k-1} + z_k}{2}\right) - \epsilon, g\left(\frac{z_{k-1} + z_k}{2}\right) + \epsilon \right],$$

where the Lebesgue measure of RHS is

$$\sum_{k=1}^{K} 2(z_k - z_{k-1})\epsilon = 4\gamma\epsilon.$$

Since ϵ is arbitrary, we conclude that

$$\lambda\left(\{(z, g(z)) : z \in [x_1 - \gamma, x_1 + \gamma]\}\right) = 0.$$

By letting $\gamma \uparrow \gamma_x^1$, we obtain

$$\lambda\left(\{(y_1, g(y_1)) : y_1 \in I_x\}\right) = 0.$$

Finally, note that $I_x \times J_x$ exists for each x in S_0, $\{I_x \times J_x\}_{x \in S_0}$ is an open cover of S_0. Since $\mathbb{R}^2$ is Lindelof, there exists a countable subcover $\{I_n \times J_n\}_{n=1}^{\infty}$ of S_0 (Gemignani (1990), Chapter 7). Therefore, we have

$$\begin{aligned}\lambda(S_0) &= \lambda\left(\bigcup_{n=1}^{\infty} (S_0 \cap (I_n \times J_n))\right) \\ &\leqslant \sum_{n=1}^{\infty} \lambda\left(S_0 \cap (I_n \times J_n)\right) \\ &\leqslant \sum_{n=1}^{\infty} \lambda\left(\{f(y) = 0\} \cap (I_n \times J_n)\right) \\ &= \sum_{n=1}^{\infty} \lambda\left(\{(y_1, g(y_1)) : y_1 \in I_n\}\right) = 0,\end{aligned}$$

which completes the proof. □

Acknowledgments

The authors thank the Editor and two reviewers for their valuable comments and suggestions which substantially improved an earlier version of this paper. The authors also thank Dr. Ignacio Rodriguez-Iturbe and Dr. Ann Carla Staver for their advice in the process of the data analysis. This research was supported in part by the Simons Foundation Mathematics and Physical Sciences - Collaboration Grants for Mathematicians Program Award #499650.

References

Adler, R. J. and Taylor, J. E. (2007). *Random Fields and Geometry*. Springer-Verlag, New York. MR2319516

Asnera, G. P. and Warner, A. S. (2003). Canopy shadow in IKONOS satellite observations of tropical forests and savannas. *Remote. Sens. Environ.* **87** 521–533.

Azzalini, A. and Torelli, N. (2007). Clustering via nonparametric density estimation. *Stat. Comput.* **17** 71–80. MR2370969

Cattelan, M. and Varin, C. (2018). Marginal logistic regression for spatially clusteredbinary data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 939–959. MR3832258

Chen, Z., Mohanty, B. P. and Rodriguez-Iturbe, I. (2017). Space-time modeling of soil moisture. *Adv. Water. Resour.* **109** 343–354.

Flores, E. and Leon, J. R. (2010). Level sets of random fields and applications: Specular points and wave crests. *Int. J. Stoch. Anal.* **2010** 1–22. MR2678922

Gemignani, M. C. (1990). *Elementary Topology*, 2 ed. Dover Publications, New York. MR1088253

Iglauer, S., Favretto, S., Spinelli, G., Schena, G. and Blunt, M. J. (2010). X-ray tomography measurements of power-law cluster size distributions for the nonwetting phase in sandstones. *Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys.* **82** 056315.

Kefi, S., Guttal, V., Brock, W. A., Carpenter, S. R., Ellison, A. M. and Livina, V. N. e. a. (2014). Early warning signals of ecological transitions: Methods for spatial patterns. *PLoS ONE* **9** e92097.

Li, B. (2006). A new approach to cluster analysis: The clustering-function-based method. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 457–476. MR2278335

Martin, H. G. and Goldenfeld, G. (2006). On the origin and robustness of power-law species–area relationships in ecology. *PNAS* **103** 10310—10315.

McNicholas, P. D. (2016). Model-based clustering. *J. Classification* **33** 331–373. MR3575621

Menardi, G. and Azzalini, A. (2014). An advancement in clustering via nonparametric density estimation. *Stat. Comput.* **24** 753–767. MR3229695

Pascual, M., Roy, M., Guichard, F. and Flierl, G. (2002). Cluster size distributions: Signatures of self-organization in spatial ecologies. *Phil. Trans. R. Soc. Lond. B* **357** 657–666.

Rodriguez-Iturbe, I., Chen, Z., Staver, C. and Levin, S. A. (2019). Tree clusters in savannas result from islands of soil moisture. *PNAS* **116** 6679–6683.

Scanlon, T. M., Caylor, K. K., Levin, S. A. and Rodriguez-Iturbe, I. (2007). Positive feedbacks promote power-law clustering of Kalahari vegetation. *Nature* **449** 209–212.

Sole, R. (2007). Scaling laws in the drier. *Nature* **449** 151–153.

Song, W., Wang, J., Satoh, K. and Fan, W. (2006). Three types of power-law distribution of forest fires in Japan. *Ecol. Model.* **196** 527–532.

Staver, A. C., Asner, G. P., Rodriguez-Iturbe, I., Levin, S. A. and Smit, I. (2019). Spatial patterning among savanna trees in high-resolution, spatially extensive data. *PNAS* **116** 10681–10685.

Steinwart, I. (2015). Fully adaptive density-based clustering. *Ann. Statist.* **43** 2132–2167. MR3396981

Ulrich, Z. (1984). A general rice formula, palm measures, and horizontal-window conditioning for random fields. *Stochastic Process. Appl.* **17** 265–283. MR0751206

Wootton, J. T. (2001). Local interactions predict large-scale pattern in empirically derived cellular automata. *Nature* **413** 841–844.

Worsley, K. J. (1995). Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *Ann. Statist.* **23** 640–669. MR1332586

Worsley, K. J. (1997). The geometry of random images. *Chance* **9** 27–40.