ASSESSING DISSIMILARITY OF RANDOM SETS THROUGH CONVEX COMPACT APPROXIMATIONS, SUPPORT FUNCTIONS AND ENVELOPE TESTS

VESNA GOTOVAC[1], KATEŘINA HELISOVÁ[✉,2] AND IVO UGRINA[3]

[1]Department of Mathematics, Faculty of Science, University of Split, 21000 Split, Croatia; [2]Department of Mathematics, Faculty of Electrical Engineering, Czech Technical University in Prague, 162 27 Prague 6 – Dejvice, Czech Republic; [3]Faculty of Pharmacy and Biochemistry, University of Zagreb, 10000 Zagreb, Croatia
e-mail: vgotovac@pmfst.hr, helisova@math.feld.cvut.cz, ivougrina@gmail.com

(Received February 24, 2016; revised November 16, 2016; accepted November 17, 2016)

ABSTRACT

In recent years random sets were recognized as a valuable tool in modelling different processes from fields like biology, biomedicine or material sciences. Nevertheless, the full potential of applications has not still been reached and one of the main problems in advancement is the usual inability to correctly differentiate between underlying processes generating real world realisations. This paper presents a measure of dissimilarity of stationary and isotropic random sets through a heuristic based on convex compact approximations, support functions and envelope tests. The choice is justified through simulation studies of common random models like Boolean and Quermass-interaction processes.

Keywords: approximations, dissimilarity, envelope tests, random sets, stochastic geometry, support functions
.

INTRODUCTION

The need for continuous improvement in fields like biology, biomedicine or material sciences urges us to develop and explore new techniques and models in applications. Over few years, one of these models received a lot of attention due to its strong mathematical foundations and general nature.

As a model, random sets can describe and explain many events. Some examples are dynamics of cells in organisms (Mrkvička and Mattfeldt, 2011; Hermann *et al.*, 2015), particles in materials (Helisová, 2014; Neumann *et al.*, 2016) or the presence of different plants in ecosystems (Diggle, 1981; Møller and Helisová, 2010).

Although three dimensional applications are of interest, applications of random sets are mainly done through two-dimensional modelling. We are often interested only in the projection of objects to the plane (*e.g.*, ground area of plants or trees) or we study only cross-sections of a mass which create planar formations and suppose that the behaviour of the studied object is stationary in the third dimension (*e.g.*, organic cells or material particles).

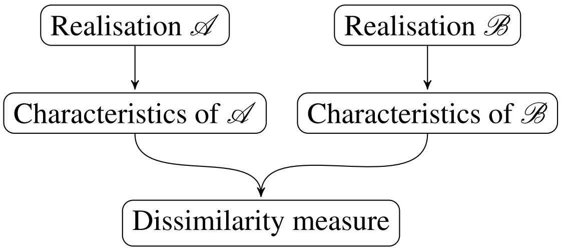
Random sets are represented by different models ranging from simple and intuitive ones to those that are highly complex and specialized. One of the basic models is the Boolean model (Chiu *et al.*, 2013). Due to its simplicity many theoretical results can be derived for it but, unfortunately, the model is not sufficient in many applications. For example, in Diggle (1981) researchers tried to model a data set concerning heather plants and the model proved to be unsatisfactory due to some interactions between bushes.

An example of a sophisticated model is Quermass-interaction process. This model introduces dependencies on geometrical characteristics of sets like area or boundary. It was first studied by Kendall *et al.* (1999) and later extended by Møller and Helisová (2008).

As with classical statistics, being able to specify a particular model (*e.g.*, distribution in hypothesis testing) is of great benefit. However, it is not necessarily feasible. Mrkvička and Mattfeldt (2011) explored tumour cells data with the Boolean model with the aim to use estimated parameters for detecting the differences between mammary cancer and masthopatic tissue. This model proved to be unsatisfactory since only 7% of images of mammary cancer and 1% of images of mastopathy were compatible with the model on the significance level of 0.05. Later in Hermann *et al.* (2015), the Quermass-interaction process was fitted to the same data, but the compatibility was again rejected in many cases (only 27.5% cases of mammary cancer and 6% cases of mastopathic tissue were accepted with a significance level of 0.05).

There are some instances where knowledge about the underlying model is not necessary, although it could be beneficial. For example, in classical hypothesis testing we could use rank based statistics without specifying strict probability distributions. Translated to the language of random sets, one is interested only in distinguishing between two sets seen as realisations of the same or different models. Exactly this problem, restricted only to stationary and isotropic random sets, will be the main topic explored in the rest of the paper.

In its most general form the problem can be described by the following diagram:


Realisation $\mathscr{A}$
Realisation $\mathscr{B}$
Characteristics of $\mathscr{A}$
Characteristics of $\mathscr{B}$
Dissimilarity measure

Given two realisations of stationary and isotropic random sets one would like to find characteristics useful when making a decision about strength of *dissimilarity* of the underlying processes that have generated those sets. Strength of dissimilarity will be represented as a number $\alpha \in [0,1]$, where the lower value of α means lower belief that the realisations come from the same underlying processes.

Notice that we have posed the problem starting with realisations of random sets. Contrary to the classical point measurements, where points can be described purely by numbers (up to an error), describing realisations of random sets is much harder. Especially in practice where one usually obtains an image in raster format (like JPG or PNG), *i.e.*, a grid of pixels (usually black and white) representing an observed realisation of a random set. Also, realisations of random sets are often of rugged shape and it is hard to work with them since even the description of the realisations is complicated.

Due to these reasons, it seems plausible to try to approximate realisations by sets of more appealing characteristics. One of the approaches, and the one taken in this article, is to use convex compact sets due to some useful properties describing them. Therefore, we shall assess dissimilarity of the underlying process generating realisations through assessing dissimilarity of convex compact sets approximating our realisations.

The approach is graphically presented in Fig. 1 and can be described in the following steps: starting with realisations of random sets (images in practice) find a suitable covering with simple (well described) sets like balls (step 1), break the covering into convex compact sets (step 2), derive some characteristics of this approximation (step 3) and use these characteristics to assess the dissimilarity of the underlying processes. To the best of our knowledge, this approach had, until now, not been taken into consideration in the known literature.

It is really important to stress that we are not interested in assessing the similarity of images representing realisations of random sets like in the literature on analysis of binary images (*e.g.*, Hall, 1985; Ohser and Mücklich, 2000, or Pratt, 2001) but on assessing the dissimilarity of underlying processes.

The paper is organised as follows. In the *Material and Methods* section we provide basic definitions, methods and procedures. Roughly, we define support functions, introduce a procedure for covering a set by discs, define Voronoi tessellation on the union of discs, introduce the theory behind envelope tests and give a few remarks on free parameters. In the *Results* section we present a simulation study and numerical results based on the methodology from the *Material and Methods* section. In the last section we give an overview of the presented results with comments and remarks.

MATERIAL AND METHODS

BASICS OF RANDOM SETS

In this section, we recall some terms related to the basic theory of random sets. Definitions 2-5 concerning basic terms can be found in (Chiu *et al.*, 2013) while Definition 6 of Quermass-interaction process in this form is taken from (Møller and Helisová, 2008). Readers familiar with this topic can skip to the following section.

