Image Anal Stereol 2021;40:127-140 doi: 10.5566/ias.2600
Original Research Paper

TWO-STEP METHOD FOR ASSESSING SIMILARITY OF RANDOM SETS

VESNA GOTOVAC ĐOGAŠ[1], KATEŘINA HELISOVÁ✉,[2], BOGDAN RADOVIĆ[2], JAKUB STANĚK[3], MARKÉTA ZIKMUNDOVÁ[4] AND KATEŘINA BREJCHOVÁ[2]

[1]University of Split, Croatia, [2]Czech Technical University in Prague, Czech Republic, [3]Charles University, Czech Republic, [4]University of Chemistry and Technology, Czech Republic
e-mail: vgotovac@pmfst.hr, heliskat@fel.cvut.cz, radovbog@fel.cvut.cz, stanekj@karlin.mff.cuni.cz, zikmundm@vscht.cz, brejcka1@fel.cvut.cz
(Received July 9, 2021; revised November 5, 2021; accepted November 7, 2021)

ABSTRACT

The paper concerns a new statistical method for assessing dissimilarity of two random sets based on one realisation of each of them. The method focuses on shapes of the components of the random sets, namely on the curvature of their boundaries together with the ratios of their perimeters and areas. Theoretical background is introduced and then, the method is described, justified by a simulation study and applied to real data of two different types of tissue - mammary cancer and mastopathy.

Keywords: connected component, curvature, similarity, *N*-distance, random set.

INTRODUCTION

In the last years, modelling and statistical analyses of random sets have become very popular. It has been studied both from theoretical point of view (Matheron, 1975; Molchanov, 2005; Serra, 1982) as well as from the practical side, because they have many applications in biology (Moeller and Helisová, 2010), medicine (Hermann et al., 2015), material sciences (Neumann et al., 2016) and other branches. They can describe and explain many events, for example behaviour of cells in organisms, particles in materials, presence of different plants etc. Therefore, mathematical methods dealing with random sets must be developed and improved.

Usually, when we are given a realisation of a random set, we try to find, based on the realisation, a model in order to make further statistical analyses. However, there are situations when the knowledge about the concrete model is not necessary, because the aim is to decide whether two realisations are similar in some sense, i.e. whether they may come from the same process, e.g. we need only to distinguish between two types of cells in tissue from microscopic pictures, recognise different tendency of growth of some plants, detect defects in materials etc.

In the presented paper, we focus on planar random sets, nevertheless, the method described here can be easily extended to more dimensional spaces. Although there exist classical tools for comparing random sets like covariance function or contact distribution function (Chiu et al., 2013), functions on morphological operations, namely dilation, erosion, opening and closing (Serra, 1982), etc., there are situations when these characteristics are not sufficient to distinguish between two realisations. E.g. we can obtain the same estimate of contact distribution function for different shapes or, on the other hand, two realisations consisting of components of the same shapes but different mutual distances have very different estimate of the covariance function etc., so they cannot be used when the main objects of interest are the shapes of the components, as in this paper. Another disadvantage of the mentioned characteristics is that for one realisation, we obtain only one function. Then, it is difficult to formulate the task of comparing two random sets when we have only one realisation of each of them. In this case, it would be helpful to have data consisting of more functions for each realisation.

New methods taking this into account have been developed in the last five years (Debayle et al., 2021; Gotovac, 2019; Gotovac Đogaš and Helisová, 2021; Gotovac et al., 2016). In Debayle et al. (2021), morphological skeletons (Serra, 1982) of compared realisations are constructed and then, similarity of two realisations is defined through a function describing mass growth around selected points of the skeletons. This method shows the highest power in simulation study, compared to the methods in the remaining three papers. However, it is closely tied to the placement of components in realisations, which is not always desirable. Further approach can be found in Gotovac (2019), where the author also focuses on similarity of shapes and positions of components in realisations, but the positions can be omitted in special cases. The considered components are either the connected components or some more specific set components obtained by further decomposing of the observed set (e.g. cells in a tissue which are

connected but they can be determined in their binary image). The positions of the components are described by the so-called neighbourhood tessellation of the observation window constructed from the components and by using the Hausdorff metric (Chiu et al., 2013). The samples of pairs of the components and neighbouring tessellation cells are compared using the test based on the $\mathscr{N}$-distance (Klebanov, 2006), where a suitable kernel involving the symmetric difference of the sets is constructed. For omitting dependence on locations, the components can be considered without their neighbourhoods. A disadvantage of the method is that the test of similarity in this sense is weak, especially when the neighbourhoods are omitted, so in order to achieve better accuracy, the positions of the components must be taken into account again. An approach independent of the positions of the components is introduced in Gotovac et al. (2016) and improved in Gotovac Đogaš and Helisová (2021). The authors of the papers distinguish between two realisations via a heuristic approach based on approximation of the components by unions of convex compact sets using Voronoi tessellation (Chiu et al., 2013) with respect to a special hard-core point process on the realisations and consequent comparison of the support functions of the convex compact cells of the tessellation using the envelope test from Myllymäki et al. (2017) and the test based on $\mathscr{N}$-distance (sometimes also called the kernel test (Gretton et al., 2012)), respectively. The authors in Gotovac Đogaš and Helisová (2021) and Gotovac et al. (2016) declare that the method focuses on the structure of the components like clustering or repulsion tendencies, generating rounded or angular objects, long and thin or short and thick formations, etc., because it recognises how much mass is concentrated on the boundaries of the components and what is the approximate shape of the boundary. The results of simulation are satisfactory, however, there is weakness in the heuristic procedure because the approximation of the components is not unique, but random and there are quite large differences between the shapes of the original realisations and their approximations. Moreover, some free parameters must be chosen, which can affect the results.

In the present paper, we focus on testing similarity of two realisations of random sets, where the similarity is given by similar shapes of their components. However, instead of the heuristic approximation of realisations, we use description by uniquely defined characteristics of components in the realisations, namely the ratio of the perimeter and the area of each component, and the curvature (Bullard et al., 1995) of the boundary of each component.

The paper is organised as follows. The section "Methods" introduces basic terms and approaches, while the subsection "Theoretical background" summarises definitions and already existing theoretical results concerning curvatures of planar curves and statistical testing via the $\mathscr{N}$-distance theory, and in the subsection "Methodology", we present our new matters, namely we define similarity of two random sets and describe the procedure of assessing similarity from two realisations. This is the main result of the paper. In the section "Results", the procedure is first justified by a simulation study, and then applied to real histological data. The section "Discussion" is dedicated to comparison of the new results to the results obtained by the previous methods.

METHODS

THEORETICAL BACKGROUND

Curvature of a planar curve

The definition and claims in this section can be found in Bullard et al. (1995).

Definition 2.1 *Consider a smooth 2D curve c parameterised by a parameter $\varphi \in [0,\phi] \subset \mathbb{R}$, i.e. $c(\varphi) = (x(\varphi), y(\varphi))$. Then the curvature κ of c is defined as*

$$\kappa(c(\varphi)) = \frac{x'(\varphi)y''(\varphi) - x''(\varphi)y'(\varphi)}{(x'^2(\varphi) + y'^2(\varphi))^{3/2}}.$$

It means that $\kappa(c(\varphi)) = \pm 1/R(\varphi)$, where $R(\varphi)$ is the radius of the osculating circle touching the curve in the point $[x(\varphi), y(\varphi)]$ and the choice between "+" and "−" is determined by the local convexity convention.

