Image Anal Stereol 2019;38:185-199 doi:10.5566/ias.2017
Original Research Paper

SIMILARITY BETWEEN RANDOM SETS CONSISTING OF MANY COMPONENTS

Vesna Gotovac

Faculty of Science, University of Split, 21000 Split, Croatia
e-mail: vgotovac@pmfst.hr

(Received September 28, 2018; revised June 13, 2019; accepted June 14, 2019)

ABSTRACT

Random sets play an essential role in modelling several phenomena in biology, medicine and material science. However, sometimes it is hard to describe them using a specific model. Therefore it can also be difficult to classify them or to compare their realisations. This contribution proposes a similarity measure between two random sets whose realisations consist of many components based on just one realisation of each of them. The similarity measure is obtained in a non-parametric way taking into account the shapes and the positions of the components. The procedure is justified by a simulation study and consequently applied to real biomedical data of histological images of mammary tissue.

Keywords: kernel test, non-parametric statistics, random tessellation, similarity measure.

INTRODUCTION

In recent years, random sets have been developed as a valuable tool for modelling various phenomena in many fields of science such as dynamics of cells in organisms (Mrkvička and Mattfeldt, 2011; Hermann *et al.*, 2015), presence of different plants in ecosystems (Diggle, 1981; Møller and Helisová, 2010) or particles in materials (Helisová, 2014; Neumann *et al.*, 2016). In order to describe and explain these events, a wide range of random set models are introduced.

A significant class of random set processes are germ-grain models (Chiu *et al.*, 2013), where the random set is formed as a union of simple random sets (grains) whose reference points (germs) form a point process. The most simple model of this kind is the Boolean model (Chiu *et al.*, 2013), where the process of the germs forms a Poisson point process, the grains are independently distributed and their distribution does not depend on the germs. Since this model does not capture any dependencies between grains, some more sophisticated models have had to be developed. Let us mention for example Quermass-interaction processes (Kendall *et al.*, 1999) used in the simulation study in the presented paper, where the interactions among the grains are described through geometric characteristics of the whole union, which makes the model very flexible.

In order to compare or classify the random set processes based on their realisations, a natural way is to fit a particular class of parametric models to the realisations and compare the obtained parameters. Unfortunately, it can be hard to find a suitable model in some situations. For example, Mrkvička and Mattfeldt (2011) try to fit a random-disc Boolean model to data of mammary tissue with the aim to distinguish between mammary cancer and masthopatic tissue, but they notice that only a few percents of the realisations can be described in this way. Neither the Quermass-interaction process fitted to the same data in Hermann *et al.* (2015) appears to be satisfactory. Therefore we have decided to construct a similarity measure of random sets in a non-parametric way. The main idea is to obtain a value describing the degree of belief that two processes have the same feature of interest so that it can be then further used for comparison and classification purposes.

A similarity measure of random sets based on just two realisations is constructed by Gotovac *et al.* (2016) and improved by Gotovac and Helisová (2019). The authors define a similarity measure through the inner structure of the random sets. More precisely, they approximate the realisations by a union of the discs with identical radii which are tessellated to obtain two groups of convex compact cells. Since the inner structures of the random sets are characterised by the shapes of the obtained random convex compact cells, the distributions of those two groups of cells are compared. Although the procedure gives satisfactory results in some situations, there are some features that cannot be captured. For example, in special cases when having many small connected components in the input realisations, it may be sensitive to small changes in approximations because small differences in the approximations can significantly change the shape of the original realisation. Moreover, it does not take into account the positions of the components.

This paper proposes a similarity measure between two realisations that focuses on the resemblance of the distribution of set components and their positions. It works with all the components completely visible within the observation window with no need for their further approximations. This method is flexible in the sense that one can decide based on which set components to compare the realisations. For example, if we want to construct the similarity of random sets based on the similarity of the distribution of their connected components, the components of interest are obtained by decomposing the set into its connected components. When the nature of the problem imposes assessing the similarity of some more specific set component, we can further decompose the observed set into those smaller set components of interest. The idea is as follows:

1. After isolating the components of interest, the underlying random set process is considered as the disjoint union of those components (usually rugged shaped) centred in their centroids. Although it seems tempting to use the point process of centroids for investigating the position of the shapes, this approach has some disadvantages. First, there is a loss of information due to the edge effects (i.e. the components not completely visible in the observation window could have their centroid within the observation window, but we do not have that information), and second, we would have to investigate the shapes and their positions independently. Therefore we apply the following steps.
2. We define a neighbourhood of each connected component as the set of all points that are nearer to that connected component than to any other connected component in the Hausdorff metric, so we construct a random tessellation of the observation window. Its purpose is to reveal the position of the connected components.
3. We sample pairs of connected components and their neighbourhoods from each realisation and compare them by a permutation version of the goodness of fit test based on $\mathfrak{N}$-distances (Klebanov, 2006) for which we introduce a particular kernel.
4. The similarity between two realisations is calculated as the *p*-value of the aforementioned test applied on the groups of ordered pairs of centred components with their neighbourhoods.

The procedure is graphically represented in Fig. 3.

The paper is organised as follows. In the *Materials and Methods* section, we recall the basic definitions concerning random sets, $\mathfrak{N}$-distances and permutation test, introduce neighbourhood tessellation of the set and explain the methodology of construction of the similarity measure in details. Then in the *Results* section, we present the results of the simulation study and apply the methodology to real data concerning histological images of mammary tissue. Finally, we give a summary of the paper with a brief discussion and comments in the *Discussion* section.

MATERIALS AND METHODS

RANDOM SETS

In this section, we present some basics concerning the theory of random sets. Definitions 1-4 are taken from (Chiu *et al.*, 2013) while Definition 5 of Quermass-interaction process in this form can be found in (Møller and Helisová, 2008).

Let $\mathscr{F}$ be the family of closed sets and $\mathscr{C}$ the family of compact set of the topological space $\mathbb{R}^d$ with the standard topology $\mathscr{G}$.

Definition 1. *Let* (Ω,Σ,P) *be a probability space. A mapping* $\mathbf{X}:\Omega\to\mathscr{F}$ *is a* random closed set *if, for every compact set* $K\in\mathscr{C}$

$$\{\omega\in\Omega:\mathbf{X}(\omega)\cap K\neq\emptyset\}\in\Sigma.$$

Definition 2. The distribution $P_{\mathbf{X}}$ of a random closed set $\mathbf{X}$ *is given by the relation* $P_{\mathbf{X}}(F) = P(\{\omega\in\Omega:\mathbf{X}(\omega)\in F\})$ *for* $F\in\mathscr{B}(\mathscr{F})$, *where* $\mathscr{B}(\mathscr{F})$ *is the Borel sigma algebra on* $\mathscr{F}$ *generated by topology* $\mathscr{G}$.

Definition 3. *A random closed set* $\mathbf{X}$ *is* stationary *if its distribution is invariant under translation, i.e. for all* $v\in\mathbb{R}^d$, *the distribution of* $\mathbf{X}+v=\{u+v,v\in\mathbf{X}\}$ *is the same as that of* $\mathbf{X}$.

For $A,B\subset\mathbb{R}^d$ let us denote by $A\oplus B:=\{x+y: x\in A, y\in B\}$ and by $|A|$ the n-dimensional Lebesgue measure of the set A.

Definition 4. *Let* $Y=\{y_1,y_2,\ldots\}$ *be a stationary Poisson point process in* $\mathbb{R}^d$ *and* $\{\mathbf{B}_1,\mathbf{B}_2,\ldots\}$ *be a sequence of independent identically distributed random compact sets in* $\mathbb{R}^d$ *that are independent of* Y. *If* $\mathbb{E}|\mathbf{B}_1\oplus K|<\infty$ *for all compact sets* K, *then the random set*

$$\mathbf{B}=\cup_{n=1}^{\infty}(y_n+\mathbf{B}_n)$$

is called Boolean model.

Definition 5. *Consider a planar random disc Boolean model, i.e. the Boolean model with* $\mathbf{B}_1$ *being a disc in* $\mathbb{R}^2$ *with random radius.* The Quermass-interaction process *is a random set whose probability measure is absolutely continuous with respect to the probability* *measure of the given Boolean model . The density of its probability measure with respect to the probability measure of the given Boolean model is of the form*

$$f_\theta(\mathbf{b}) = \frac{1}{c_\theta}\exp\{\theta_1 A(U_\mathbf{b}) + \theta_2 L(U_\mathbf{b}) + \theta_3\chi(U_\mathbf{b})\}$$

for each finite disc configuration $\mathbf{b} = \{\mathbf{b}_1 \dots, \mathbf{b}_n\}$*, where* $A = A(U_\mathbf{b})$ *is the area,* $L = L(U_\mathbf{b})$ *is the perimeter,* $\chi = \chi(U_\mathbf{b})$ *is Euler-Poincaré characteristic (i.e. the number of connected components minus the number of holes) of the union* $U_\mathbf{b} = \cup_{i=1}^n \mathbf{b}_i$*,* $\theta = (\theta_1, \theta_2, \theta_3)$ *is 3-dimensional vector of parameters and* c_θ *is the normalising constant.*

NEIGHBOURHOODS OF COMPONENTS

We address the problem of describing the position of the components of the set by introducing a random tessellation of an observation window. This tessellation splits up an observation window based on the components $C_i, i \in I$ in a way that each cell N_i contains the original component C_i and all points in the observation window that are closer to C_i than any other $C_j, j \in I \setminus \{i\}$ in Hausdorff metric. Let us recall that *Hausdorff metric* on $\mathscr{F}$ is defined by

$$d_H(A,B) = \max\left\{\sup_{x\in A}\inf_{y\in B}\|x-y\|, \sup_{y\in B}\inf_{x\in A}\|x-y\|\right\},$$
$$A,B \in \mathscr{F},$$

where we denote $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$.