Definition 1. *The set* $\mathbf{K} \subset \mathbb{R}^d$ *is said to be* convex *if* $cx + (1-c)y \in \mathbf{K}$ *for all* $x, y \in \mathbf{K}$ *and all* $0 < c < 1$. *The set* $\mathbf{K} \subset \mathbb{R}^d$ *is said to be* compact *if it is closed and bounded.*

Definition 2. *Let* $(\Omega, \mathscr{F}, P)$ *be a probability space,* $\mathscr{C}$ *the system of closed sets in* $\mathbb{R}^d$ *and* $\mathfrak{C} = \sigma\{\mathscr{C}^K : K \text{ is a compact subset of } \mathbb{R}^d\}$, *where* $\mathscr{C}^K = \{C \in \mathscr{C} : C \cap K \neq \emptyset\}$. *Then a* random closed set $\mathbf{X}$ *in* $\mathbb{R}^d$ *is a measurable mapping from* $(\Omega, \mathscr{F})$ *to* $(\mathscr{C}, \mathfrak{C})$.

Definition 3. The distribution $P_{\mathbf{X}}$ of a random set $\mathbf{X}$ *is given by the relation* $P_{\mathbf{X}}(F) = P(\{\omega \in \Omega : \mathbf{X}(\omega) \in F\})$ *for* $F \in \mathfrak{C}$.

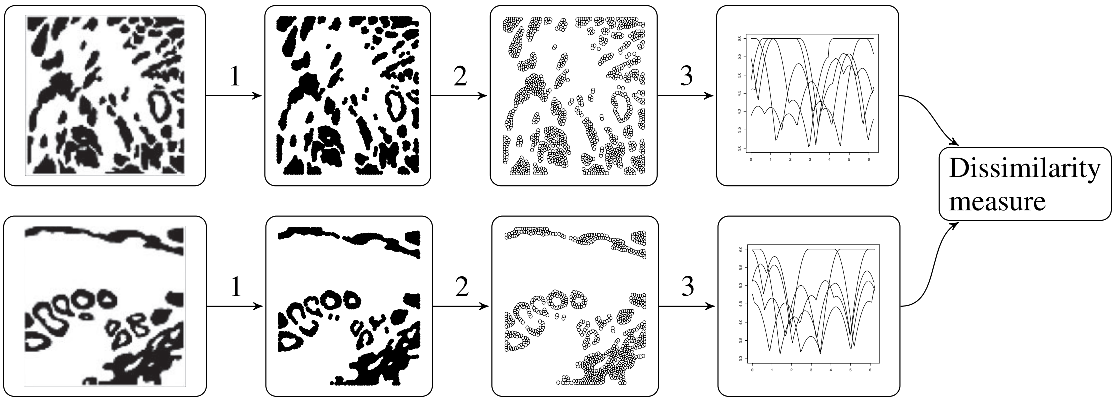


Fig. 1. *Steps in assessing dissimilarity of the underlying processes generating realizations through dissimilarity convex compact sets approximating realizations. The original images (left) were kindly provided by Mrkvička and Mattfeldt (2011).*

Definition 4. *A random 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* $\mathbf{X}$*. A random set* $\mathbf{X}$ *is* isotropic *if its distribution is invariant under rotation. If a random set is both stationary and isotropic, it is called* motion invariant.

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 d-dimensional Lebesgue measure of the set A.

Definition 5. *Let* $Y = \{y_1, y_2, \dots\}$ *be a stationary Poisson point process in* $\mathbb{R}^d$ *and* $\{\mathbf{B}_1, \mathbf{B}_2, \dots\}$ *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*

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

is called Boolean model.

Definition 6. *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.* Quermass-interaction process *is a random set whose probability measure is absolutely continuous with respect to the probability measure of the given Boolean model and 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.*

SUPPORT FUNCTION OF A CONVEX COMPACT SET

In the following text, convex compact sets have a central role in our approach and we are restricted only to the planar case, *i.e.*, the convex compact sets in $\mathbb{R}^2$. One of the most important basic concepts related to convex compact sets is the support function.

Definition 7. *For a (random) convex compact set* $\mathbf{X}$ *define its support function as*

$$h_\mathbf{X}(u) = \sup_{x \in \mathbf{X}} \langle u, x \rangle, \quad u \in \partial b(0,1),$$

where $\partial b(0,1)$ *is the unit sphere in* $\mathbb{R}^2$*, i.e., the circle with radius 1 and the centre in the origin.*

Fig. 2 illustrates the support function of two sets, a disc and a square centred in the origin.

The importance of the support function in our approach is derived from the following result on the equality of distributions of two random convex compact sets via their support functions (for the proof please take a look at Lavie (2000)).

Theorem 8. *For two random convex compact sets* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *it holds*

$$\mathbf{X}_1 =^{(\mathscr{D})} \mathbf{X}_2 \Leftrightarrow (h_{\mathbf{X}_1}(u_i))_{i \in I} =^{(\mathscr{D})} (h_{\mathbf{X}_2}(u_i))_{i \in I}$$

for all finite index sets I *and* $(u_n)_{n \in \mathbb{N}}$ *dense subset of* $\partial b(0,1)$.

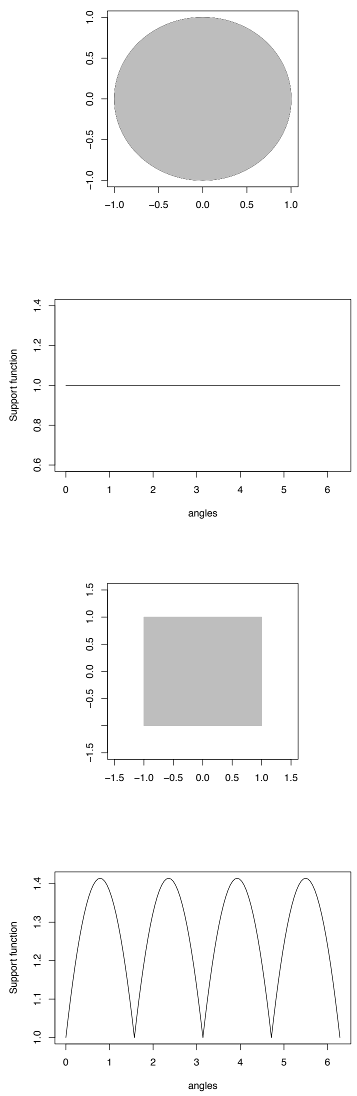


Fig. 2. *Examples of the support function for a disc and a square, respectively.*

COVERING A SET BY DISCS WITH IDENTICAL RADII

As mentioned in the *Introduction*, our aim is to divide a planar data set $\mathbf{S}$ (*e.g.*, Fig. 3a) into a family of convex compact sets.

In practice we usually start with a digital binary image $\mathbf{I}$ containing the observed set, *i.e.*, a picture consisting of black and white pixels, as an approximation of the initial set and its complement. This approximation can be achieved by pixelisation through a simple structure, where pixels are squares with side length Δ and the binary image $\mathbf{I}$ containing an arbitrary planar set $\mathbf{S}$ is a matrix $M = [m_{i,j}]$ of black and white pixels (or the matrix of 1's and 0's) such that the pixel $m_{i,j}$ is black if and only if its centre lies in the set $\mathbf{S}$, *i.e.*, if and only if $(i \cdot \Delta - \Delta/2, j \cdot \Delta - \Delta/2) \in \mathbf{S}$, and it is white otherwise (Fig. 3b). Let us denote by $\mathbf{D}$ the approximation of $\mathbf{S}$ in the digital image $\mathbf{I}$.

Having a digital approximation $\mathbf{D}$ of our initial set $\mathbf{S}$, it can be useful to approximate the shape of the given set $\mathbf{S}$ by a geometrical object $\mathbf{A}$ with suitable mathematical properties. One of the approaches, and the one taken in this article, would be to cover the pixelised version with discs of identical radii. This can be achieved by utilising a customized version of maximal Poisson-disc sampling algorithm (Ebeida *et al.*, 2011).