Let us assume that the curve c is continuous, closed (i.e. $c(0) = c(\phi)$) and it does not intersect itself (i.e. $c(\varphi_1) = c(\varphi_2) \Rightarrow \varphi_1 = \varphi_2$). Consider a (connected) planar set X whose boundary is given by the curve c. It can be shown that for the curvature $\kappa(z)$ evaluated in a given point $z \in c$ and for a disc $b(z,r)$ with the center in z and a radius r small enough, it holds that

$$\kappa(z) \approx \frac{3A^*_{b(z,r)}}{r^3} - \frac{3\pi}{2r} = \frac{3\pi}{r}\left(\frac{A^*_{b(z,r)}}{A_{b(z,r)}} - \frac{1}{2}\right), \qquad (1)$$

where $A_{b(z,r)}$ is the area of the disc $b(z,r)$ and $A^*_{b(z,r)}$ is the area of $b(z,r) \cap X$.

This is used below in the section "Methodology" when estimating the curvature of boundary of binary

image of the set X. The center of discretised disc b with the radius r pixels is placed to the boundary pixel in which we want to evaluate the curvature, and approximate the ratio $A^*_{b(z,r)}/A_{b(z,r)}$ by the number of pixels of the disc b inside the set X divided by the number of all pixels forming the disc b.

Testing equality in distribution based on $\mathscr{N}$-distance of probability measures

In this paper, the procedure of testing equality in distribution of random variables and random functions comes from the theory of $\mathscr{N}$-distances, which is briefly recalled in the following paragraphs. More details concerning this topic can be found in Klebanov (2006).

Let $\mathscr{X}$ be a nonempty set. Consider a negative definite kernel $\mathscr{L} : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$.

Definition 2.2 *The negative definite kernel $\mathscr{L}$ is called* strongly negative definite kernel *if for an arbitrary probability measure μ and an arbitrary $f : \mathscr{X} \to \mathbb{R}$ such that $\int_{\mathscr{X}} f(x)d\mu(x) = 0$ holds and $\int_{\mathscr{X}}\int_{\mathscr{X}} \mathscr{L}(x,y)f(x)f(y)d\mu(x)d\mu(y)$ exists and is finite, the relation*

$$\int_{\mathscr{X}}\int_{\mathscr{X}} \mathscr{L}(x,y)f(x)f(y)d\mu(x)d\mu(y) = 0$$

implies that $f(x) = 0$ μ-a.e.

For a map $\mathscr{L} : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$, denote $\mathscr{B}_{\mathscr{L}}$ the set of all measures μ such that $\int_{\mathscr{X}}\int_{\mathscr{X}} \mathscr{L}(x,y)d\mu(x)d\mu(y)$ exists.

Theorem 2.1 (Klebanov, 2006) *Let $\mathscr{L}(x,y) = \mathscr{L}(y,x)$. Then*

$$\begin{aligned}\mathscr{N}(\mu,\nu) =& 2\int_{\mathscr{X}}\int_{\mathscr{X}} \mathscr{L}(x,y)d\mu(x)d\nu(y) \\ &- \int_{\mathscr{X}}\int_{\mathscr{X}} \mathscr{L}(x,y)d\mu(x)d\mu(y) \\ &- \int_{\mathscr{X}}\int_{\mathscr{X}} \mathscr{L}(x,y)d\nu(x)d\nu(y) \geq 0\end{aligned} \tag{2}$$

holds for all measures $\mu,\nu \in \mathscr{B}_{\mathscr{L}}$ with equality in the case $\mu = \nu$ only, if and only if $\mathscr{L}$ is a strongly negative definite kernel.

In the following text, the term $\mathscr{N}(\mu,\nu)$ from Eq. 2 is called the $\mathscr{N}$-distance of the measures μ and ν. The approach to testing equality of distributions given by the (probability) measures μ and ν is described below.

First, let the measures μ and ν correspond to distributions of real random variables. Suppose we have observations $x_1,\dots,x_{m_1} \in \mathbb{R}$ from the distribution μ and $y_1,\dots,y_{m_2} \in \mathbb{R}$ from the distribution ν. The $\mathscr{N}$-distance of the measures μ and ν is estimated as

$$\begin{aligned}\hat{\mathscr{N}}_1 =& \frac{2}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \mathscr{L}(x_i,y_j) \\ &- \frac{1}{m_1^2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_1} \mathscr{L}(x_i,x_j) - \frac{1}{m_2^2}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \mathscr{L}(y_i,y_j),\end{aligned} \tag{3}$$

where we use the Euclidean distance as the negative definite kernel $\mathscr{L}$. The value $\hat{\mathscr{N}}_1$ plays the role of test statistic. Then, we use Monte Carlo permutation test, i.e. we make s permutations of all observed values $x_1,\dots,x_{m_1},y_1,\dots y_{m_2}$, split each permutation into two groups of the lengths m_1 and m_2, and, analogously to Eq. **??**, we calculate $\hat{\mathscr{N}_i}$ for the i-th permutation, $i = 2,\dots,s+1$. Then the p-value of the test is

$$p = \frac{\sharp\{i \in \{2,\dots,s+1\} : \hat{\mathscr{N}}_i \geq \hat{\mathscr{N}}_1\} + 1}{s+1}. \tag{4}$$

When the measures μ and ν correspond to distributions of random functions, then testing the equality of μ and ν runs analogously as above, only with the difference in the choice of the negative definite kernel $\mathscr{L}$. Here, we use the kernel introduced in Gotovac Đogaš and Helisová (2021), constructed especially for random functions as follows. We evaluate testing functions $t^{(1)}$ and $t^{(2)}$ in discrete arguments $u_1,\dots,u_n$, $n \in \mathbb{N}$. Then the negative definite kernel is

$$\begin{aligned}&\mathscr{L}(t^{(1)},t^{(2)}) = \\ &= \sum_{m=1}^{D}\sum_{\{k_1,\dots,k_m\}\subseteq\{1,\dots,n\}} \left(\sum_{l=1}^{m}\left(t^{(1)}(u_{k_l}) - t^{(2)}(u_{k_l})\right)^2\right)^{1/2},\end{aligned} \tag{5}$$

where D is a chosen constant specifying the depth of dependence (more precisely, it allows testing the equality of finite-dimensional distributions of random functions $t^{(1)}$ and $t^{(2)}$ for the dimensions less than or equal to D). The estimate of the $\mathscr{N}$-distance of the functions $t^{(1)}$ and $t^{(2)}$ based on the random samples $t_i^{(1)}$, $i = 1,\dots,m_1$, and $t_j^{(2)}$, $j = 1,\dots,m_2$, respectively, is evaluated as

$$\begin{aligned}\hat{\mathscr{N}}_1 =& \frac{2}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \mathscr{L}(t_i^{(1)},t_j^{(2)}) \\ &- \frac{1}{m_1^2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_1} \mathscr{L}(t_i^{(1)},t_j^{(1)}) \\ &- \frac{1}{m_2^2}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \mathscr{L}(t_i^{(2)},t_j^{(2)}).\end{aligned} \tag{6}$$

Then, we again apply the Monte Carlo permutation test, i.e. we make s permutations of all functions $t_1^{(1)}(u),\dots,t_{m_1}^{(1)}(u),t_1^{(2)}(u),\dots t_{m_2}^{(2)}(u)$ in order to obtain $\hat{\mathscr{N}_i}$, $i=2,\dots,s+1$, and evaluate the p-value as described above by Eq. 4.

METHODOLOGY

Testing characteristics

In this paper, we define the similarity of random sets through their components, namely through the curvature of their boundaries and the ratios of their perimeters and areas.