Definition 6. *Consider a finite union of disjoint random sets* $\{C_1 \dots, C_n\}$ *within the observation window* W*. Every set* C_j *generates the neighbourhood*

$$N_j = \{y \in W\ : d_H(\{y\}, C_j) \le d_H(\{y\}, C_k) \textit{ for all } j \neq k\}.$$

The system $\mathscr{T}$ *of all sets* N_j*,* $j = 1, \dots, n$ *is called the* neighbourhood tessellation on a union of sets.

An example of neighbourhood tessellations of a union of components is represented in Fig. 1.

NEGATIVE DEFINITE KERNELS AND $\mathfrak{N}$-DISTANCES

In this section, we briefly introduce the reader to the theory of $\mathfrak{N}$-distances, the distances between probability distributions that are built using negative definite kernels. We also provide some valuable examples that will be used in the construction of the similarity measure. All definitions, Propositions 8, 9 and 14, Theorem 10 and Example 11 can be found in (Klebanov, 2006) and Example 12 is taken from (Gotovac *et al.*, 2017).

Definition 7. *Let* $\mathscr{X}$ *be a non-empty set. A map* $L : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$ *is called a* negative definite kernel *if for any* $n \in \mathbb{N}$*, arbitrary* $c_1, ..., c_n \in \mathbb{C}$ *such that* $\sum_{j=1}^n c_j = 0$ *and arbitrary* $x_1, ..., x_n \in \mathscr{X}$*, it holds*

$$\sum_{j=1}^n\sum_{k=1}^n L(x_j,x_k)c_j\overline{c_k} \le 0. \qquad (1)$$

Proposition 8. *If* L *is a real function satisfying* $L(x,y) = L(y,x)$ *for all* $x,y \in \mathscr{X}$*, then* L *is a negative definite kernel if and only if* (1) *holds for arbitrary real numbers* $c_1, \dots, c_n$ *under condition* $\sum_j^n c_j = 0$.

Suppose that L is a real continuous function, and denote by $\mathscr{P}_L$ the set of all probability measures P on $\mathscr{X}$ for which the integral

$$\int_\mathscr{X}\int_\mathscr{X} L(x,y)dP(x)dP(y)$$

exists.

Proposition 9. *Let* L *be a real continuous function on* $\mathscr{X} \times \mathscr{X}$ *satisfying*

$$L(x,y) = L(y,x),\ x,y \in \mathscr{X}.$$

The inequality

$$\begin{aligned}&2\int_\mathscr{X}\int_\mathscr{X} L(x,y)\mathrm{d}P_1(x)dP_2(y)\\ &-\int_\mathscr{X}\int_\mathscr{X} L(x,y)dP_1(x)dP_1(y)\\ &-\int_\mathscr{X}\int_\mathscr{X} L(x,y)dP_2(x)dP_2(y) \ge 0\end{aligned}$$

holds for all $P_1, P_2 \in \mathscr{P}_L$ *if and only if* L *is a negative definite kernel.*

Definition 10. *A map* $L : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ *is called a* strongly negative definite kernel *if* L *is a negative definite kernel and for an arbitrary probability measure* $Q \in \mathscr{P}$ *and an arbitrary* $c : \mathscr{X} \to \mathbb{R}$ *such that* $\int_\mathscr{X} c(x)dQ(x) = 0$ *and* $\int_\mathscr{X}\int_\mathscr{X} L(x,y)c(x)c(y)dQ(x)dQ(y)$ *exists, relation*

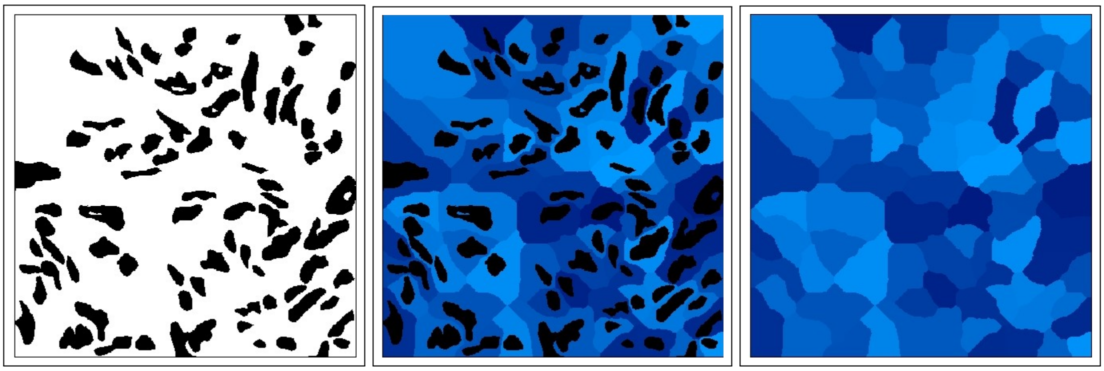
$$\int_\mathscr{X}\int_\mathscr{X} L(x,y)c(x)c(y)dQ(x)dQ(y) = 0$$

implies $c(x) = 0$ *Q-almost everywhere.*

Theorem 11. *Let* L *be a strongly negative definite kernel on* $\mathscr{X} \times \mathscr{X}$ *satisfying* $L(x,y) = L(y,x)$ *and* $L(x,x) = 0$ *for all* $x,y \in \mathscr{X}$*. Let* $\mathscr{N} : \mathscr{P}_L \times \mathscr{P}_L \to \mathbb{R}$ *be defined by*

$$\begin{aligned}\mathscr{N}(P_1,P_2) =&2\int_\mathscr{X}\int_\mathscr{X} L(x,y)dP_1(x)dP_2(y)\\ &-\int_\mathscr{X}\int_\mathscr{X} L(x,y)dP_1(x)dP_1(y)\\ &-\int_\mathscr{X}\int_\mathscr{X} L(x,y)dP_2(x)dP_2(y).\end{aligned}$$

Then $\mathfrak{N} = \mathscr{N}^{1/2}$ *is a distance on* $\mathscr{P}_L$*.*



(a) Union of the disjoints components (b) Components and neighbourhoods (c) Neighbourhood tessellation

Fig. 1: Neighbourhood tessellations of union of components

Example 12. *If $\mathscr{X} = \mathbb{R}^d$, we have that*

$$L(x,y) = \|x-y\|^r,$$

for $0<r<2$ is a strongly negative definite kernel on $\mathbb{R}^d \times \mathbb{R}^d$.

For $C_1, C_2 \subset \mathbb{R}^d$ let us denote $C_1 \Delta C_2 := \{x \in \mathbb{R}^d : x \in C_1, x \notin C_2 \text{ or } x \notin C_1, x \in C_2\}$ the symmetric difference between sets C_1 and C_2.

Example 13. *If $\mathscr{X} = \mathscr{C}$ the family of all compact sets in $\mathbb{R}^d$ we have that*

$$L(C_1,C_2) = \mu^{r/2}(C_1 \Delta C_2),$$

where $0<r\leq 2$ and μ is an arbitrary finite measure on $\mathscr{C}$, is a negative definite kernel on $\mathscr{C}\times\mathscr{C}$ (Gotovac et al., 2017).

Let us further on take $\mu(D) = \mu_A(D)$, where $\mu_A(D)$ is the area of the set $D \in \mathscr{C}$. Suppose we have discretised versions of the sets C_1 and C_2, i.e. the sets are represented as the matrices of zeros and ones, denoted by M_{C_1} and M_{C_2} respectively. Then

$$L_A(C_1,C_2) = \mu_A^{r/2}(C_1 \Delta C_2) = ||M_{C_1} - M_{C_2}||_F^r,$$

where $||\cdot||_F$ stands for Frobenius matrix norm, is by Example 12 a strongly negative definite kernel on the set of all matrices for $0<r<2$. Fig. 2 presents the difference between two matrices representing discretised sets whose Frobenius norm approximates the square root of the area of the symmetric difference between the original sets.