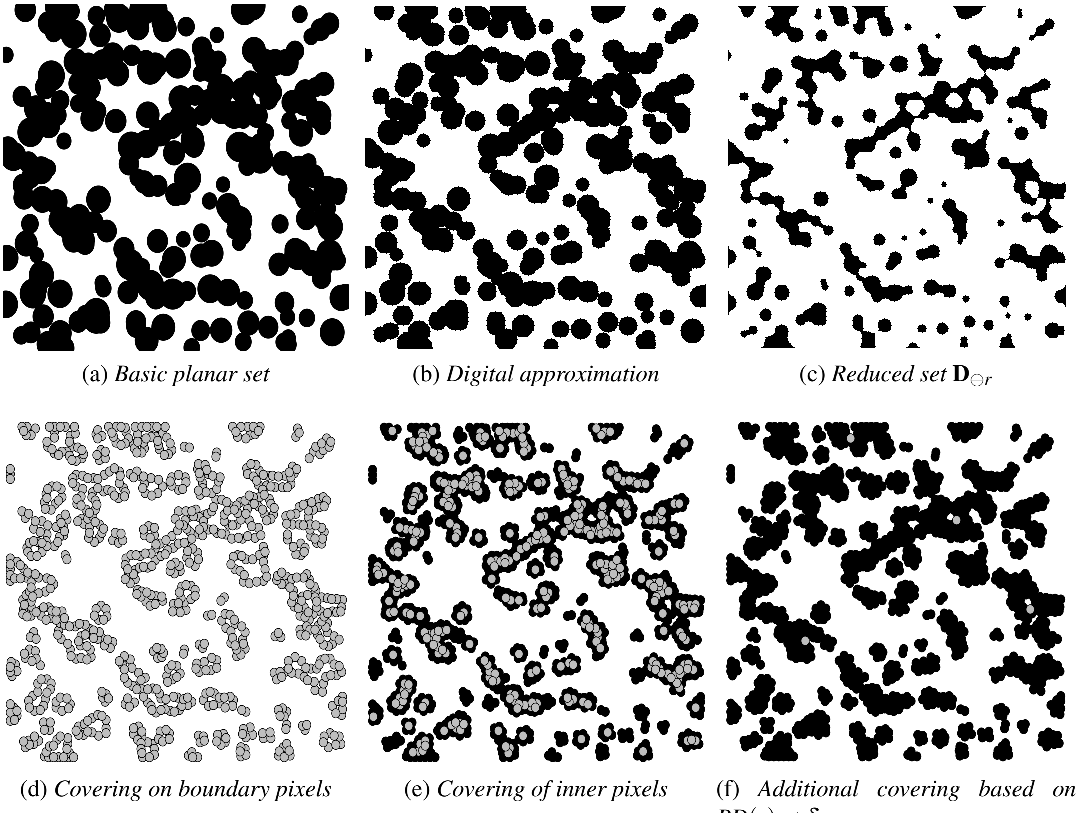
In the maximal Poisson-disc sampling algorithm, set $\mathbf{D}$ is covered by discs $b(x_i, r)$ with centres x_i and radii r so that the discs centres come from the Poisson-disc sampling process, *i.e.*, the point process $X = \{x_i; i = 1, \ldots, n\}$ is constructed point by point satisfying the following conditions:

1. $\forall x_i, x_j \in X, x_i \neq x_j : ||x_i - x_j|| \geq r,$
2. If $\mathbf{D}_{i-1} = \mathbf{D} \setminus \cup_{j=1,\ldots,i-1} b(x_j, r)$, then $\forall x_i \in X, \forall \mathbf{B} \subset \mathbf{D}_{i-1}$:
$$P(x_i \in \mathbf{B}) = \frac{|\mathbf{B}|}{|\mathbf{D}_{i-1}|} \,,$$
3. $\forall x \in \mathbf{D} \; \exists x_i \in X : ||x - x_i|| < r.$

It is obvious that in this approach the area of approximating set is larger than the area of the original set $\mathbf{D}$. Depending on the choice of radius r this can lead to many white pixels of the digital image $\mathbf{I}$ becoming black, therefore leading to an increase in geometrical properties like area or boundary length. To account for this we introduce a customized version of the maximal Poisson-disc sampling (called CMPD in future) such that we first reduce the set $\mathbf{D}$ by a radius r, *i.e.*, we construct
$$\mathbf{D}_{\ominus r} = \{u : b(u, r) \subseteq \mathbf{D}\} \,,$$
(Fig. 3c). Now, we cover the newly obtained set $\mathbf{D}_{\ominus r}$ by the maximal Poisson-disc sampling obtaining a set $\hat{\mathbf{D}}_{\ominus r}$. In order to keep the original shape as precise as possible, we start covering on the border of $\mathbf{D}_{\ominus r}$, *i.e.*, the centres are first sampled from the boundary pixels (Fig. 3d). After covering all the boundary pixels we choose the centres from the remaining black pixels of $\mathbf{D}_{\ominus r}$ (Fig. 3e).

However, since set $\hat{\mathbf{D}}_{\ominus r}$ will be a subset of the starting set $\mathbf{D}$ we can try to cover what has not yet been covered, *i.e.*, $\tilde{\mathbf{D}} = \mathbf{D} \backslash \hat{\mathbf{D}}_{\ominus r}$, using the conditions 1 and 2 from the maximal Poisson-disc sampling and adjusted condition 3:

1. $\forall x_i, x_j \in X, x_i \neq x_j : ||x_i - x_j|| \geq r,$



(a) *Basic planar set* (b) *Digital approximation* (c) *Reduced set* $\mathbf{D}_{\ominus r}$

(d) *Covering on boundary pixels* (e) *Covering of inner pixels* (f) *Additional covering based on* $PD(r) < \delta$

Fig. 3. *Stepwise covering of a planar set by discs of identical radii*

2.If $\mathbf{D}_{i-1} = \tilde{\mathbf{D}} \setminus \cup_{j=1,\dots,i-1} b(x_j, r)$, then $\forall x_i \in X, \forall \mathbf{B} \subset \mathbf{D}_{i-1}$:

$$P(x_i \in \mathbf{B}) = \frac{|\mathbf{B}|}{|\mathbf{D}_{i-1}|} ,$$

3.$PD(r) < \delta$,

where $PD(r)$ is a measure of pixel difference defined below by (2) and δ is some threshold. In words, this can be described as following the same principal of covering as with the maximal Poisson-disc sampling except that we do not want to completely cover the original set but only up to a predefined similarity (measured by PD). Therefore, we add more discs as long as $PD(r) < \delta$ is valid (Fig. 3f).

It should be noted that depending on the choice of δ covering with a predefined radius r might not be possible at all. Also, adding a disc to $\mathbf{D} \setminus \hat{\mathbf{D}}_{\ominus r}$ might fail the $PD(r) < \delta$ condition for the first random centre (disc) while for some other centres it could work. In the rest of this paper, we have decided to stop with the first disc failing, therefore not trying with other/additional (random) centres. This choice was based on our computational restrictions (feasibility).

Simulation study presented in *Results* section justifies (in the sense of pixel difference (Eq. 2), see below) the choice of covering starting on the boundary.