Consider a connected random set $\mathbf{X}$, i.e. the random set whose realisations are connected. Denote $B_{\mathbf{X}}$ the boundary of $\mathbf{X}$ and $\kappa_{\mathbf{X}}(z)$ the (random) curvature in the point $z\in B_{\mathbf{X}}$. From Eq. 1, we can see that for a disc $b(z,r)$ with suitable chosen radius r,

$$\kappa_{\mathbf{X}}(z)\propto\frac{A^*_{b(z,r),\mathbf{X}}}{A_{b(z,r)}},$$

where $A_{b(z,r)}$ is the area of the disc $b(z,r)$ and $A^*_{b(z,r),\mathbf{X}}$ is the area of $b(z,r)\cap\mathbf{X}$. Therefore, we focus only on the ratio of these two areas. Denote

$$O_{\mathbf{X},b(z,r)}=\frac{A^*_{b(z,r),\mathbf{X}}}{A_{b(z,r)}}$$

and define the function

$$\tilde{\kappa}_{\mathbf{X},r}(u)=|B_{\mathbf{X}}|^{-1}\int_{B_{\mathbf{X}}}\mathbf{1}\{O_{\mathbf{X},b(z,r)}\le u\}dz,\quad u\in[0,1],$$

which is basically an analogy of the distribution function of the curvature at points on the boundary, but it is evaluated for all boundary points, so it describes the distribution for strongly dependent values. The object of our interest is the function, analogous to density function, describing the distribution of the curvature along the boundary, i.e.

$$t_{\mathbf{X},r}(u)=\tilde{\kappa}'_{\mathbf{X},r}(u). \tag{7}$$

Finally, denote $R_{\mathbf{X}}$ the random variable describing the ratio of the perimeter and the area of the random set $\mathbf{X}$.

Definition 2.3 *Two connected random sets* $\mathbf{X}$ *and* $\mathbf{Y}$ *are considered to be similar if the distributions of* $\lim_{r\to0}t_{\mathbf{X},r}$ *and* $\lim_{r\to0}t_{\mathbf{Y},r}$ *as well as the distributions of* $R_{\mathbf{X}}$ *and* $R_{\mathbf{Y}}$ *are equal.*

In practice, we usually observe realisations X and Y of the random sets $\mathbf{X}$ and $\mathbf{Y}$, respectively, in the form of binary images, so we need to adjust the task of assessing dissimilarity of the realisations consisting of black and white pixels. The pixels play the role of units in the following sections. The ratio of the perimeter and the area is then simply given by the number of boundary pixels divided by the number of all pixels of the component. For evaluating of the function describing the curvature, fix a radius $r\in\mathbb{N}$, denote P the set of all pixels of the binary image X, $z_1,\dots,z_n$ all boundary pixels, and for each boundary pixel z_i, define

$$K(z_i)=\frac{\sharp\{p\in P:p\in b(z_i,r)\cap X\}}{\sharp\{p\in P:p\in b(z_i,r)\}}.$$

Then, the approximation of the function $t_{\mathbf{X},r}(u)$ from Eq. 7 is

$$t(u)=\frac{\sharp\{i\in\{1,\dots,n\}:K(z_i)\in[u-1/l,u)\}}{n} \tag{8}$$

for $u=\frac{1}{l},\frac{2}{l},\dots,1$, which plays the role of testing function.

Testing similarity of connected random sets

Consider two samples, namely $X_1,\dots,X_{m_1}$ and $Y_1,\dots,Y_{m_2}$, of realisations of connected random sets $\mathbf{X}$ and $\mathbf{Y}$, respectively. We want to test the null hypothesis that $\mathbf{X}$ and $\mathbf{Y}$ are similar. First we evaluate the ratios $R_{X_1},\dots,R_{X_{m_1}},R_{Y_1},\dots,R_{Y_{m_2}}$ of the perimeters and areas of the corresponding realisations. Based on these values, we estimate the $\mathscr{N}$-distance of the ratios of $\mathbf{X}$ and $\mathbf{Y}$ by Eq. **??**, where we set $x_i=R_{X_i}$, $i=1,\dots,m_1$ and $y_j=R_{X_j}$, $j=1,\dots,m_2$. Let us denote it $\hat{\mathscr{N}}_1^R$. Then, we evaluate the testing functions $t(u)$ from Eq. 8, which describe the boundary curvatures $t_{X_1}(u),\dots,t_{X_{m_1}}(u),t_{Y_1}(u),\dots,t_{Y_{m_2}}(u)$, calculate the $\mathscr{N}$-distance of the functions corresponding to $\mathbf{X}$ and $\mathbf{Y}$, respectively, using Eq. 6 and Eq. **??**, and denote this $\mathscr{N}$-distance as $\hat{\mathscr{N}}_1^t$. The couple $(\hat{\mathscr{N}}_1^R,\hat{\mathscr{N}}_1^t)$ is the test statistic. Here, we use the Monte Carlo permutation test described above, i.e. we make s permutations of all realisations $X_1,\dots,X_{m_1}$ and $Y_1,\dots,Y_{m_2}$, and split them into two groups of the sizes m_1 and m_2, respectively, in order to obtain $(\hat{\mathscr{N}}_i^R,\hat{\mathscr{N}}_i^t)$, $i=2,\dots,s+1$, and evaluate the p-value as

$$p=\frac{\sharp\{i\in\{2,\dots,s+1\}:\hat{\mathscr{N}}_i^R\ge\hat{\mathscr{N}}_1^R\wedge\hat{\mathscr{N}}_i^t\ge\hat{\mathscr{N}}_1^t\}+1}{s+1}. \tag{9}$$

Similarity of random sets consisting of more components

Usually in practice, we have the data in the form of realisations consisting of more than one component. If we can suppose that the components are independent and come from the same distribution, then we can define similarity of two random sets it the way that they are considered to be similar, if their components are similar in the meaning of Definition 2.3. It is used in simulation study below in the section "Simulation study".

Nevertheless, the independence of the components can be supposed in very specific cases, e.g. in germ-grain models (Chiu et al., 2013) in which the intensity of germs is low with respect to the volume of grains. However, the components are usually dependent. In order to avoid this complication, we make suitable random samples of the components in each realisation. The size of such samples is discussed below in the section "Simulation study". Just note that since we use permutation version of the test, the condition of independence of the components can be weakened to their exchangeability.

In this way, we obtain two samples of components which are then used as the input samples $X_1, \dots, X_{m_1}$ and $Y_1, \dots, Y_{m_2}$ from the section "Testing similarity of connected random sets".

RESULTS

SIMULATION STUDY

In the simulation study, we first focus on four models which illustrate the usage of the procedure. The models and approach to simulation can be found in Debayle et al. (2021), Gotovac (2019), Gotovac Đogaš and Helisová (2021) and Gotovac et al. (2016). For all considered models, we simulate 200 realisations and compare 100 vs 100 realisations of the same models as well as 100 vs 100 realisations of different models. Since the outputs of the tests are the p-values, we obtain 100 p-values for each couple. Some of their histograms are shown and commented below. Note that p-value close to zero means that the equality of distributions of the corresponding testing functions is rejected. Thus, the p-value should be concentrated close to zero when comparing realisations of different models, while it should be uniformly distributed in the interval $[0,1]$ when comparing realisations of the same models.