Proposition 14. *If a negative definite kernel $L: \mathscr{X}\times\mathscr{X} \to \mathbb{R}$ satisfies the conditions $L(x,x)=0$, $L(x,y) = L(y,x)$ for all $x,y \in \mathscr{X}$, then there exists a real Hilbert space and a family $(a_x)_{x\in\mathscr{X}}$ of its elements such that*

$$L(x,y) = \left\|a_x - a_y\right\|^2,\ x,y \in \mathscr{X}.$$

Proposition 15. *Let $\mathscr{X}_1$ and $\mathscr{X}_2$ be arbitrary non-empty sets and $L_1 : \mathscr{X}_1 \times \mathscr{X}_1 \to \mathbb{R}$ and $L_2 : \mathscr{X}_2 \times \mathscr{X}_2 \to \mathbb{R}$ negative definite kernels on $\mathscr{X}_1$ and $\mathscr{X}_2$, respectively, satisfying conditions $L_1(x_1,x_1) = L_2(x_2,x_2) = 0$, $L_1(x_1,y_1) = L(y_1,x_1)$ and $L_2(x_2,y_2) = L_2(y_2,x_2)$ for all $x_1,y_1 \in \mathscr{X}_1$, $x_2,y_2 \in \mathscr{X}_2$. Let $\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2$ and define $L : \mathscr{X}\times\mathscr{X} \to \mathbb{R}$ by*

$$L\left((X_1,X_2),(X_1',X_2')\right) = \sqrt{L_1\left(X_1,X_1'\right) + L_2\left(X_2,X_2'\right)},$$

for $(X_1,X_2),(X_1',X_2') \in \mathscr{X}$. Then L is a negative definite kernel on $\mathscr{X}$.

Proof. Note that from Proposition 14 it follows that L_1 and L_2 are non-negative functions, so L is well defined.

For $n \in \mathbb{N}$, arbitrary $c_1,\ldots,c_n \in \mathbb{R}$ such that $\sum_{j=1}^n c_j = 0$ and arbitrary $\left(X_1^{(1)},X_2^{(1)}\right),\ldots,\left(X_1^{(n)},X_2^{(n)}\right) \in \mathscr{X}$ it holds

$$\begin{aligned}
&\sum_{j=1}^n\sum_{k=1}^n L\left(\left(X_1^{(j)},X_2^{(j)}\right),\left(X_1^{(k)},X_2^{(k)}\right)\right)c_jc_k \\
&= \sum_{j=1}^n\sum_{k=1}^n \sqrt{L_1\left(X_1^{(j)},X_1^{(k)}\right) + L_2\left(X_2^{(j)},X_2^{(k)}\right)}c_jc_k \\
&\leq \sum_{j=1}^n\sum_{k=1}^n \left(1 + L_1\left(X_1^{(j)},X_1^{(k)}\right) + L_2\left(X_2^{(j)},X_2^{(k)}\right)\right)c_jc_k \\
&\leq \sum_{j=1}^n\sum_{k=1}^n L_1\left(X_1^{(j)},X_1^{(k)}\right)c_jc_k + \\
&+ \sum_{j=1}^n\sum_{k=1}^n L_2\left(X_2^{(j)},X_2^{(k)}\right)c_jc_k \leq 0.
\end{aligned}$$

The first inequality results from the fact that for a non-negative real number x it holds $\sqrt{x} \leq 1+x$. Consequently, following inequality (2) and Proposition 8, we conclude that L is a negative definite kernel. ■

Example 16. *Suppose that $\mathscr{X} = \mathscr{C} \times \mathscr{C}$ and that we only have available discretised versions of the sets. Then we can use following kernel*

$$\mathscr{L}_A\left((C_1,N_1),(C_2,N_2)\right) = \sqrt{\mu_A\left(C_1 \Delta C_2\right) + \mu_A\left(N_1 \Delta N_2\right)}$$
$$= \sqrt{||M_{C_1} - M_{C_2}||_F^2 + ||M_{N_1} - M_{N_2}||_F^2},$$

where $(C_1,N_1),(C_2,N_2) \in \mathscr{C} \times \mathscr{C}$ and M_{C_1}, M_{N_1}, M_{C_2}, $M_{N_2} \in \{0,1\}^{n \times m}$ are the matrices consisting of zeros and ones representing the approximations of the sets C_1, N_1, C_2 and N_2, respectively.

SIMILARITY MEASURE BASED ON $\mathfrak{N}$-DISTANCES

Similarity measures are used to represent the nearness of two objects. In statistics and related fields they play an important role in cluster analysis and alignment algorithms. The value of a similarity measure increases as two objects become more similar. In this section, we introduce a similarity measure that is inspired by the permutation version of the test of equality in distribution based on $\mathfrak{N}$-distances (Klebanov, 2006).

Consider two random elements **A** and **B** taking values in $\mathscr{X}$. Let L be a negative definite kernel on $\mathscr{X} \times \mathscr{X}$.

Suppose we have two samples $X = (X_1,\ldots,X_{m_1})$ from a random element **A** with distribution P_1 and $Y = (Y_1,\ldots,Y_{m_2})$ from a random element **B** with distribution P_2, where $P_1,P_2 \in \mathscr{P}_L$. Further on, suppose that the samples X and Y are such that the sequence of random elements $X_1,\ldots,X_{m_1},Y_1,\ldots,Y_{m_2}$ is exchangeable under the assumption that $P_1 = P_2$. The aim is to define similarity between distributions P_1 and P_2 as the p-value of a statistical test with the null hypothesis $P_1 = P_2$.

First, we evaluate the empirical estimate of the square of the $\mathfrak{N}$-distance between P_1 and P_2 from Theorem 11

$$\hat{\mathscr{N}} = \frac{2}{m_1 m_2}\sum_{j=1}^{m_1}\sum_{k=1}^{m_2} L(X_j,Y_k) - \frac{1}{m_1^2}\sum_{j=1}^{m_1}\sum_{k=1}^{m_1} L(X_j,X_k)$$
$$- \frac{1}{m_2^2}\sum_{j=1}^{m_2}\sum_{k=1}^{m_2} L(Y_j,Y_k).$$

We choose the number of permutations s (recommended to be about 1000), and then s times, we permute the considered joined sample $X_1,\ldots,X_{m_1},Y_1,\ldots,Y_{m_2}$, subsequently split them back into two samples of the length m_1 and m_2, respectively, and for each calculate empirical values of $\hat{\mathscr{N}_j}$, $j = 1,\ldots,s$. Under the assumption $P_1 = P_2$, permutations do not modify the distribution of the random variable $\hat{\mathscr{N}}$. If the distributions of the two samples differ, we expect that after the permutations, the value of the $\mathfrak{N}-$distance is smaller, so we define

$$p = \frac{\#\left\{j : \hat{\mathscr{N}_j} \geq \hat{\mathscr{N}}\right\} + 1}{s+1}$$

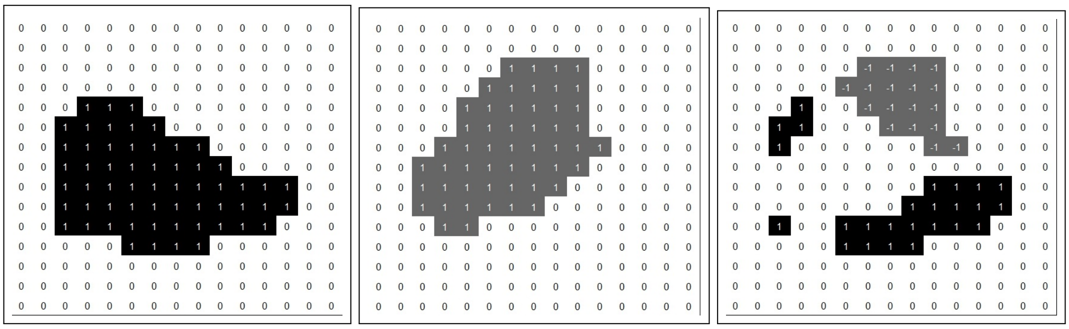
which will be the similarity between two samples (or two distributions P_1 and P_2). The smaller values of this similarity indicate lower degree of belief that these two distributions are the same. Note that this similarity p should be uniformly distributed on the segment $[0,1]$ if $P_1 = P_2$.

METHODOLOGY

Suppose we have two stationary random closed sets **X** and **Y** in $\mathbb{R}^2$, which can be expressed as $\mathbf{X} = \cup_{n\in\mathbb{N}}\left(x_n + \mathbf{C}_n^{(1)}\right)$ and $\mathbf{Y} = \cup_{n\in\mathbb{N}}\left(y_n + \mathbf{C}_n^{(2)}\right)$, where $\left(\mathbf{C}_n^{(1)}\right)_{n\in\mathbb{N}}$ and $\left(\mathbf{C}_n^{(2)}\right)_{n\in\mathbb{N}}$ are sequences of identically distributed random compact sets centred at the origin (further referred as the components), $\{x_n\}$ and $\{y_n\}$ are stationary point processes. Further on, suppose that the random sets are observed within the observation windows W_1 and W_2, respectively. We set $\left(\mathbf{N}_n^{(1)}\right)$ to be the neighbourhood tessellation of $\left\{(x_n + \mathbf{C}_n^{(1)}) \cap W_1\right\}$ and $\left(\mathbf{N}_n^{(2)}\right)$ to be the neighbourhood tessellation of $\left\{(y_n + \mathbf{C}_n^{(2)}) \cap W_2\right\}$. In order to compare **X** and **Y**, we introduce the similarity measures defined as the p-values of the statistical tests with the following null hypothesis:

- $\left(\mathbf{C}_n^{(1)}\right) \overset{d}{=} \left(\mathbf{C}_n^{(2)}\right)$, when we concentrate only on the similarity of the components shapes or
- $\left(\mathbf{C}_n^{(1)},\mathbf{N}_n^{(1)}\right) \overset{d}{=} \left(\mathbf{C}_n^{(2)},\mathbf{N}_n^{(2)}\right)$ and
- $\left(\mathbf{N}_n^{(1)}\right) \overset{d}{=} \left(\mathbf{N}_n^{(2)}\right)$, when we want to obtain the similarity based on the shapes and the positions of the components.