The choice of the pixel difference measure is based on the idea of precision from the field of binary classification and is given by the following formula

$$PD(r) = \frac{BW(r) + WB(r)}{B^{(\mathrm{orig})}} , \qquad (2)$$

where BW denotes the number of pixels which were black in the digital image I and are white after the covering, WB denotes the number of pixels which were white in the digital image I and are black after the covering, and $B^{(\mathrm{orig})}$ denotes the number of all black pixels in the original digital image $\mathbf{I}$, *i.e.*, the area of $\mathbf{D}$ (expressed in pixels).

It could be tempting to conclude that instead of the CMPD we could use the maximal Poisson-disc sampling and just remove pixels that are outside of the original set $\mathbf{D}$. Although this approach would be much simpler it would suffer from a major drawback since some covering sets would not be convex any more and convexity is one of our important assumptions to use support functions.

One problem that still remains open is the choice of the optimal radius for covering. The optimal radius would be the one where discs overlap suitably in the sense that they are not too dense (due to consequent construction of the corresponding Voronoi tessellation), the union covers the whole set satisfactorily (in the sense of condition 3 for CMPD), the radius should not be too small because it would not provide satisfactory information about inner structure of the set $\mathbf{D}$ (support functions) and the radius should not be too large since we would like to obtain sufficient number of convex compact objects (cells of the corresponding Voronoi tessellation in our case). This is needed for testing whether the cells of Voronoi tessellations corresponding to two different realisations of random sets come from the same distribution. More on the choice of radius and δ is given later in the text.

VORONOI TESSELLATION ON A DISC UNION

The final step for obtaining a family of convex compact sets from the data set $\mathbf{D}$ is to construct a tessellation over the union of discs $\mathbf{A}$ approximating the set $\mathbf{D}$. We took the inspiration from the study of power tessellation of a union of discs (for construction, properties etc., see Møller and Helisová, 2008) which is the intersection of the union of discs and Laquerre tessellation (*e.g.*, Imai *et al.*, 1985) built over discs centres, where the weights are given by the radii of the discs. Here we construct its simpler version for discs having the same radii, *i.e.*, for the case when the Laquerre tessellation corresponds to its special case, Voronoi tessellation (*e.g.*, Chiu *et al.*, 2013). Thus the definition of the tessellation we will use is the following:

Definition 9. *Consider a finite configuration of discs* $\{\mathbf{b}_1 \dots, \mathbf{b}_n\}$ *with centres* $\{c_1 \dots, c_n\}$ *and identical radii, and denote by*

$$B_i = \{y \in \mathbf{b}_i \ : \ \|y - c_i\| \leq \|y - c_j\| \, \textit{for all } j \neq i\}.$$

The system $\mathscr{B}$ *of all sets* B_i *is called the* Voronoi tessellation on a union of discs.

It is easy to see that all cells of a Voronoi tessellation are convex. Moreover, since we intersect them with discs they are compact. Thus, applying this procedure to the approximation given by discs, we obtain a family of convex compact sets from the data set $\mathbf{D}$. This approximation by a union of convex compact sets can now serve as a basis for a dissimilarity measure.

ENVELOPE TESTS

In this section, we introduce a tool (test) used for testing equality of distributions of random geometrical objects recently developed by Myllymäki *et al.* (2016) and Mrkvička *et al.* (2015). Also, we describe the customization of the test to fit our needs.

Consider a group of $s+1$ geometrical objects, where k-th object, $k = 1,...,s+1$, is described by characteristics $T_k(\varphi)$ for $\varphi \in I$, where I is a finite index set, so $T_k(\varphi)$ may be considered as a vector of finite dimension. We suppose that $T_k(\varphi)$, $k = 2,...,s+1$, come from the same distribution, and we would like to test the hypothesis that $T_1(\varphi)$ also comes from the same distribution as $T_k(\varphi)$, $k = 2,...,s+1$.

Suppose that $T_k(\varphi)$, $k = 1,...,s+1$, are evaluated on some a priori chosen mesh of values φ_j, $j = 1,\dots,n$. Let $r_{1,j},\dots,r_{s+1,j}$ denote raw ranks of $T_1(\varphi_j),\dots,T_{s+1}(\varphi_j)$ such that the smallest $T_k(\varphi_j)$ for a fixed j has rank 1, the second smallest $T_k(\varphi_j)$ has rank 2, etc., and the largest $T_k(\varphi_j)$ has rank $s+1$. In case of ties, the raw ranks are averaged within ties. The point-wise ranks are calculated as

$$R_{k,j} = \min\left\{r_{k,j}, s+1-r_{k,j}\right\}.$$

The extreme rank measure R_k of the vector T_k is the minimum of the point-wise ranks $R_{k,j}$, *i.e.*,

$$R_k = \min_{j\in\{1,\dots,n\}} R_{k,j} \, .$$

Now, for a given significance level α, we can define a test as follows. First, we determine an appropriate low rank $R_{(\alpha)}$ which is the smallest value in $\{R_1,\dots,R_{s+1}\}$ for which

$$\sum_{k=1}^{s+1} 1(R_k \leq R_{(\alpha)}) \geq \alpha(s+1)\,.$$

Secondly, for

$$I_\alpha = \left\{k \in \{1,\dots,s+1\} : R_k \geq R_{(\alpha)}\right\}$$

we define values

$$T^{(\alpha)}_{low}(\varphi_j) = \min_{k\in I_\alpha} T_k(\varphi_j), \quad T^{(\alpha)}_{upp}(\varphi_j) = \max_{k\in I_\alpha} T_k(\varphi_j)\,.$$

The following results from Myllymäki *et al.* (2016) now lead to the definition of the rank envelope test and present its statistical justification:

$$P\left(T_1(\varphi_j) \notin \left[T^{\alpha}_{low}(\varphi_j), T^{(\alpha)}_{upp}(\varphi_j)\right] \text{ for any j} \,|H_0\right) \leq \alpha$$

and

$$P\left(T_1(\varphi_j) \notin \left(T^{\alpha}_{low}(\varphi_j), T^{(\alpha)}_{upp}(\varphi_j)\right) \text{ for any j} \,|H_0\right) > \alpha,$$

where H_0 is a simple null hypothesis.

Therefore, if the observed vector T_1 leaves envelope at some point, *i.e.*, $R_1 < R_{(\alpha)}$ the null hypothesis is rejected at significance level α. If the observed vector lies completely inside this envelope, *i.e.*, $R_1 > R_{(\alpha)}$, the null hypothesis is not rejected at significance level α. If the observed vector coincides in some point with the border of the envelope, *i.e.*, $R_1 = R_{(\alpha)}$ the rejection of the null hypothesis remains undecided.

Testing can also be conducted through a customized version of p-values which are defined by assigning an extreme rank measure R_k to each of the vectors T_k, such that the lowest rank corresponds to the most extreme values of statistic:

$$p_- = \frac{1}{s+1}\sum_{k=1}^{s+1}\mathbf{1}(R_k < R_1)\,,$$

$$p_+ = \frac{1}{s+1}\sum_{k=1}^{s+1}\mathbf{1}(R_k \le R_1)\,.$$

From the definition of p-values it is easy to see that $R_1 < R_{(\alpha)}$ is equivalent to $p_+ \le \alpha$ (null hypothesis rejected), $R_1 > R_{(\alpha)}$ is equivalent to $p_- > \alpha$ (null hypothesis not rejected) and $R_1 = R_{(\alpha)}$ is equivalent to $p_- \le \alpha < p_+$ (the rejection of the null hypothesis remains undecided).