In this simulation study, we moreover have to choose the radius of the disc used to estimate the curvature. Briefly said, too large disc does not detect localised changes in curvature in the sense that it can capture more than one interface, on the other hand, too small disc has large error due to discretisation. All realisation images in our simulation study are in resolution of 400×400 pixels (units). Some recommendations on how to choose a suitable radius can be found in Bullard et al. (1995). Based on these conditions and personal consultation with Matěj Lébl (Institute of Information Theory and Automation, Czech Academy of Science), we use the radii $r = 3$ and $r = 5$. Note that the results of the simulation study are very similar for both radii. The presented histograms show the p-values of the tests using $r = 5$ (with one exception mentioned below).

Examples of realisations of the illustrating models are shown in Fig. 1. The first picture is a realisation of the random disc Boolean model used in previous studies. The second realisation is simulated so that in (another) realisation of the Boolean model, each connected component is deleted with probability 1/2. It is called the reduced Boolean model in the following paragraphs. The third realisation is formed by disjoint squares whose ratio of the perimeter and the area comes from the same distribution as the ratio for the Boolean model (namely from the empirical distribution obtained from 100 realisations of the Boolean model). We call it the square model in the following paragraphs. The fourth realisation is simulated as the process of disjoint rectangles with the same distribution of perimeters as the square perimeters, while one side has fixed length of 4 pixels (note that in this case, we use the disc with radius $r = 3$ only for estimating the boundary curvature). It is called the rectangle model in the following paragraphs.

We want to show that the method does not distinguish between the Boolean model and reduced Boolean model since it is based only on the similarity of the components, but it distinguishes between the Boolean models and the square model due to the boundary curvature, as well as between the square model and the rectangle model due to the ratios of the perimeter and the area of the components. Indeed, we can see it in the first column of Fig. 2. The histogram of p-values shows approximately uniform distribution when comparing the Boolean model and reduced Boolean model, but the p-values are very close to zero when testing the Boolean model vs the square model and the square model vs the rectangle model. Moreover, we test the equality in distribution of the curvature functions and of the ratios of perimeters and areas separately. The histograms of the p-values of the test for the ratios are shown in the second column of Fig. 2, and the

Fig. 1. *Example of realisation of the Boolean, the reduced Boolean, the square and the rectangle model, respectively.*

p-values of the test for the curvature functions are shown in the third column. It is natural that the p-values are approximately uniformly distributed for both tests of the Boolean model vs reduced Boolean model. The meaning of the procedure is clear from other histograms. There, we can observe the equal distributions of the ratios of perimeters and areas, but clearly different distributions of the curvature functions when comparing the Boolean model and the square model, and conversely the agreement of the distribution of the curvature functions and different distributions of the ratios of perimeters and areas when testing the square model vs the rectangle model. Thus, when we want to distinguish between realisations with differently shaped components, both characteristics must be taken into account.

Next thing we observe in our simulation study is that when we have all components of the realisation in the sample, the p-values are greater than we expect. It is seen when comparing the same models. The histograms of p-values are located to the right, while they become more uniform when we sample less components for testing, see Fig. 3. It is the effect of dependence of components in each realisation. In realisations with densely placed components, the shape of one component affects the shape of another one, so they form something like a puzzle. Such sets of components are then more similar then sets formed by independent components. From histograms in Fig. 3, we conclude that in our case, we can take a sample of 10 components from realisation of Boolean model to eliminate the effect of dependence of the components.

Further, we consider models compared in the previous papers. Except the Boolean model mentioned above, which appears in all the mentioned publications, we consider a model of partially repulsive particles (called the repulsive model in the following paragraphs and a model of particles forming clusters (called the cluster model in the following paragraphs). Realisations of both these models are simulated as realisations of the random disc Quermass-interaction process (Moeller and Helisová, 2008) with suitable chosen parameters. More details about the parameters can be found in Gotovac et al. (2016) and Gotovac (2019). Note that the same repulsive model is employed for simulation studies in all above mentioned papers, similarly as the Boolean model, while the same cluster model is used only in Gotovac (2019). In Debayle et al. (2021), Gotovac Đogaš and Helisová (2021) and Gotovac et al. (2016), another cluster model is considered, which is not suitable for our current study since its realisations consist of too few components. The fourth model is the Boolean model with grains to be ellipses (called the ellipse model in the following paragraphs). It appears, similarly as the cluster model, only in Gotovac (2019), because in the other papers, its application would not be interesting. Examples of realisations of the four models are shown in Fig. 4.

First, we test the similarity of the same models. For the Boolean model, we can see in Fig. 3 that the p-values are approximately uniformly distributed for samples of the size between 10 and 20 components. Such a large sample can be viewed as a sample of weakly dependent components. Therefore, we make samples of 10 and 20 components from each realisation of the remaining models. The histograms in Fig. 5 shows that the p-values are uniformly distributed when testing the samples of 20 components for the repulsive model and for the ellipse model, while for the cluster model, the sample is not rare enough, it has uniformly distributed p-values for the samples of 10 components.

Based on this observation, we use the samples of 10 components for testing similarity of different models. Their histograms are shown in Fig. 6. We can see that the p-values are more or less close to zero, but the rejection of the similarity hypothesis is not very convincing.

We assume that this is due to the small number of components in the test sample, but we cannot create a larger sample from our simulated realisations because of the interdependence of the components.