Let us now propose the procedure for sampling and testing the aforementioned hypothesis when we have just two realisations of random sets available. As mentioned in the *Introduction*, the procedure of comparing two realisations of random sets works in the following steps (Fig. 3):



(a) Example of a discretised set M_{C_1} (b) Example of a discretised set M_{C_2} (c) Difference $M_{C_1} - M_{C_2}$

Fig. 2: Difference between matrices representing discretised sets.

First, we determine the components of interest in each realisation. It can be done simply by considering the connected components of the realisations as usually done when we do not have enough information about the input data, or we can take into account the nature of the data and split them due to the required purpose as done for the example here in the *Application to real data* section.

Then we construct neighbourhood tessellations with respect to the obtained components as described in the *Neighbourhoods of components* section.

After the construction of the neighbourhood tessellations, to avoid edge effects, we exclude the connected components intersecting the boundary of the observation window, i.e. the ones that are not completely visible (coloured by dark grey in Fig. 3 while the considered components are coloured in black). Furthermore, if we moreover want to compare the positions of the components, we exclude all the neighbourhoods touching the boundary together with their connected components (coloured by grey in Fig. 3 while the neighbourhoods to be considered are coloured in blue). In this way, we identify each realisation with a realisation of a marked point process, where the collection of centroids corresponding to the components form the point process (red dots in Fig. 3) and the corresponding components and neighbourhoods, respectively, play the role of marks. Then assuming that the observed random sets are stationary, all the centred components together with their centred neighbourhoods have identical joined distribution. Furthermore, suppose that those two realisations are generated by the same random set, then two collections of pairs of the components with their neighbourhoods form an exchangeable sequence of random pairs.

Finally, we construct the similarity measure between two realisations of such marked point processes using the approach described in the *Similarity measure based on $\mathfrak{N}$-distances* section.

Suppose we have a sample of m_1 centred components $C_1^{(1)}, \ldots, C_{m_1}^{(1)}$ paired with its centred neighbourhoods $N_1^{(1)}, \ldots, N_{m_1}^{(1)}$ from the first realisation and m_2 centred components $C_1^{(2)}, \ldots, C_{m_2}^{(2)}$ paired with its centred neighbourhoods $N_1^{(2)}, \ldots, N_{m_2}^{(2)}$ from the second realisation.

The approach requires us to choose the appropriate negative definite kernel that captures the most important features of the shapes (i.e. this negative definite kernel should obtain smaller values when evaluated on two sets having similar features of interest). So if we want to focus only on the similarity of the component shapes (or only on the neighbourhoods shapes) we can consider a negative definite kernel in the form from Example 13, where we chose kernel L_A. In case we want to incorporate also the position of the components, we can use the kernel $\mathscr{L}_A$ from Example 16 on ordered pairs of components and their neighbourhoods.

Using these kernels we calculate the p-value of the permutation test based on $\mathfrak{N}$-distances , which presents the measure of similarity between two realisations of random sets.

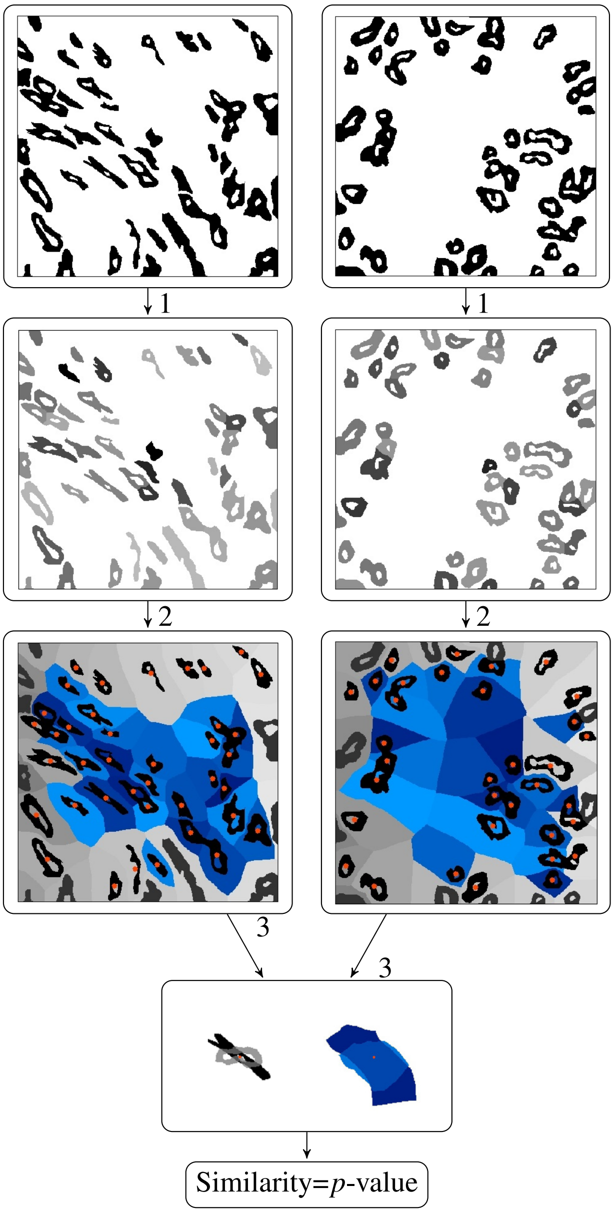


Fig. 3: Steps in calculating the similarity between two random sets. *Step 1.:* Isolate the components of interest from each realisation (different components are coloured in different shades of grey). *Step 2.:* Construct neighbourhood tessellations with respect to obtained components (coloured in different shades of blue and light grey). Those suitable for sampling are components coloured in black with their centroids coloured red and neighbourhoods coloured in different shades of blue, i.e. ones not intersecting the boundary of the observation window. *Step 3.:* Calculate the similarity of two collections of pairs of centred components and their neighbourhoods using the procedure from *Similarity measure based on $\mathfrak{N}$-distances* section.

RESULTS

SIMULATION STUDY

We applied the methodology described in the previous section to different simulated realisations of random sets to see if we can differentiate between them based on the shape of the connected components and their position. The first one is the random-disc Boolean model (see Definition 4) with centres of discs in the window 25×25, the intensity of the disc centres equal to 0.4 and the uniform distribution of radii on the interval $(0.5, 1)$ (see Fig. 4c). The second one is the random-ellipse Boolean model (see Definition 4) with centres of ellipses in the window 25×25, the intensity of the ellipse centres equal to 0.4 and uniform distribution of semi-major axes on the interval $(0.5, 1)$ and semi-minor axes on interval $(0.2, 0.7)$ (see Fig. 4b). The third one is the Quermass-interaction process (see Definition 5) with the parameters θ_1 = 0.62, θ_2 = -0.86 and θ_3 = 0.7 with respect to the random-disc Boolean model mentioned above. Since this process produces realisations with larger area and smaller perimeter compared to the reference process, it tends to create clusters (see Fig. 4d). Therefore, we will refer to it as the cluster process in the rest of the text. The fourth data set is simulated as Quermass-interaction process with parameters θ_1 = -1, θ_2 = 1 and θ_3 = 0 with respect to the same random-disc Boolean model. It prefers smaller area and larger perimeter than the reference random-disc Boolean modes, so its realisations are usually small non-overlapping components (see Fig. 4a) and therefore the process will be referenced as the repulsive process in the rest of the text. We have simulated 200 realisations for each of the mentioned processes, all realisations are transformed to matrices of 400 $\times$ 400 black and white pixels which play the role of the input data. Just note that the input random-disc Boolean and repulsive random sets models are the same as the ones used in (Gotovac *et al.*, 2016) and (Gotovac and Helisová, 2019).

To explore the sensitivity of the methodology between two different processes as well as within the classes of the same processes, the following approach is taken. First, the input matrices are divided into two groups so that we have 100 pairs of realisations for the random-disc Boolean model, random-ellipse Boolean model, cluster process and repulsive process, respectively, and the similarity is studied within these groups separately for different processes. Additionally, we considered 100 pairs of realisations of different processes, namely random-disc Boolean vs random-ellipse Boolean, random-disc Boolean vs cluster random-disc Boolean vs repulsive, random-ellipse Boolean vs cluster, random-ellipse Boolean vs repulsive, and cluster vs repulsive processes, and studied the similarity again.

Using the methodology from the *Methodology* section, for each realisation, the connected components are isolated, and their neighbourhoods are constructed (see Fig. 4e, Fig. 4f, Fig. 4g and Fig. 4h). We use connected components that are not touching the boundary when we construct a similarity based on the components shapes and their connected components together with their neighbourhoods that are also not touching the boundary of the observation window for the construction of the similarity based on the shapes of connected components and their positions. Note that the realisations of the random-disc Boolean model have on average 32 connected components and 25 neighbourhoods not touching the boundary. From the random-ellipse Boolean model realisations, an average of 73 connected components and 58 neighbourhoods were sampled. The repulsive process realisations have 91 connected components and 83 neighbourhoods on average, and the cluster model realisations have on average 33 connected components and 20 neighbourhoods not touching the boundary.