The width of the interval between p_+ and p_- is

$$p_+ - p_- = \frac{1}{s+1}\sum_{k=1}^{s+1}\mathbf{1}(R_k = R_1) \le \frac{2n}{s+1}\,.$$

and in the case of lots of ties it can be large. This problem can sometimes be solved by additional ordering as described by Myllymäki *et al.* (2016). It consists in replacing ranks R_k by vector ranks $\mathbf{R}_k$ which allow finer ordering among ranks so that in some cases it eliminates ties. Consider the vectors of point-wise ordered ranks $\mathbf{R}_k = (R_{k,(1)},\dots,R_{k,(n)})$, where $\{R_{k,(1)},\dots,R_{k,(n)}\} = \{R_k(\varphi_1),\dots,R_k(\varphi_n)\}$ and $R_{k,(j_1)} \le R_{k,(j_2)}$ whenever $j_1 \le j_2$. Then

$$\begin{aligned}&\mathbf{R}_{k_1} \prec \mathbf{R}_{k_2} \Leftrightarrow \exists n_0 \le n:\\ &\quad R_{k_1,(j)} = R_{k_2,(j)}\ \forall j < n_0\ \&\ R_{k_1,(n_0)} < R_{k_2,(n_0)}\,.\end{aligned} \quad (3)$$

Using $\mathbf{R}_k$ instead of R_k, $k = 1,\dots,s+1$, for calculation of the width $p_+ - p_-$ of the p-value interval, it may theoretically happen that the value is the same as for R_k, but in the most cases it is significantly narrower.

Now lets describe the application of this test to our data.

In our case, we compare two realisations of random sets in the way that first, each of the sets is covered by discs with identical radii, then the corresponding Voronoi tessellation is constructed, and finally, m non-neighbouring cells from each of the Voronoi tessellations are sampled. The sampling is carried out so that the first cell B_1 is chosen randomly uniformly from all the cells in the corresponding tessellation and having k sampled cells, the $(k+1)$-th cell B_{k+1} is chosen randomly uniformly from the set of cells $\{B_j : B_j \cap B_{k+1} = \emptyset\}$ for $k = 1,\dots,m-1$. It means that our data consist of $2\times m$ characteristics of convex compact sets, more precisely of $2\times m$ support functions related to the centres of the corresponding covering discs and evaluated in some (equidistant) partition of $(0,2\pi]$. Thus we do not explicitly have only one characteristic $T_1(\varphi)$ to be compared to $T_k(\varphi)$, $k = 2,\dots,s+1$. Therefore we decided to use a permutation version of the above-mentioned test to attain functional ANOVA procedure (Mrkvička *et al.*, 2015). It works as follows.

We form matrices H_A and H_B by putting in their rows discretized support functions from the first set and the second set, respectively. We denote by $H_A(\varphi_j)$ and $H_B(\varphi_j)$, respectively, the $j-$th column of matrices H_A and H_B while the column represents values of support functions of sample cells evaluated at angles $\varphi_j = j\frac{2\pi}{n}$, $j = 1,2,\dots,n$. Further, denote by $\overline{H_A(\varphi_j)}$, $\overline{H_B(\varphi_j)}$, $\mathrm{Var}(H_B(\varphi_j))$ and $\mathrm{Var}(H_A(\varphi_j))$ the mean values and variances, respectively, of each column in each matrix.

Assume that there exist non random functions $\mu_A(\varphi)$ and $\mu_B(\varphi)$ such that

$$H_{A,i}(\varphi) = \mu_A(\varphi) + e_{A,i}(\varphi)\,,$$

$$H_{B,i}(\varphi) = \mu_B(\varphi) + e_{B,i}(\varphi)$$

for $i = 1,\dots,m$, where $e_{A,i}(\varphi)$ and $e_{B,i}(\varphi)$ are i.i.d. samples from distribution $G(\varphi)$ for every φ, where $G(\varphi)$ has zero mean and finite variance. We want to test the hypothesis

$$H_0 : \mu_A(\varphi) - \mu_B(\varphi) \equiv 0.$$

The characteristic $T_1(\varphi_j)$ is the mean difference of support functions $H_A(\varphi_j)$ and $H_B(\varphi_j)$ normalised to unit variance, *i.e.*,

$$T_1(\varphi_j) = \frac{\overline{H_A(\varphi_j)} - \overline{H_B(\varphi_j)}}{\sqrt{\mathrm{Var}(H_A(\varphi_j)) + \mathrm{Var}(H_B(\varphi_j))}}$$

for $j = 1,2,\dots,n$. Characteristics $T_k(\varphi_j)$ for $k = 2,\dots,s+1$ are given by

$$T_k(\varphi_j) = \frac{\overline{H_1'(\varphi_j)} - \overline{H_2'(\varphi_j)}}{\sqrt{\mathrm{Var}(H_1'(\varphi_j)) + \mathrm{Var}(H_2'(\varphi_j))}}\,,$$

where H_1', and H_2' are obtained by randomly permuting rows of matrix $\begin{bmatrix} H_A \\ H_B \end{bmatrix}$ and then splitting back into two matrices with equal number of rows.

An example of graphical output of the test is shown in Fig. 4. The introduced plots come from the simulation study described in details below in *Results* section.

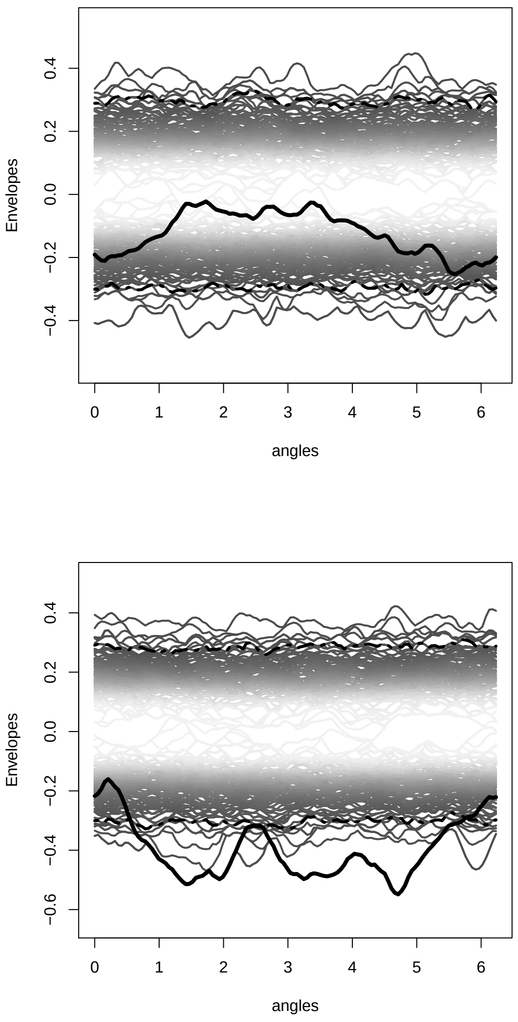


Fig. 4. *Graphical representation of envelopes for number of permutations $s = 4999$ when comparing two Boolean models, where $[p_-, p_+) = [0.1726, 0.1774)$ (left), and Cluster model with Repulsive model, where $[p_-, p_+) = [0.0054, 0.0102)$ (right), for details concerning the mentioned models Simulated data section. Dot-dashed black line represents envelope for $\alpha = 0.05$.*

ALIGNMENT OF SUPPORT FUNCTIONS

Since we assume that the distribution of original random set is isotropic, the same should be true for the distribution of a randomly chosen cell in the Voronoi tessellation of its coverage. However, the mean value of support functions of two cells having identical shape but being rotated by different angles could significantly differ from their original support function. In order to avoid this "loss of information", we cluster and align support functions of cells that are "almost equal" if transformed under rotation. Alignment is achieved through an adjusted version of agglomerative hierarchical clustering (James *et al.*, 2013) with average linkage and correlation as the dissimilarity measure.