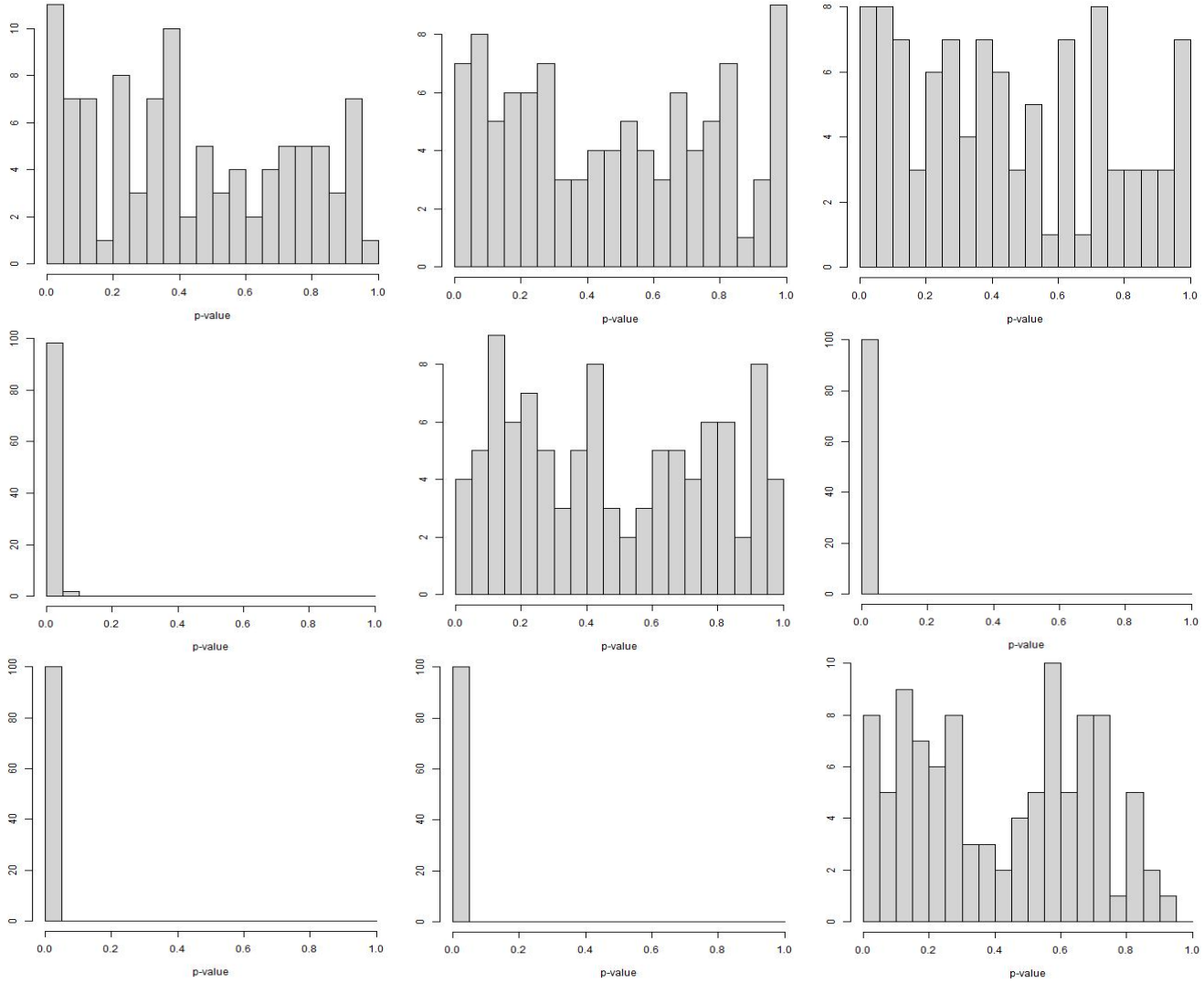

Fig. 2. *Histograms of p-values when testing the Boolean model vs the reduced Boolean model (the first row), the Boolean model vs the square model (the second row) and the square morel vs the rectangle model using both characteristics, i.e. the curvatures of the boundary and the ratios of the area and perimeter of the components (the first colomn), only the ratios (the second columns) and only the curvatures (the third columns).*

Therefore, we try to apply the boot strap method. We mix all the components from each model together and, when testing the similarity of the two models, we randomly select 100 components of one model and 100 components of the second one and calculate the p-value of the similarity test. We repeat this approach one hundred times to get 100 p-values for construction of histogram. The histograms are shown in Fig. 7. We can see that except comparing the repulsive model and the cluster model, almost all p-values are less than 0.05 now. The reason for larger p-values in the case of the repulsive model and the cluster model is the fact that many components in these models are formed by isolated discs that come from the same distribution.

APPLICATION TO REAL DATA

Finally, we apply the procedure to real data kindly provided by the authors of Mrkvička and Mattfeld (2011). We work with binary images of two different types of mammary tissue, namely 8 images of mastopathy tissue and 8 images of mammary cancer, see Fig. 8 and 9. 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 same images are analysed in Gotovac (2019).

The images are in resolution of 512×5120 pixels. Each image is made by pasting together ten square pictures in the resolution 512×512 pixels, which have been provided to us. Since this resolution is similar to the resolution of the realisations used for

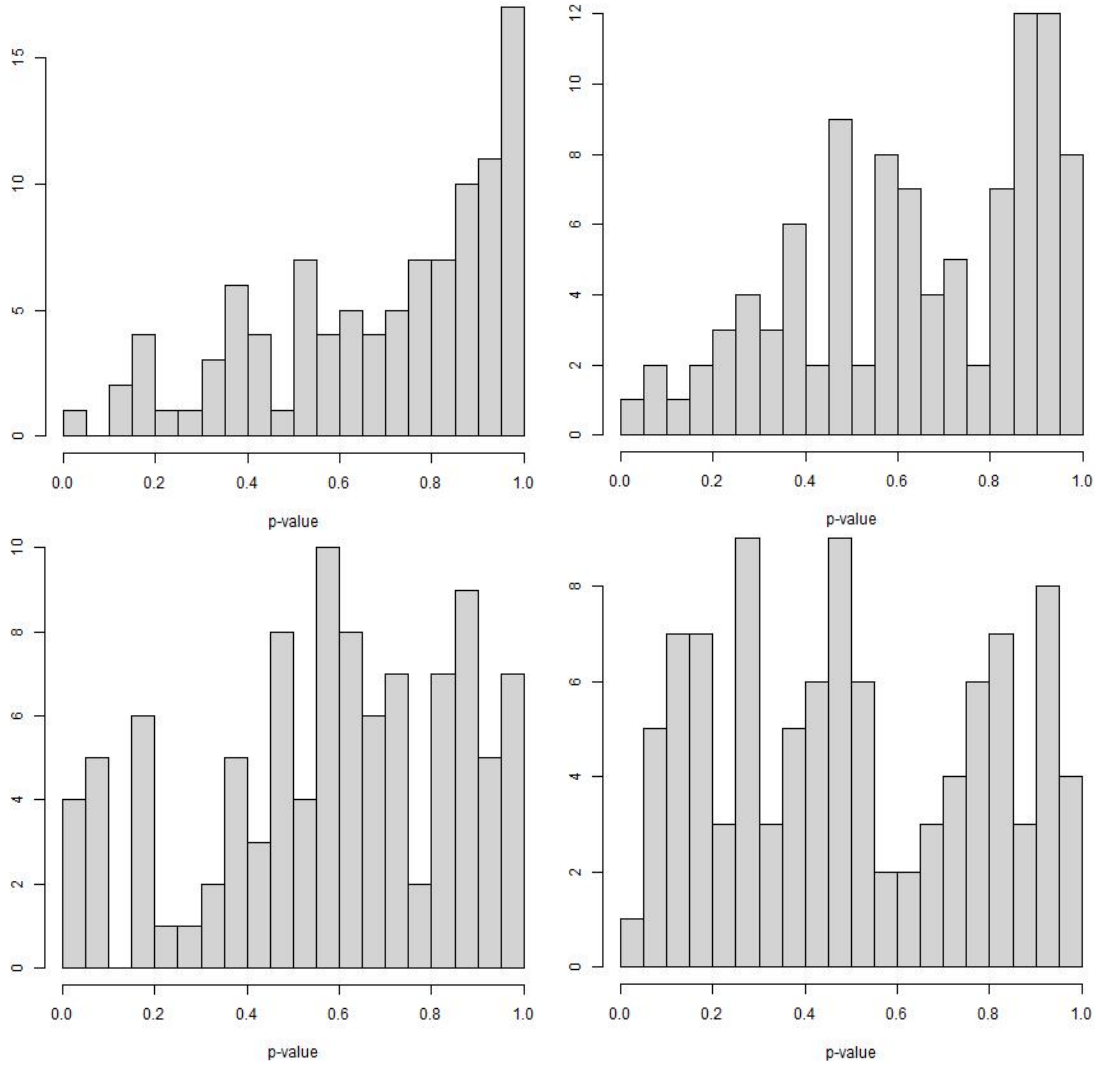

Fig. 3. *Histograms of p-values when testing the Boolean model vs the Boolean model using 50 (upper left), 30 (upper right), 20 (lower left) and 10 (lower right) components from each realisation.*

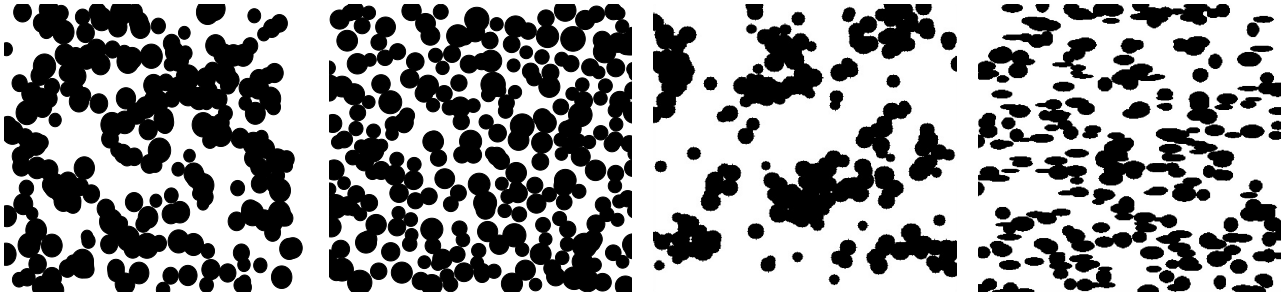

Fig. 4. *Example of realisation of the Boolean, the repulsive, the cluster and the ellipse model, respectively.*