In order to compare the methodology presented in this paper with the existing methodology for comparing the inner structure of the sets from Gotovac *et al.* (2016) and Gotovac and Helisová (2019), the realisations were approximated by the union of the convex compact sets following the recommendations from Gotovac *et al.* (2016). In more details, for comparison of repulsive vs random-ellipse Boolean, random-ellipse-Boolean vs random-disc Boolean and random-ellipse Boolean vs cluster, the realisations were approximated by the union of the discs with the radii of 4 pixels. When comparing repulsive vs random-disc Boolean and repulsive vs cluster, we used approximation by discs with radii 5 and for comparison of random-disc Boolean and cluster the radii of the discs in approximation was 7.

The Voronoi tessellation of the obtained union of the discs was constructed, which resulted in two groups of convex compact cells for each comparison. For each realisation, 100 non-neighbouring convex compact cells were sampled. For comparing two samples of convex compact cells, we used the permutation test based on the $\mathfrak{N}$-distances. The kernel that we used it this permutation test was introduced in Gotovac and Helisová (2019) and it was selected as the one that showed the best empirical power when testing the equality in the distribution of convex compact cells.

Fig. 5 shows the histograms of p-values when comparing the same processes using the method introduced in this paper. We observe that those p-values are approximately uniformly distributed. Fig. 6 and Fig. 7 present the histograms of the p-values when comparing different processes when using the methodology from present paper together with the histograms of the p-values when using the methodology from Gotovac *et al.* (2016) and Gotovac and Helisová (2019) as described above. The percentages of the p-values that are less or equal to 0.05 obtained by the method presented in this paper are higher than the ones obtained by the other method, with the exception of comparing repulsive vs cluster model using only connected components.

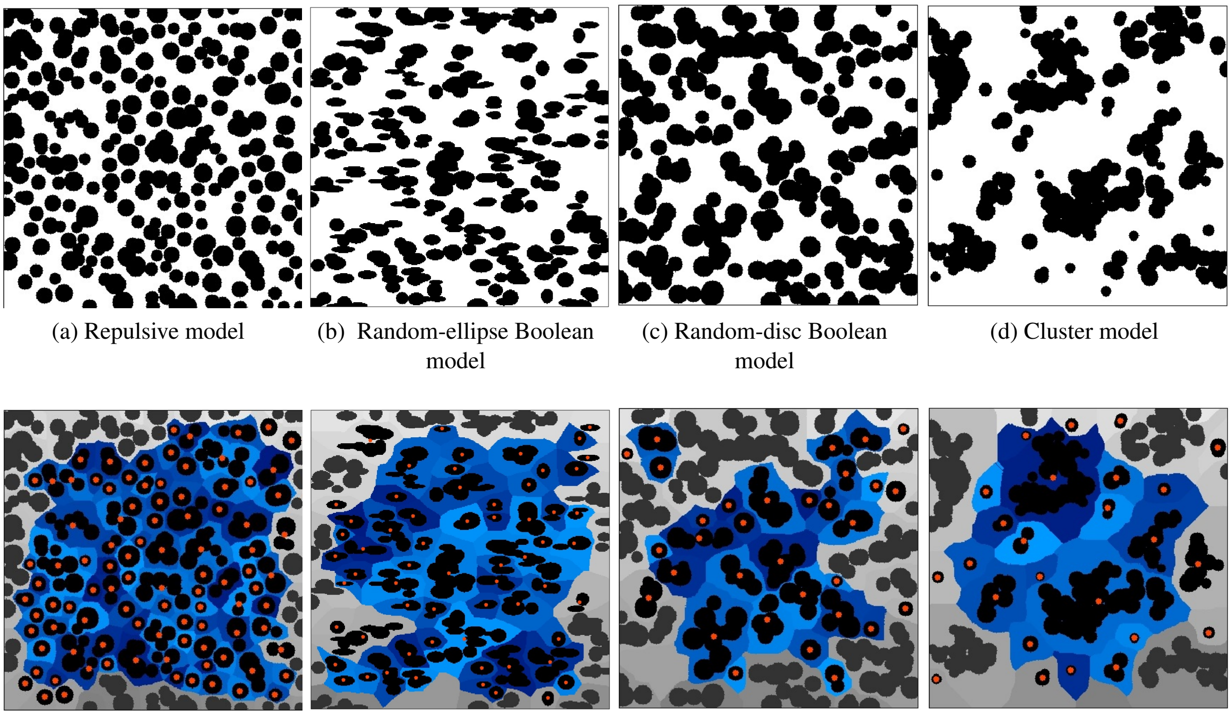
The weaker results of the powers obtained when comparing cluster model with other models using the methodology presented in this paper are due to the smaller sample sizes of connected components in cluster samples and only a few connected components which are significantly larger and can be seen as the outliers when testing equality in distribution. Also, for some realisations, the larger components characterising the cluster model are touching the boundary, so they were excluded from the sampling. In these cases one can consider using the methodology from Gotovac *et al.* (2016) and Gotovac and Helisová (2019) instead of the methodology presented in this paper.

APPLICATION TO REAL DATA

We apply the procedure also to data of mammary tissue with the aim to distinguish between mammary cancer and masthopatic tissue. Note that the breast contains a branching system of ducts spanning down from the nipple to glands. The tissue between the ducts and glands is made of fat and fibrous tissue of differing proportions. The morphology of this tissue may indicate various malignant or benign changes.

We consider 8 samples of mastopathic breast tissue (called Masto in the sequel, see Fig. 8) and 8 samples of mammary cancer tissue (called Mamca in the sequel, see Fig. 9) that were kindly provided by the authors of (Mrkvička and Mattfeldt, 2011). Each sample consists of 10 sub-samples formed by matrices of 512 × 512 black and white pixels. The samples present histological images of cross sections of the ducts branches, where the black areas represent the surrounding tissue between ducts and glands.

The aim is to compare the samples based on the shapes of the ducts and their surrounding tissue which are the components of interest. So, we proceed as follows. First, we isolate the components of interest. We decompose the samples so that the holes, i.e. the white pixels surrounded by the black pixels, can be interpreted as the ducts. So, heuristically, we assign to



(a) Repulsive model

(b) Random-ellipse Boolean model

(c) Random-disc Boolean model

(d) Cluster model

(e) Repulsive model with neighbourhood tessellation

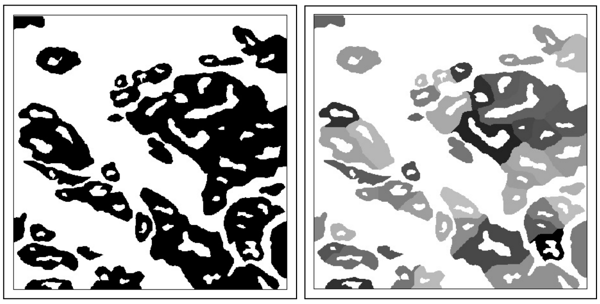
(f) Random-ellipse Boolean model with neighbourhood tessellation

(g) Random-disc Boolean model with neighbourhood tessellation

(h) Cluster model with neighbourhood tessellation

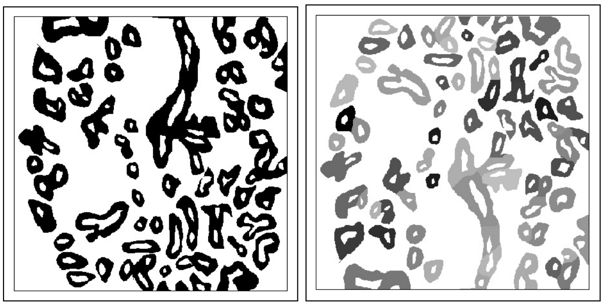
Fig. 4: Examples of realisations of simulated repulsive model (a) and its decomposition to connected components and neighbourhood tessellation (e), random-ellipse Boolean model (b) and its decomposition to connected components and neighbourhood tessellation (f), random-disc Boolean model (c) and its decomposition to connected components and neighbourhood tessellation (g) and cluster model (d) and its decomposition to connected components and neighbourhood tessellation (h) .

each hole (duct) all black pixels (surrounding tissue) that are closer to that hole than to any other hole within the corresponding connected component. In more detail, first we decompose sample to its connected components and than each connected component is further decomposed based on the holes as described above. An example of such decomposition is presented in Fig. 10.