In more details, let us denote by $H = [h_{i,j}] = [h_i(\varphi_j)]$ a matrix, where $h_i(\varphi_j)$ stands for the value of support function for the i-th sampled cell of the approximating tessellation at the j-th angle on a mesh of angles (*e.g.*, $\varphi_j = j\frac{2\pi}{n}, j = 1, \ldots, n$). Also, let us denote with $h_{i,1:}$ the i-th row of the matrix H and with $h_{i,l:} = (h_{i,l}, h_{i,l+1}, \ldots, h_{i,n}, h_{i,1}, \ldots, h_{i,l-1})$ a shifted version of $h_{i,1:}$ by l places. Now, for rows i_1 and i_2 we can define the value

$$\widetilde{cor}(i_1, i_2) = \max\left\{\mathrm{cor}\left(h_{i_1,1:}, h_{i_2,l:}\right) : l = 1, 2, \ldots, n.\right\}.$$

Using this newly defined measure of similarity between rows of H we can build a dendrogram (or clustering tree) for rows thus obtaining the clustering between rows based on the similarity between support functions regardless of rotation. As with almost all applications of hierarchical clustering a threshold C for the correlation should be defined by a researcher to cut the dendrogram into clusters. Alignment is now performed within the obtained clusters by shifting appropriate rows to achieve maximal $\widetilde{cor}$.

Since the main problem in this paper is comparing two random sets, matrix H is defined as $H = \begin{bmatrix} H_A \\ H_B \end{bmatrix}$ where H_A and H_B are matrices representing support functions for the first and the second set, respectively. Therefore, alignment is done together for both datasets. This approach removes a problem with the choice of the starting row within a cluster to align with since all *similar* tessellations will be aligned in the same way regardless of affiliation to the first or the second random set.

The effects of such alignment when comparing realisations from different underlying processes conducted in our later simulation study can be seen in Fig. 5.

CHOOSING "FREE" PARAMETERS

Following the aforementioned remarks on radius we define a criterion for choosing the suitable radius

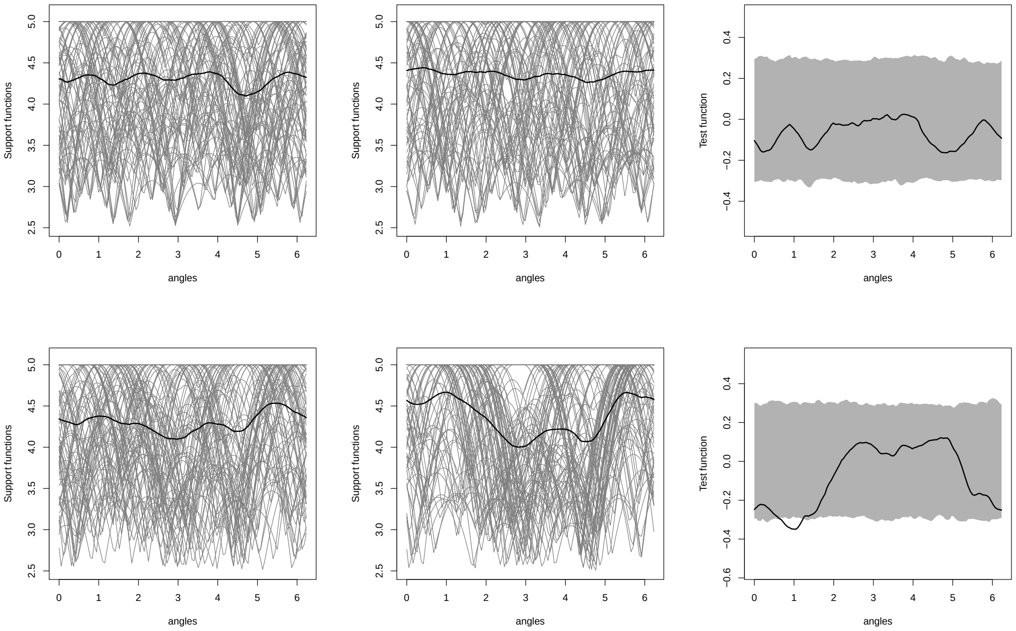


Fig. 5. *Support functions of 100 independent and randomly chosen cells from Boolean model (left) and Repulsive model (middle) with black lines representing mean values of these functions, and graphical representation of envelope test for these two models without alignment (upper line) and with alignment for C=0.9 (lower line).*

(expressed again in pixels) as

$$R = \max_{i=1,2,\dots} \{i : PD(i) \le \delta\} , \tag{4}$$

where δ is some predefined pixel difference level.

Note that the values $PD(i)$ depend on random covering, so it is a random variable and therefore the value R is random, too. In the section on simulation studies, we have a lot of input realisations of the same random set, thus we use the mean value of $PD(i)$ of all these realisations to find the suitable radius R. In real applications where usually only one realisation is available, we can take R calculated from only one covering of that realisation since our simulation study suggests that the values $PD(i)$ for a given realisation do not change significantly.

Since we are comparing two data sets $\mathbf{D}_A$ and $\mathbf{D}_B$ it may happen that the corresponding optimal radii R_A and R_B are different. In that case we define R as

$$R = \min\{R_A, R_B\} . \tag{5}$$

RESULTS

SIMULATED DATA

We applied the methodology described in *Material and Methods* section to a few preselected pairings between three different simulated realisations of random sets. One random set was a Boolean model (see Definition 5) and the other ones were random disc Quermass-interaction processes (see Definition 6) with parameters chosen so that the processes formed clusters or mutually repulsive components, respectively.

More precisely, the Boolean model used for the simulation study had centres of discs in the window 25×25, intensity of the disc centres equal to 0.4 and uniform distribution of radii on the interval $(0.5, 1)$ (see the left picture in Fig. 6). The second simulated data set was given by realisations of Quermass-interaction process with parameters $\theta_1 = 0.65$, $\theta_2 = -1.1$ and $\theta_3 = 0$ with respect to the mentioned Boolean model. Since this process produces realisations with larger area and smaller perimeter compared to the reference process, it tends to create clusters (see the middle picture in Fig. 6). Therefore, we will refer to it as the cluster process in the rest of the text. On the other hand, the third data set simulated as Quermass-interaction process with parameters $\theta_1 = -1$, $\theta_2 = 1$ and $\theta_3 = 0$ prefers smaller area and larger perimeter. Realisations are usually small non-overlapping components (see the right picture in Fig. 6) and therefore the process will be referenced as the repulsive process in the rest of the text.

We simulated 400 realisations for each of the mentioned processes. All realisations were transformed to matrices of 400×400 black and white pixels (an example of such a matrix for the Boolean model is shown in Fig. 3b) and these matrices played the role of the input data.

To explore the sensitivity of the methodology within and between classes of processes the following approach was taken. First, the input matrices were divided into two groups of 200 realisations for Boolean model, cluster process and repulsive process, respectively, and dissimilarity was studied on these groups separately for different processes. Additionally, we considered two groups of 200 realisations of different processes, namely Boolean vs cluster, Boolean vs repulsive and cluster vs repulsive process, and studied the dissimilarity again.