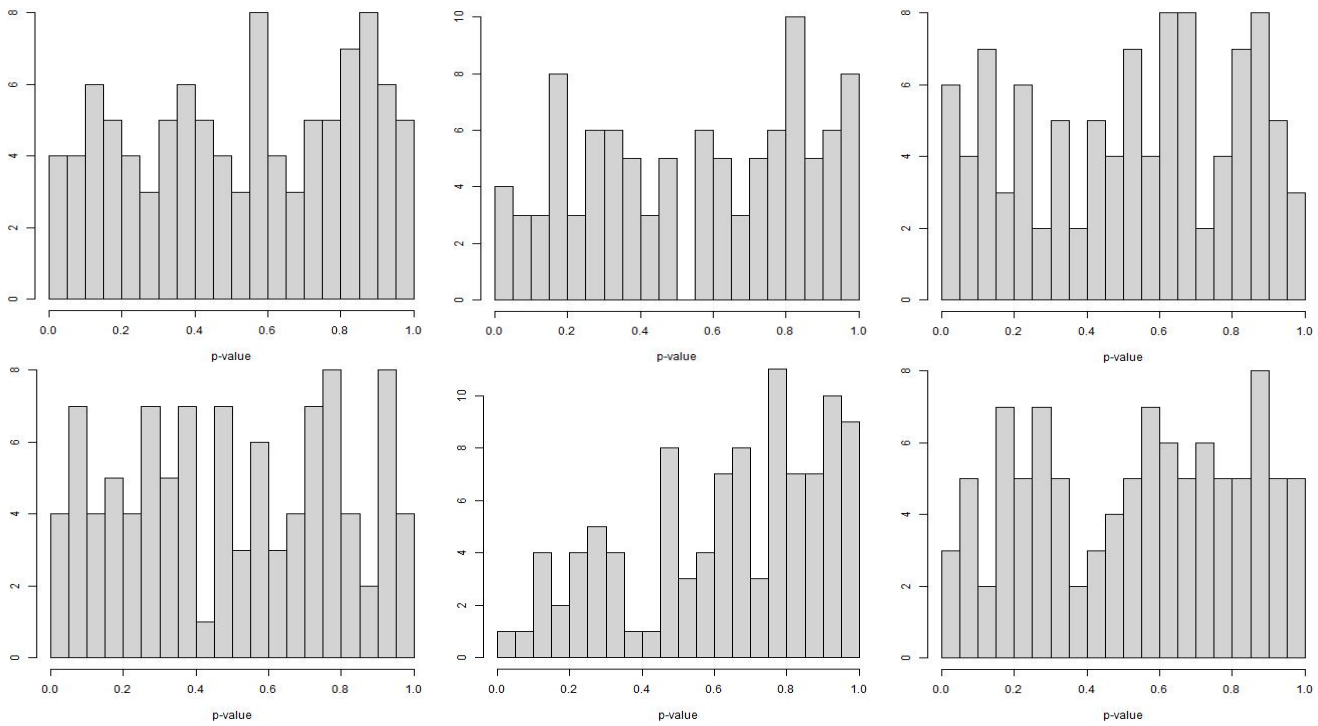

Fig. 5. *Histograms of p-values when testing similarity of the same models, namely the repulsive models (left), the cluster models (middle) and the ellipse models (right) using the samples of 10 components (upper row) and 20 components (lower row).*

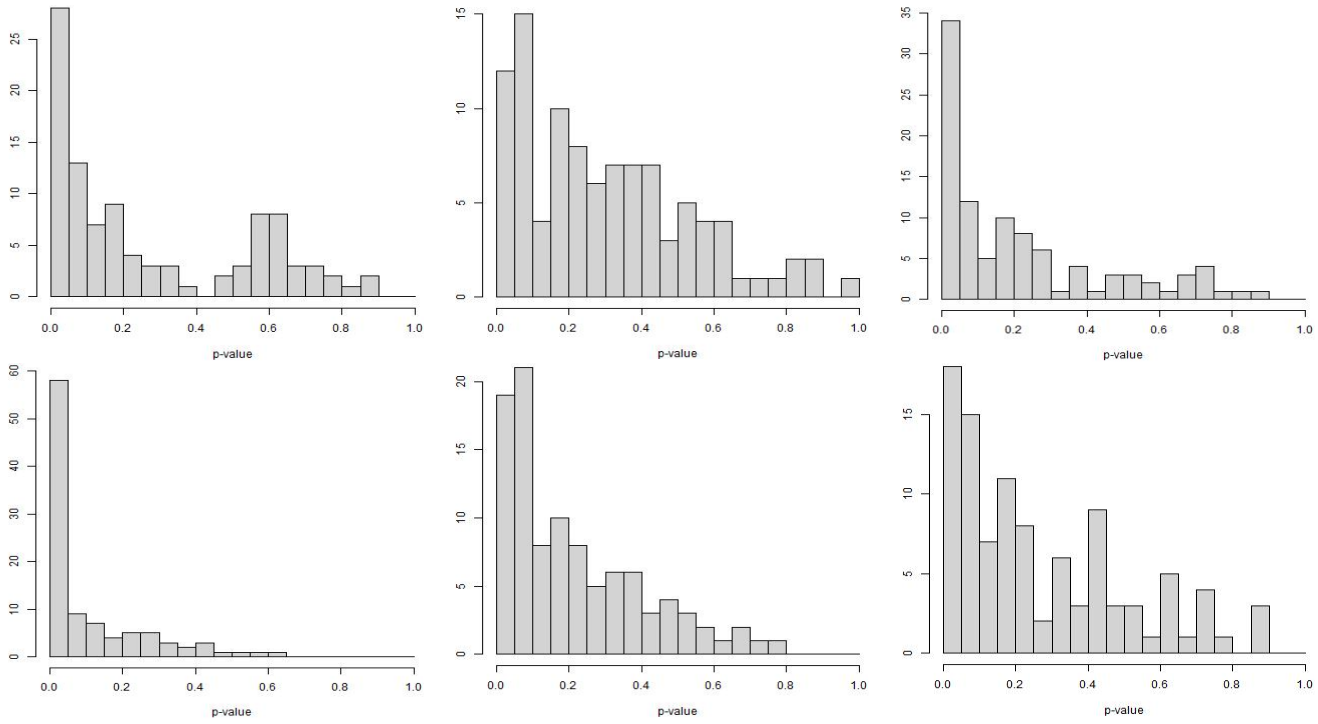

Fig. 6. *Histograms of p-values when testing similarity of the Boolean model vs the repulsive model (upper left), the Boolean model vs the cluster model (upper middle), the repulsive model vs the cluster model (upper right), the ellipse model vs the Boolean model (lower left), the ellipse model vs repulsive model (lower middle) and the ellipse model vs the cluster model (lower right) using the samples of 10 components.*