(a) Original sample of mammary cancer

(b) Decomposed sample of mammary cancer



(c) Original sample of mastopathic tissue

(d) Decomposed sample of mastopathic tissue

Fig. 10: Example of the decomposition of original set based on holes; Original samples are in sub-figures (a) and (c) and obtained component are in sub-figures (b) and (d), respectively (coloured in different shades of grey)

Then we determine the centroids of the obtained components and construct the neighbourhood tessellation. From each sample, 50 components with their neighbourhoods which are not touching the boundary of the sub-samples are then collected

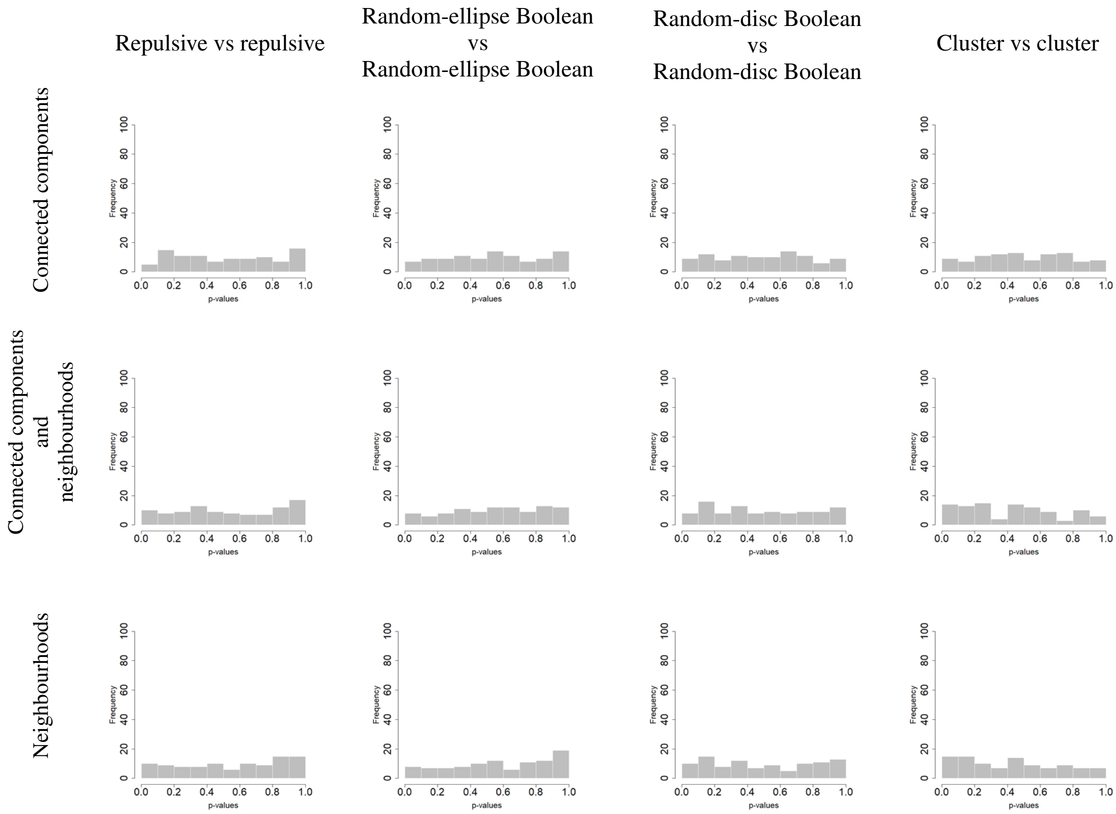


Fig. 5: Histograms of p-values for testing pairs of same simulated processes when using only components (first row), components and neighbourhood tessellations (second row) and only neighbourhood tessellations (third row). Here, we compare repulsive vs repulsive (first column), random-ellipse Boolean vs random-ellipse Boolean (second column), random-disc Boolean vs random-disc Boolean (third column), cluster vs cluster (fourth column); 100 realisations of each process were used.

randomly. For all pairs of samples, corresponding p-values of the tests are evaluated (see *Similarity measure based on* $\mathfrak{N}$*-distances* section) using only components and also using components together with the neighbourhoods. Note that in both cases, the number of permutations is $s = 999$, so that the smallest possible p-value is 0.001.

The obtained p-values from both permutation tests are presented in Table 1 for Masto vs Masto (left sub-table) and Mamca vs Mamca (right sub-table) and Table 2 for Masto vs Mamca. From the left sub-table in Table 1, we can observe that the higher similarities are in the blocks on the diagonal. It suggests that we can cluster those images based on the shapes of components into 3 clusters: the first one consisting of the samples Masto1 and Masto2, the second one of the samples Masto3, Masto4, Masto5 and Masto6, and the third one including the samples Masto7 and Masto8. The right sub-table in Table 1 shows us that Mamca samples can be divided into two groups: the first one with the samples Mamca1 to Mamca5 and the second one containing the samples Mamca6 to Mamca8. Table 2 shows that comparing two groups of different samples, we mostly obtain small p-values, which means that the procedure can distinguish between those two types of tissue.

	Masto1	Masto2	Masto3	Masto4	Masto5	Masto6	Masto7	Masto8
Mamca1	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.005	0.001 0.001
Mamca2	0.002 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.002 0.006	0.013 0.009
Mamca3	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.006	0.001 0.114
Mamca4	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.013	0.001 0.009
Mamca5	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.011 **0.055**	0.017 0.023
Mamca6	0.007 0.001	0.002 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 **0.097**
Mamca7	0.010 0.001	0.015 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.024	0.011 **0.051**
Mamca8	0.021 0.001	0.015 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.001 0.001	0.006 0.040	0.001 0.015

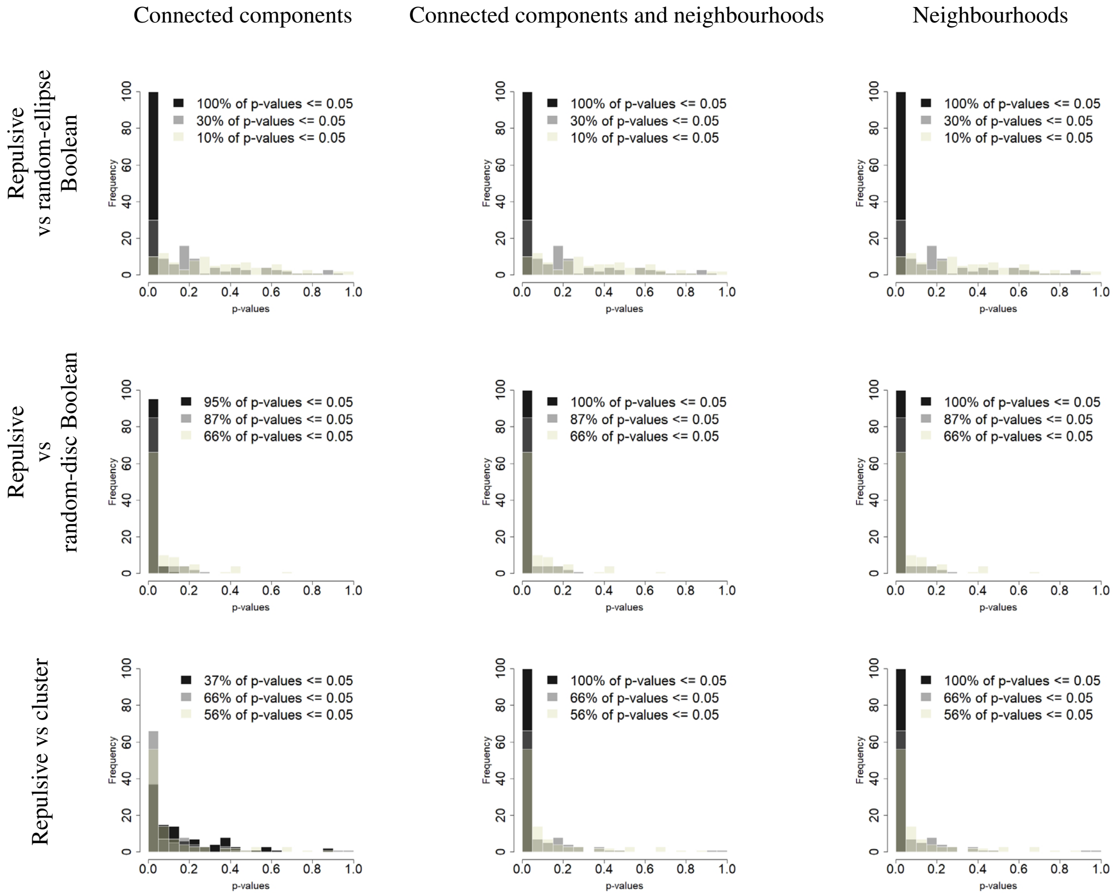


Fig. 6: Histograms of *p*-values for testing pairs of different simulated processes by the method from (Gotovac *et al.*, 2016) (histograms coloured light grey), by the method from (Gotovac and Helisová, 2019) (histograms coloured grey) and by the method from the present paper (histograms coloured in black) when using only components (first column), components and neighbourhood tessellations (second column) and only neighbourhood tessellations (third column). Here, we compare repulsive vs random-ellipse Boolean (first row), repulsive vs random-disc Boolean (second row), repulsive vs cluster (third row). 100 realisations of each process were used. Percentages of *p*-values less or equal to 0.05 are displayed in legend.

Table 2: Table of the *p*-values (similarity measures) between samples of mammary cancer tissue and mastophatic tissue; Each row consists of two sub-rows: the first sub-row presents results of the test when only components are taken into the consideration (black) and the second sub-row consists of results of the test when both shape of components and their position are considered (dark grey); *p*-values larger than 0.05 are in bold.

DISCUSSION

The presented research shows that a similarity measure can be a useful tool to differentiate between random set processes, especially in cases where there is no suitable parametric model to describe the data.

When comparing two realisations, due to the high complexity of some random sets in the applications and in order to avoid overfitting, it is recommendable to concentrate on specific features of interest. Here, we proposed a new similarity measure between random sets based on just two realisations that concentrates on the shapes of set components as well as their position.

The presented methodology can be summed in the following steps: isolating components of interest, constructing the neighbourhood tessellation (this step can be omitted if positions of components are not important), choosing features of components that are of interest and consequently finding a kernel which will be sensitive to those features, and finally applying the permutation test. Therefore, the method is very flexible since a researcher could easily modify the steps in order to adjust the method for the specific needs.