CHOOSING THE OPTIMAL RADIUS

In order to choose optimal radii, *i.e.*, to determine a suitable pixel difference level δ for approximations of data sets by unions of discs, we studied the behaviour of the approximations for different values of radii and their corresponding pixel difference levels. Data sets introduced in the *Results* section were covered by discs with identical radii using the method described in the *Material and Methods* section and taking the values of radii $r = 3,\ldots,15$ while the corresponding pixel differences $PD(r)$ were calculated. Consequently the Voronoi tessellations on unions of covering discs were constructed as described in the *Material and Methods* section. Then we studied the quality of coverings for chosen pixel differences levels $\delta = 10\%, 20\%, 30\%$ by visual comparison of the shape of original and approximating sets. Moreover, we explored the shape of cells in the corresponding Voronoi tessellations as well as their quantities.

For a realisation of the Boolean model, three different approximations using radii $r = 4,7,9$ corresponding to pixel differences levels $\delta = 10\%, 20\%, 30\%$, respectively, are presented in Fig. 7. By visual comparison of the similarity of the original set and Voronoi tessellations of its approximations, we have inferred that for a pixel difference level of 10% cells of the tessellation are too small thus do not give us enough information on the inner structure of the original set while for a pixel difference level of 30% the approximation of the original data set is not good enough in description of its shape. However, for a pixel difference level of 20% covering discs approximate original set shape more appropriately and thereby the corresponding tessellation provides good information about inner structure of the original set, too. Thus, we have chosen the pixel difference level $\delta = 20\%$.

It is easy to notice that the optimal radius gives us information about the structure of the underlying random set and it could be tempting to conclude that the difference between optimal radii in models could be used as the criterion for the final decision on dissimilarity. However, one should be careful in using it as the criterion for the final decision since different realisations of the same process may have different optimal radii and, moreover, since covering is random, even one realisation may result in significant differences of PD for the same radii, see Fig. 8, leading to different optimal radii.

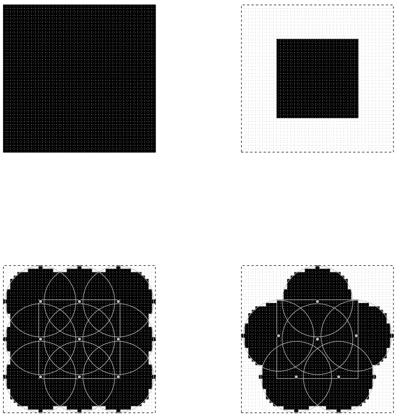


Fig. 8. *The original set in resolution 43×43, its reduced form and two ways of covering by radius 10, where in the first case, we have 268 white pixels, i.e.,* $PD(10) = 268/43^2 = 0.14$*, and in the second case, we have 606 white pixels so we obtain* $PD(10) = 606/43^2 = 0.33$.

NUMERICAL RESULTS OF ENVELOPE TEST

As mentioned in the previous subsection, we considered 400 realisations of each of the Boolean model, cluster process and repulsive process and studied the dissimilarity between realisations of the same processes as well as between realisations of different processes. Therefore, we had six different combinations of pairs in the study.

First, we had to choose the optimal radius for each of studied process. Using pixel a difference level of 20% as described in *Material and Methods* section we obtained from (4) the optimal radii $R = 7$ for the Boolean model, $R = 13$ for the cluster process and $R = 5$ for the repulsive process. Recall that these values were derived so that in (4) we considered the mean pixel difference $PD(R)$ calculated from all 400

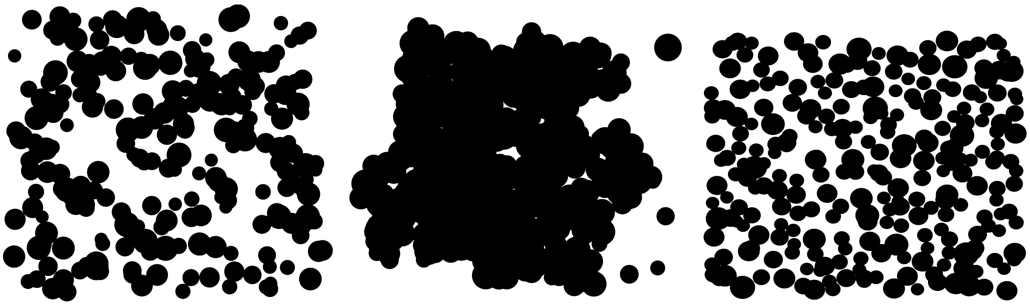


Fig. 6. *Examples of realisations of Boolean model with intensity of the disc centres 0.4 in the window* 25×25 *and 0 otherwise, and distribution of radii* $U(0.5, 1)$ *(left picture), Quermass-interaction process with parameters* $\theta_1 = 0.65$*,* $\theta_2 = -1.1$ *and* $\theta_3 = 0$ *(middle picture) and Quermass-interaction process with parameters* $\theta_1 = -1$*,* $\theta_2 = 1$ *and* $\theta_3 = 0$ *(right picture).*

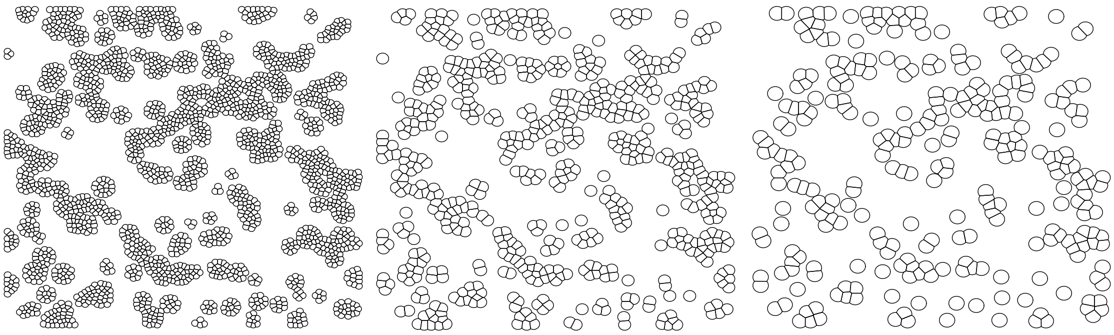


Fig. 7. *Approximations by covering using discs of* $r = 4$ *(left figure, corresponding to* $\delta = 10\%$*),* $r = 7$ *(middle figure, corresponding to* $\delta = 20\%$*) and* $r = 9$ *(right figure, corresponding to* $\delta = 30\%$*) for the planar set and its digital approximation from Fig. 3.*

realisations of the given process while we obtained the following characteristics for *PD* (in %):

	radius	*PD* mean	*PD* sd	number of realisations having *PD* < 20% when covering with the corresponding radius
Boolean	7	19.81	1.03	252 (63%)
Cluster	13	18.55	2.07	323 (81%)
Repulsive	5	19.78	0.64	284 (71%)

Due to different radii between processes Eq. 5 was applied leading to the following radii for different combinations:

	Boolean	Cluster	Repulsive
Boolean	7	7	5
Cluster		13	5
Repulsive			5

Then we applied the envelope test introduced in the *Material and Methods* section to two groups of 200 realisations for each combination obtaining appropriate p-values. More precisely, for each of the six combinations and appropriate groups, realisations were covered by discs with radii derived with Eq. 5. These coverings were translated to Voronoi tessellations. Consequently, we randomly sampled 100 independent (*i.e.*, non-neighbouring) cells of the tessellation and calculated their support functions. Finally, the envelope test as described in *the Material and Methods* section was applied to pairs of these support functions, both with and without previous alignment performed for $C = 0.9$.