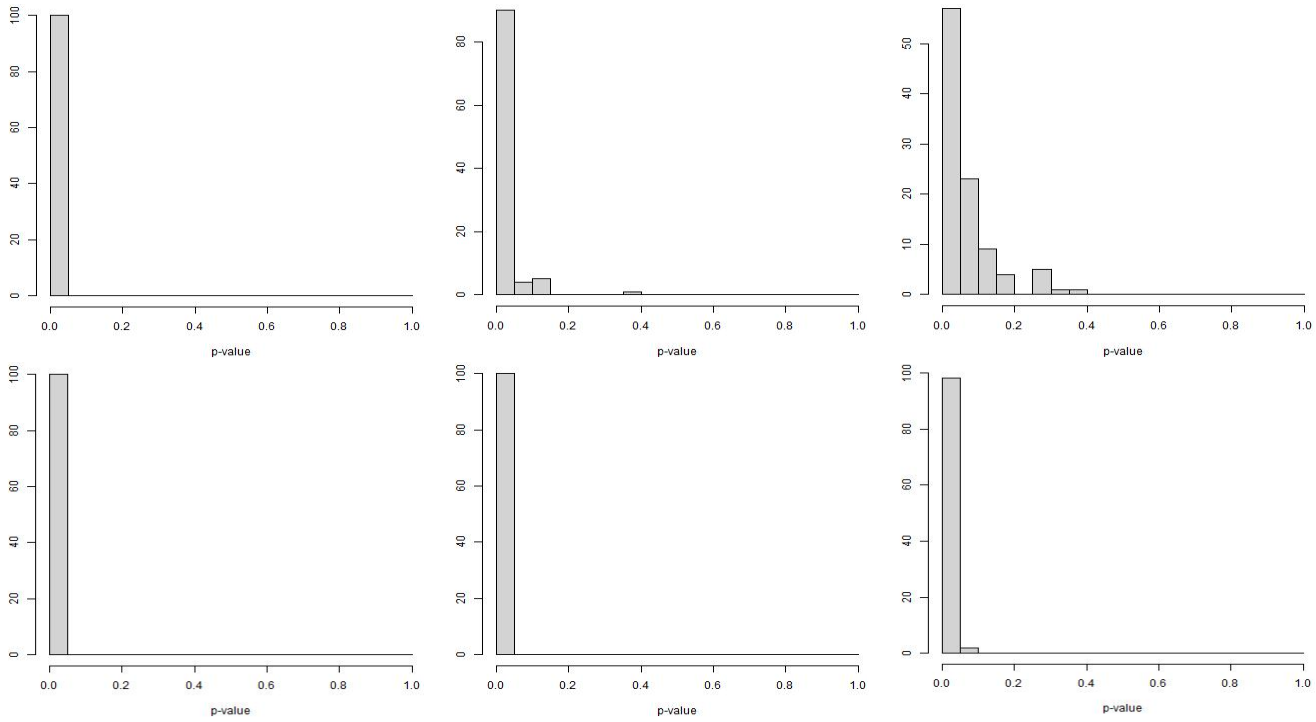

Fig. 7. *Histograms of p-values when testing similarity of the Boolean model vs the repulsive model (upper left), the Boolean model vs the cluster model (upper middle), the repulsive model vs the cluster model (upper right), the ellipse model vs the Boolean model (lower left), the ellipse model vs repulsive model (lower left) and the ellipse model vs the cluster model (lower left) using boot strap method and the samples of 100 components.*

the simulation study, and since visually, the pictures look very similar to that ones in the simulated models (in the sense of smoothness of component boundaries, distances between components etc.), we make the analysis of the real data with the same parameters as used in the simulation study, i.e. we take $R = 3$ and $R = 5$, respectively, and the samples of 10 and 20 components, respectively.

We test the similarity of the random sets represented by the images each to each. We repeat the procedure 100 times, while we evaluate the mean p-value of the test for each couple of images (including the image with itself) and the number of p-values below 0.05, which indicate significant dissimilarity on the classical level. The results for the samples of 20 components and $R = 5$ are introduced in Tab. 1 and Tab. 2. We can observe that the p-values are significantly lower and the number of p-values below 0.05 is significantly higher when comparing pairs of different types of tissue than that ones for pairs of the same types of tissue. Just note that for $R = 3$ and for the samples of 10 components from each image, the results are very similar.

DISCUSSION

A new statistical test for assessing (dis)similarity of two random sets has been constructed. It works with two realisations - one realisation of each of the random sets. The procedure focuses only on shapes of the components of the random sets, namely on the curvature of their boundaries together with the ratios of their perimeters and areas, and it does not take into account the positions of the components in the realisations, since it is very often required in practical applications.

The described method is equipped by a simulation study. The study shows that under quite mild conditions, the test has large power when distinguishing realisations of different models. The power is larger than the method in Gotovac et al. (2016) and one of the methods in Gotovac (2019), which also distinguishes realisations based on the shape of the components, and is comparable to another method in Gotovac (2019) and to the method in Debayle et al. (2021), which, however, takes into account the placement of components. Another advantage with respect to the method in Gotovac et al. (2016) is that it is not heuristic, and moreover, it does not require a lot of input parameters as needed in Gotovac et al. (2016) and Debayle et al. (2021).