This methodology was justified by a simulation

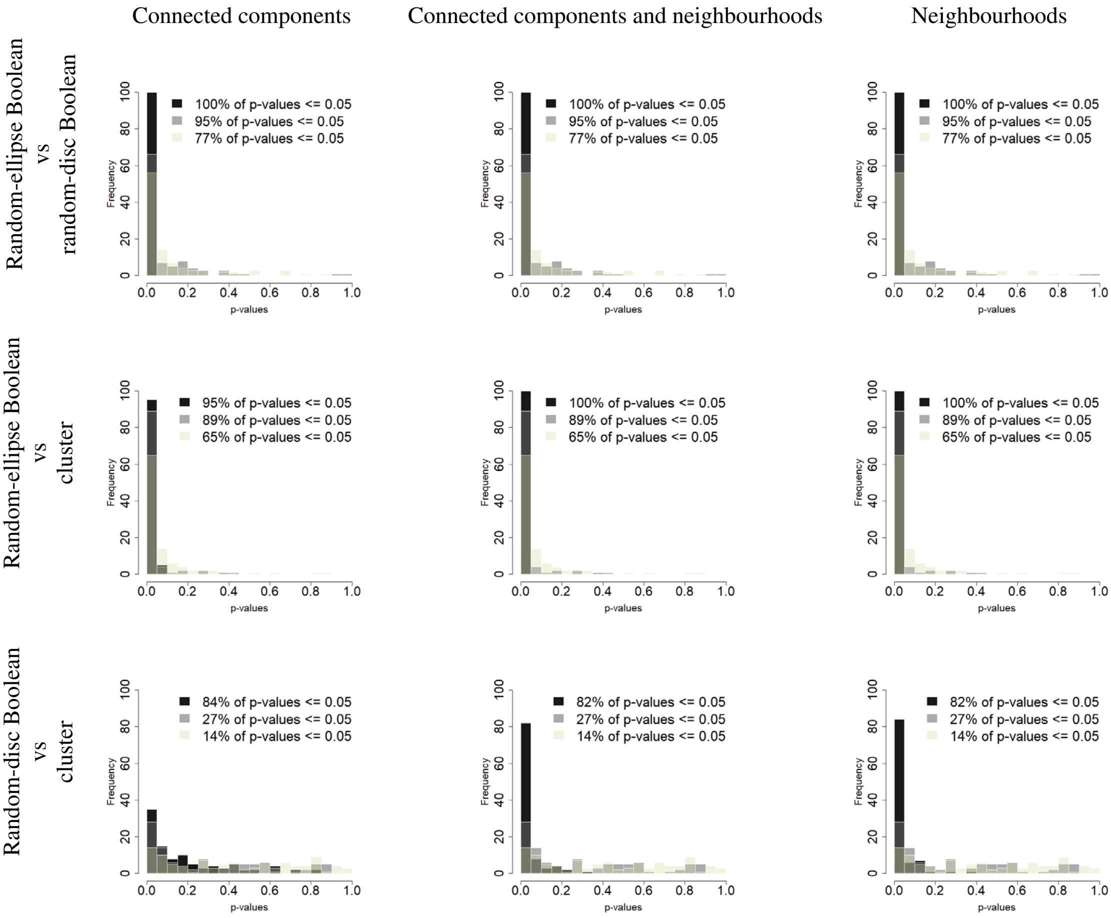
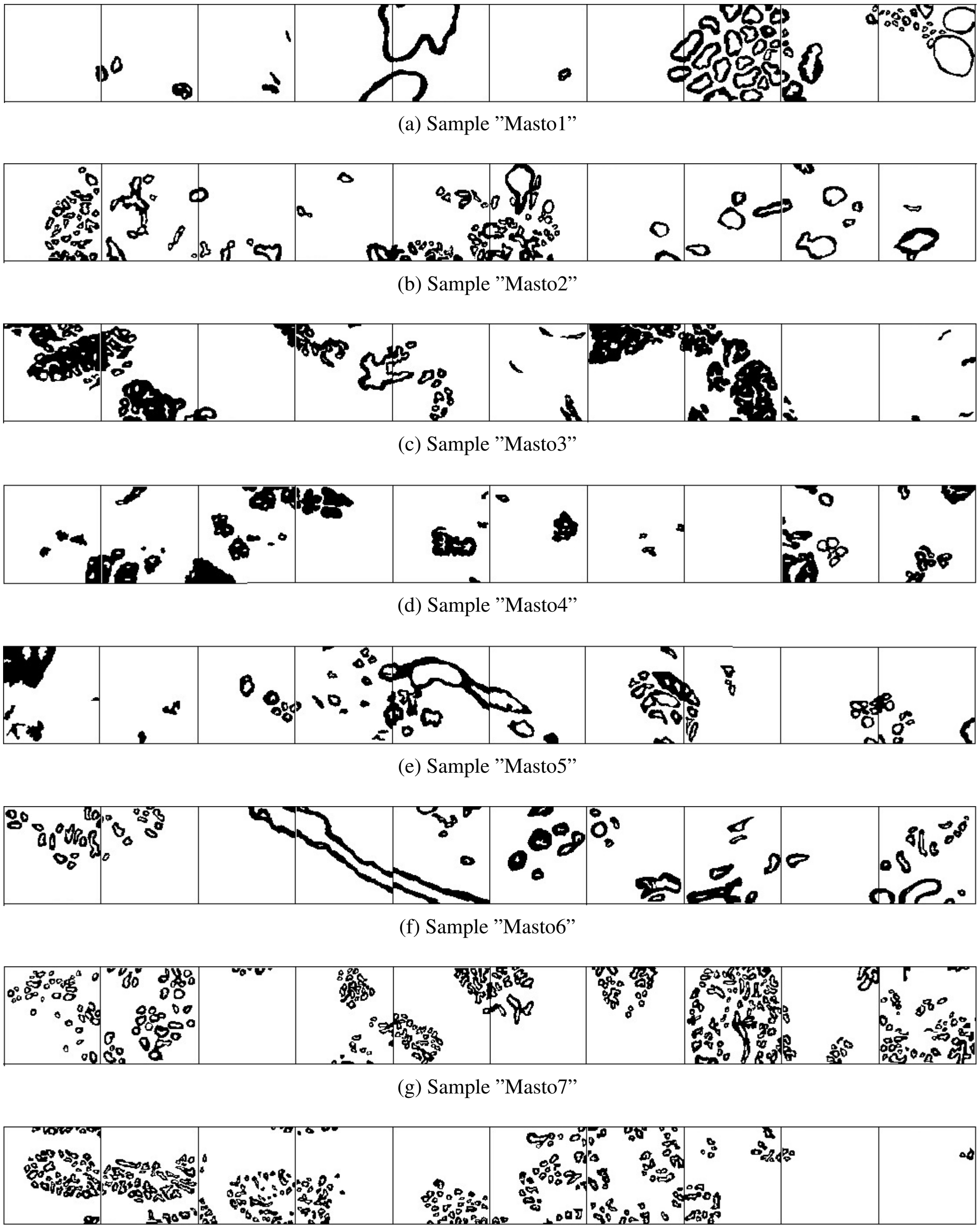


Fig. 7: Histograms of *p*-values for testing pairs of different simulated processes by the method from (Gotovac *et al.*, 2016) (histograms coloured light grey), by the method from (Gotovac and Helisová, 2019) (histograms coloured grey) and by the method from the present paper (histograms coloured in black) when using only components (first column), components and neighbourhood tessellations (second column) and only neighbourhood tessellations (third column). Here, we compare random-ellipse Boolean vs random-disc Boolean (first row), random-ellipse Boolean vs cluster (second row), random-disc Boolean vs cluster (third row). 100 realisations of each process were used. Percentages of *p*-values less or equal to 0.05 are displayed in legend.

study, where it has shown better results on simulated samples than the methodology introduced in (Gotovac *et al.*, 2016) and (Gotovac and Helisová, 2019) in cases when we have enough connected components fully visible within the observation window. The procedure was consequently applied to mastopatic breast tissue and mammary cancer data in order to distinguish between different types of tissue, as well as finding the most similar samples within the group of the same tissue type. The similarity matrix between the presented samples can be constructed using the *p*-values from Table 1, Table 2 and Table 3 and can be further used in various machine learning algorithms.