Histograms of corresponding p-values are shown in Fig. 9. We can observe that comparing groups of realisations of the same processes, with or without alignment, p-values are approximately uniformly distributed which means that the test considers the cells to be identically distributed, in other words that the inner structure is the same so the sets are similar enough. On the other hand when assessing dissimilarity of different processes, p-values are mostly very close to zero, so the hypothesis of identical distribution of the cells is significantly often rejected, so we can conclude that the similarity measure is very low, *i.e.*, the original sets are not similar. The effect of alignment can be best seen when observing histogram of p-values for testing Boolean vs repulsive models, where p-values of the test have significantly lower values after performing alignment (without alignment 48.5% of the p-values were less than 0.05 while 64.5% of p-values were less than 0.05 when using alignment). Since the p-values of test are already very close to zero when comparing Boolean vs cluster and cluster vs repulsive models (all less then 0.05), the effect of alignment is not so noticeable.

DISCUSSION

Assessing dissimilarity of random sets could have big implications in applied science but is, unfortunately, hard to achieve. Especially if theoretical results are of interest due to high complexity of some random sets.

Nevertheless, some advancement can be achieved if heuristic results are considered. In this paper we have presented one of such results. We have presented a method based on covering of random sets with unions of convex compact sets and consequent application of envelope test on support functions of these approximating sets. To justify our choice we have conducted a simulation study on different types of some common models of random sets and obtained valuable results.

The presented method, however, suffers from some problems. The main weakness, in our humble opinion, is the covering of the input data set by a union of discs, especially the choice of suitable radius and pixel difference level, *i.e.*, the measure of inaccuracy of the approximation. Here we had to make several decisions based on visual comparisons and results obtained by additional simulation studies. Despite the fact that the obtained numerical results are satisfactory, this part has the potential to be improved. One of the more straightforward approaches to try to improve the current choice of free parameters is in the case were multiple realizations for processes are available. Here, a researcher could use the machine learning approach with training and test sets to derive (optimized) best parameters.

Additionally, the workflow presented in this paper can be represented through modules: cover a set, from the covering of the original set derive convex compact sets representing an approximation, derive some characteristics of the approximation and use these characteristics for the dissimilarity measure. Therefore, the workflow is suitable to changes only within modules. If a better covering algorithm can be achieved or a different test devised they can easily be used instead of the original modules.

One of the common problems in practice, that was not investigated as a special case in this paper, is the existence of edge-effects, *i.e.*, the case when the observation window contains only parts of the original set and the other parts lying outside the window are not observed (an example is presented in Fig. 3a). In our simulation study we tried to avoid this problem by ignoring the cells lying on the edge of the observation window, sampling only those that were fully contained in the window. For random sets with significant parts observed on the edges appropriate adjustments should be devised. We have left this for future work.

Also, in future work additional ways for testing equality of distributions of convex compact approximations might be applied and further simulation studies of the current model would be beneficial to better understand properties of the introduced method. This will, hopefully, lead us to suggest suitable modifications or convenient alternatives to the current method.

ACKNOWLEDGEMENTS

This work was supported by The Czech Science Foundation (GAČR), project No.13-05466P.

REFERENCES

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

Diggle PJ (1981). Binary mosaics and the spatial pattern of heather. Biometrics 37:531–9.

Ebeida MS, Davidson AA, Patney A, Knupp PM, Mitchell SA, Owens JD (2011). Efficient maximal Poisson-disk sampling. In: ACM SIGGRAPH 2011 Papers. New York: ACM. pp. 49:1–49:12.

Hall P (1985). Counting methods for inference in binary mosaics. Biometrics 41:1049–52.

Helisová K (2014). Modeling, statistical analyses and simulations of random items and behavior on material

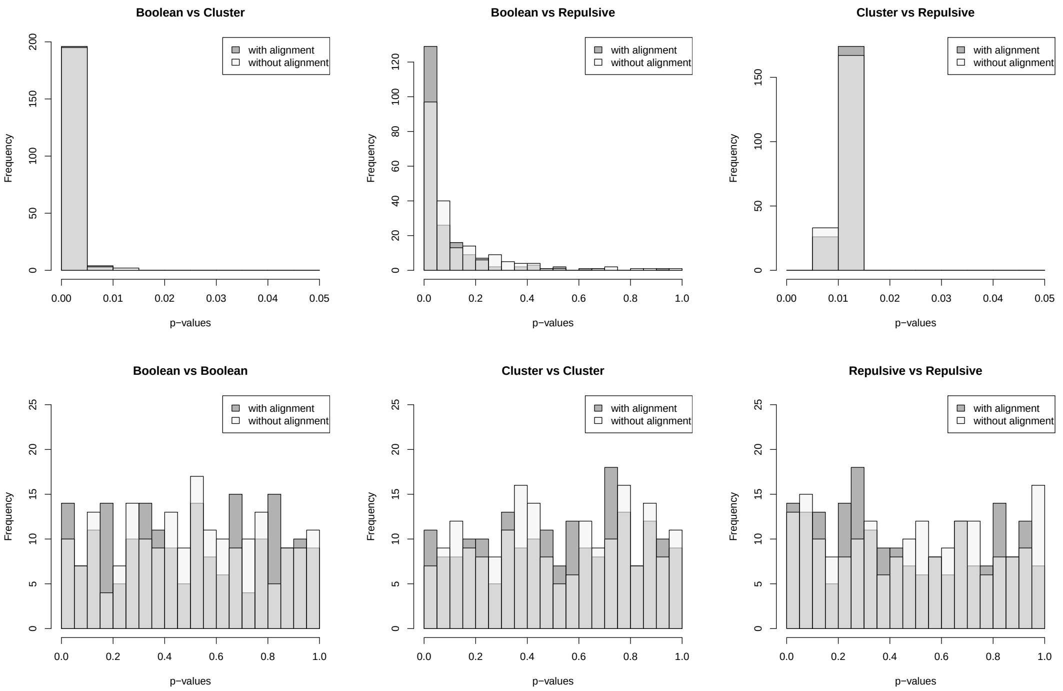


Fig. 9. *Histograms of p-values for testing all pairs of simulated processes, i.e., Boolean-cluster (upper left), Boolean-repulsive (upper middle), cluster-repulsive (upper right), Boolean-Boolean (lower left), cluster-cluster (lower middle) and repulsive-repulsive (lower right); 200 realisations of each process were used.*

surfaces. In: TMS 2014 Suppl Proc. Hoboken: John Wiley & Sons. pp. 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 Med 34:2636–61.

Imai H, Iri M, Murota K (1985). Voronoi diagram in the laguerre geometry and its applications. SIAM Comp 14:93–105.

James G, Witten D, Hastie T, Tibshirani R (2013). An introduction to statistical learning. New York: Springer.

Kendall WS, Van Lieshout MNM, Baddeley AJ (1999). Quermass-interaction processes: Conditions for stability. Adv Appl Probab 31:315–42.

Lavie M (2000). Characteristic function for random sets and convergence of sums of independent random sets. Acta Math Viet 25:87–99.

Mrkvička T, Mattfeldt T (2011). Testing histological images of mammary tissues on compatibility with the Boolean model of random sets. Image Anal Stereol 30:11–8.

Mrkvička T, Myllymäki M, Hahn U (2015). Multiple Monte Carlo testing with applications in spatial point processes. arXiv: 1506.01646.

Myllymäki M, Mrkvička T, Grabarnik P, Seijo H, Hahn U (2016). Global envelope tests for spatial processes. J Roy Stat Soc B (in press). doi: 10.1111/rssb.12172

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.

Ohser J, Mücklich F (2000). Statistical analysis of microstructures in materials science. Chichester: John Wiley & Sons.

Pratt WK (2001). Digital image processing: PIKS Inside. 3rd Ed. New York: John Wiley & Sons.