Fig. 8. *Samples of mastopathy breast tissue.*

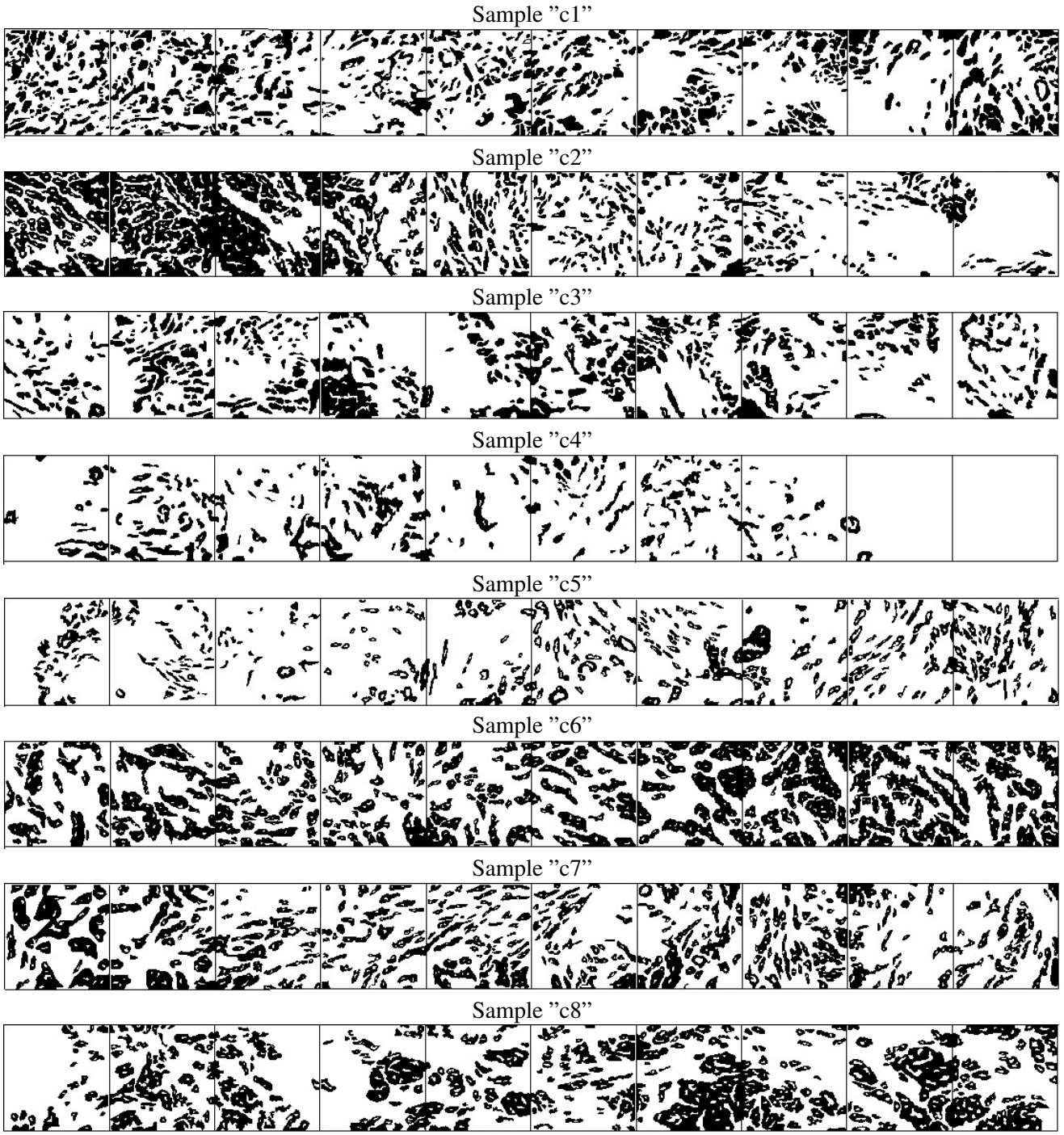

Fig. 9. *Samples of mammary cancer.*

Table 1. *Mean p-values (rounded to 2 decimal places) when comparing the corresponding samples 100 times. The values related to couples of different types of tissue are underlined.*

	m1	**m2**	**m3**	**m4**	**m5**	**m6**	**m7**	**m8**	**c1**	**c2**	**c3**	**c4**	**c5**	**c6**	**c7**	**c8**
m1	.81	.46	.35	.02	.19	.41	.04	.04	.01	.02	.04	.10	.02	.03	.04	.06
m2		.60	.10	.00	.05	.38	.07	.03	.01	.04	.06	.20	.03	.05	.06	.11
m3			.87	.35	.57	.20	.00	.00	.00	.00	.01	.01	.00	.01	.00	.00
m4				.90	.13	.04	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00
m5					.78	.32	.00	.00	.00	.00	.01	.01	.00	.01	.00	.01
m6						.65	.03	.01	.01	.03	.03	.14	.01	.09	.04	.10
m7							.59	.49	.14	.16	.10	.36	.45	.02	.19	.15
m8								.61	.10	.19	.10	.28	.40	.01	.15	.13
c1									.53	.38	.32	.17	.27	.11	.35	.26
c2										.55	.43	.41	.35	.17	.50	.37
c3											.57	.29	.18	.31	.47	.38
c4												.57	.25	.20	.35	.40
c5													.55	.03	.35	.16
c6														.54	.24	.40
c7															.51	.42
c8																.54

Table 2. *The number of p-values bellow* .05 *when comparing the corresponding samples 100 times. The values related to couples of different types of tissue are underlined.*

	m1	**m2**	**m3**	**m4**	**m5**	**m6**	**m7**	**m8**	**c1**	**c2**	**c3**	**c4**	**c5**	**c6**	**c7**	**c8**
m1	0	4	9	92	29	5	81	81	97	92	82	57	88	79	80	70
m2		3	67	100	71	13	71	84	95	86	78	37	86	65	74	57
m3			0	5	1	29	99	100	100	100	98	98	100	96	99	98
m4				0	47	77	100	100	100	100	100	100	100	96	100	100
m5					0	19	100	100	100	100	97	96	100	95	99	93
m6						0	87	96	96	91	89	54	97	62	84	62
m7							1	5	40	36	64	16	6	91	29	50
m8								1	52	29	60	15	17	97	33	50
c1									6	12	19	43	20	63	13	28
c2										2	13	8	20	52	3	14
c3											2	18	47	18	5	12
c4												1	21	39	20	12
c5													1	89	23	44
c6														7	35	16
c7															2	12
c8																3

A small complication occurs when components of compared realisations are too close to each other, because the method assumes independence of the components, which is not satisfied when the components are too densely placed. However, the problem has an easy solution - only a sample of the components can be considered instead of all components. In this case, the dependence is reduced and the method works very well. Concerning the sample size, it is very individual for different data. For a suitable choice, it is possible to apply, for example, the pre-analysis described in the section "Simulation study" and shown in Fig. 3, i.e. from one model (or from one realisation in practice), repeatedly select two samples of m components for a chosen number m, test the similarity of the samples, and check for which m the corresponding p-values are distributed uniformly.

Finally, the procedure is applied to real data. The data consists of pictures of two different types of tissue, namely the tissue of mammary cancer and mastopathic tissue. We have 8 pictures of each type. The aim is to distinguish between different types and to assess the pictures of the same type of the tissue as similar. Considering how different the images of the same type appear to be at first glance and how difficult it is to identify specific distinguishing features for different types, the results of the method are very good. For comparison to previous results, note that in the presented method, the type 2 error is slightly higher (i.e. the test power is slightly lower), but the type 1 error is significantly less than that ones observed in Gotovac (2019), where the similarity of realisations of random sets was tested using the same images.

From the above observations, we can conclude that the new method works very well and has a high potential to be a useful tool for comparing realisations of random sets.

Acknowledgment

Supported by The Czech Science Foundation, project No. 19-04412S, and by the Grant Agency of the Czech Technical University in Prague, project No. SGS21/056/OHK3/1T/13.

REFERENCES

Bullard JV, Garboczi EJ, Carter WC, Fuller ER Jr. (1995). Numerical methods for computing interfacial mean curvature. Comput Mater Sci. Vol. 4: 103–16.

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

Debayle J, Gotovac Đogaš V, Helisová K, Staněk J, Zikmundová M (2021). Assessing similarity of random sets via skeletons. Methodol Comput Appl Probab. Vol. 23: 471–490.

Gotovac V (2019). Similarity between random sets consisting of many components. Image Anal Stereol. Vol. 38: 185–99.

Gotovac Đogaš V, Helisová K (2021). Testing equality of distributions of random convex compact sets via theory of N-distances. Methodol Comput Appl Probab. Vol. 23: 503–526.

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

Gretton A, Borgwart KM, Rash MJ, Scholkopf B, Smola A (2012). A Kernel Two-Sample Test. J Mach Learn Res. Vol. 13: 723–73.

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. Vol. 34.18: 2636–61.

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

Matheron G (1975). Random Sets and Integral Geometry. John Wiley & Sons Inc, New-York.

Molchanov I (2005). Theory of random sets. Springer, New York.

Møller J, Helisová K (2008). Power diagrams and Interaction processes for unions of discs. Adv in Appl Probab. Vol. 40: 321–47.

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

Mrkvička T, Mattfeldt T (2011). Testing histological images of mammary tissues on compatibility with the Boolean model of random sets. Image Anal Stereol. Vol. 30(1): 11–18.

Myllymäki M, Mrkvička T, Grabarnik P, Henri Seijo H, Hahn U (2017). Global envelope tests for spatial processes. J R Stat Soc, Ser B (Stat Methodol). Vol. 79: 381–404.

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. Comput Mater Sci. Vol. 118: 353–64.

Serra J (1982). Image Analysis and Mathematical Morphology. Vol.2: Theoretical Advances. Academic Press.