Furthermore, since this method is concentrated on the distribution of the components and their neighbourhoods, it would not be so sensitive to possible outliers. For example, if we have a few components that are significantly larger than the others in the sample, this similarity measure could oversee it. So, it is not recommended to use this procedure when these outliers play an important role in distinguishing between sets. As an alternative in these cases, one can use methodology from (Gotovac *et al.*, 2016) or (Gotovac and Helisová, 2019). Also, since the larger components are more likely to be touching the boundary of the observation window, they can be excluded from the sampling with a higher probability. In these cases, a visual inspection is recommended in order to make sure that all the important features of the realisation are preserved after the removing the components on the boundary. The safer alternative is to check whether the mean area of the sampled components is greater or equal to the mean area of the parts of the omitted components visible within the observation window. A more detailed simulation study



(a) Sample ”Masto1”

(b) Sample ”Masto2”

(c) Sample ”Masto3”

(d) Sample ”Masto4”

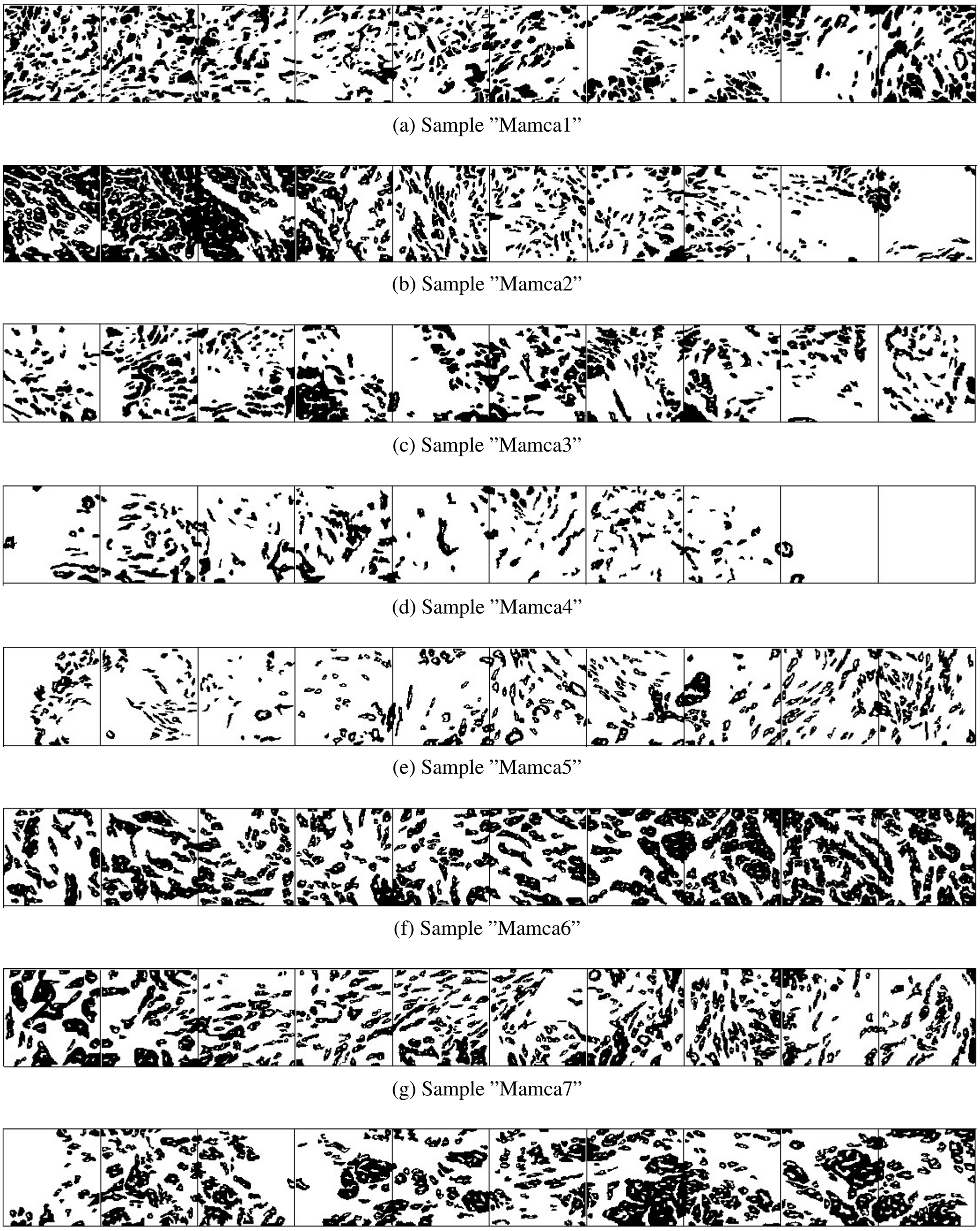
(e) Sample ”Masto5”

(f) Sample ”Masto6”

(g) Sample ”Masto7”

(h) Sample ”Masto8”

Fig. 8: Samples of masthopatic breast tissue kindly provided by Mrkvička and Mattfeldt (2011)



(a) Sample ”Mamca1”

(b) Sample ”Mamca2”

(c) Sample ”Mamca3”

(d) Sample ”Mamca4”

(e) Sample ”Mamca5”

(f) Sample ”Mamca6”

(g) Sample ”Mamca7”

(h) Sample ”Mamca8”

Fig. 9: Samples of mammary cancer kindly provided by authors of (Mrkvička and Mattfeldt, 2011)

Table 1: Tables of the p-values (similarity measures) between samples of masthopatic tissue (left) and between samples of mammary cancer tissue (right) ; Each row consists of two sub-rows: the first sub-row presents results of the test when only components are taken into the consideration (black) and the second sub-row consists of results of the test when both shape of components and their position are considered (dark grey); p-values larger than 0.05 are in bold.

	Masto2	Masto3	Masto4	Masto5	Masto6	Masto7	Masto8
Masto1	**0.199** **0.092**	0.001 0.026	0.004 0.004	0.024 0.017	0.039 0.023	0.001 0.001	0.001 0.002
Masto2		0.001 0.022	0.001 0.004	0.003 0.012	0.002 **0.075**	0.001 0.001	0.001 0.001
Masto3			**0.726** 0.001	0.037 0.007	0.042 0.039	0.001 0.001	0.001 0.001
Masto4				**0.507** **0.666**	**0.281** **0.226**	0.001 0.001	0.001 0.001
Masto5					**0.686** **0.393**	0.001 0.001	0.001 0.001
Masto6						0.001 0.001	0.001 0.001
Masto7							**0.160** **0.376**

	Mamca2	Mamca3	Mamca4	Mamca5	Mamca6	Mamca7	Mamca8
Mamca1	**0.052** **0.079**	**0.260** 0.046	**0.149** 0.001	0.026 0.001	0.001 0.001	0.001 0.001	0.001 0.001
Mamca2		**0.341** 0.004	**0.489** 0.001	**0.181** 0.001	0.001 0.001	**0.051** **0.085**	0.005 0.037
Mamca3			**0.386** 0.021	**0.163** **0.080**	0.001 0.001	0.001 0.004	0.002 0.001
Mamca4				**0.117** **0.095**	0.001 0.001	0.001 0.001	0.001 0.001
Mamca5					0.001 0.001	0.049 0.005	0.001 0.003
Mamca6						**0.681** **0.681**	**0.146** **0.067**
Mamca7							**0.054** **0.198**

on the samples of the processes from the *Simulation study* section showed that this bias in the sampling did not significantly affect the distribution of obtained p-values. However, in situations when one sample has larger components that are omitted and the other sample does not have those components, the bias in the sampling can lead to false high similarly values.

ACKNOWLEDGEMENTS

The research was supported by The Czech Science Foundation, project No. 19-04412S. The author is grateful to Kateřina Helisová and Lev Klebanov for their kind attention and useful advises.

REFERENCES

Chiu SN, Stoyan D, Kendall WS, Mecke J (2013). Stochastic geometry and its applications. John Wiley & Sons, New York.

Diggle PJ (1981). Binary mosaics and the spatial pattern of heather. Biometrics 37.3:531-9.

Gotovac V, Helisová K (2019+). Testing equality in distribution of random convex compact sets via theory of $\mathfrak{N}$-distances and its application to assessing similarity of general random sets. arXiv:1801.02090

Gotovac V, Helisová K, Klebanov LB, Volchenkova IV (2017). A new definiton of random sets. arXiv:1712.09452

Gotovac V, Helisová K, Ugrina I (2016). Assessing dissimilarity of random sets through convex compact approximations, support functions and envelope tests. Image Anal Stereol 35:181–93.

Helisová K (2014). Modeling, statistical analyses and simulations of random items and behavior on material surfaces. Supplemental UE: TMS 2014 Conference Proceedings, February 16-20, 2014, San Diego, California, USA. 461–8.

Hermann P, Mrkvička T, Mattfeldt T, Minárová M, Helisová K, Nicolis O, Wartner F, Stehlík M (2015). Fractal and stochastic geometry inference for breast cancer: a case study with random fractal models and Quermass-interaction process. Stat in Med 34.18:2636–61.

Klebanov LB (2006). $\mathfrak{N}$-distances and their applications. Karolinum Press, Charles University, Prague.

Kendall WS, Van Lieshout MNM, Baddeley AJ (1999). Quermass-interaction processes: Conditions for stability. Adv Appl Probab 31:31542.

Mrkvička T, Mattfeldt T (2011). Testing histological images of mammary tissues on compatibility with the Boolean model of random sets. Image Anal Stereol 30.1:11–8.

Molchanov I (2017). Theory of Random Sets. Springer, London.

Møller J, Helisová K (2008). Power Diagrams and Interaction Processes for Unions of Discs. Adv Appl Probab 40:321–47.

Møller J, Helisová K (2010). Likelihood inference for unions of interacting discs. Scand Stat 37:365–81.

Neumann M, Staněk J, Pecho OM, Holzer L, Beneš V, Schmidt V (2016). Stochastic 3D modeling of complex three-phase microstructures in SOFC-electrodes with completely connected phases. Comp Mat Sci 118: 353–64.