Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Radioelectronics

Similarity of Random Sets

Bachelor thesis

Bogdan Radović

Field of study: Electronics and Communications
Supervisor: doc. RNDr. Kateřina Helisová, Ph.D.

Prague, 2021

BACHELOR‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name:	**Radović Bogdan**
Personal ID number:	**483887**
Faculty / Institute:	**Faculty of Electrical Engineering**
Department / Institute:	**Department of Radioelectronics**
Study program:	**Electronics and Communications**

II. Bachelor’s thesis details

Bachelor’s thesis title in English:

Similarity of Random Sets

Bachelor’s thesis title in Czech:

Podobnost náhodných množin

Guidelines:

1. Study a literature concerning theory of random sets and already existing results about their similarity.
2. Summarise already existing results and suggest a new statistical method for assessing similarity of random sets based on their realisations in the form of binary images.
3. Implement a program solution of the suggested method to simulated data sets.
4. Compare the results obtained by the method to selected previous procedures.

Bibliography / sources:

[1] Chiu S.N., Stoyan D., Kendall W.S., Mecke J. (2013): Stochastic Geometry and Its Applications. Wiley, Chichester.
[2] Gotovac V., Helisová K., Ugrina I. (2016): Assessing dissimilarity of random sets through convex compact approximatios, support functions and envelope tests. Image Analysis and Stereology 35, 181-193.
[3] Debayle J., Gotovac Đogaš V., Helisová K., Staněk J., Zikmundová M. (2021+): Assessing similarity of random sets via skeletons. Methodology and Computing in Applied Probability, DOI: 10.1007/s11009-020-09785-y

Name and workplace of bachelor’s thesis supervisor:

doc. RNDr. Kateřina Helisová, Ph.D., Department of Mathematics, FEE

Name and workplace of second bachelor’s thesis supervisor or consultant:

Date of bachelor’s thesis assignment: **21.01.2021** Deadline for bachelor thesis submission: __________

Assignment valid until: **30.09.2022**

doc. RNDr. Kateřina Helisová, Ph.D. Supervisor’s signature	doc. Ing. Josef Dobeš, CSc. Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the bachelor’s thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor’s thesis, the author must state the names of consultants and include a list of references.

.	
Date of assignment receipt	Student’s signature

Declaration

I hereby declare that I completed the presented thesis independently and that all used sources are quoted in accordance with the Methodological Instructions that cover the ethical principles for writing an academic thesis.

In Prague, 2021

....
Bogdan Radović

Acknowledgements

I would like to express my gratitude to the supervisor of this thesis, doc. RNDr. Kateřina Helisová, Ph.D., for such an interesting assignment and for all the help she provided, especially for keeping me on the right course during this thesis. Also, I'd like to thank Vesna Gotovac Đogaš, Ph.D., from the University of Split, for sharing her materials and results with me. Finally, gratitude has to be expressed to my beloved friends and family, whom all inspire me and motivate me as well.

Abstract

In recent years, random sets have become a very important tool for modelling various phenomena in biology, geology, medicine, material sciences etc. Usually, we try to find a suitable model for their observed realisations, but there are situations when it is not necessary, because our goal is only to compare the realisations and decide whether they come from the same process without knowledge of the process. This thesis aims to summarise the methods that already exist and propose a new algorithm for assessing the (dis)similarity of random sets. The suggested two-step algorithm focuses on the components inside the pattern rather than on pattern as a whole and evaluates respective border curvatures and ratios of perimeters and areas for each component. Finally, a simulation study has been performed that justifies the proposed procedure.

Keywords: Convex compact set, Curvature, Envelope test, N-distance, Permutation test, Random set, Similarity, Stochastic geometry

Abstrakt

Náhodné množiny jsou v posledních letech velmi důležitým prostředkem pro modelování různých jevů v biologii, geologii, v lékařství, materiálových vědách atd. Obvykle je snahou nalézt pro jejich realizace vhodný model, ale jsou situace, kdy to není nutné, neboť je cílem pouze porovnat dvě realizace a rozhodnout, zda pocházejí ze stejného procesu, i bez znalosti tohoto procesu. Cílem této práce je shrnout již existující metody a navrhnout nový algoritmus pro rozlišení náhodných množin. Navržený dvoukrokový algoritmus je soustředěn na komponenty uvnitř vzorů a vyhodnocuje příslušná zakřivení hranice a poměry obvodů a ploch pro každou komponentu. Nakonec je provedena simulační studie, která navrhovaný postup ověřuje.

Klíčová slova: Konvexní kompaktní množina, křivost, obálkový test, N-vzdálenost, permutační test, náhodná množina, podobnost, stochastická geometrie

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the last years, modelling and statistical analysis of random sets have been rapidly developing due to the fact that neither traditional methods for comparing random sets (e.g. covariance function or contact distribution function) nor technologically advanced image-processing tools (i.e. dilation, erosion, opening and closing) seem to be satisfactory for the problem of distinguishing between different natural processes. Geometrical patterns that occur in nature, for example, the position of trees in a forest or the distribution of cancer cells or soil cracks, are very complicated and, in most cases, random. The randomness and variability of data constantly forces scientists to develop new methods for processing the data.

In order to describe a particular phenomenon, we do not need to have knowledge about the generating process. Instead, it is enough to be able to determine whether two realisations come from the same process or not. The problem arises due to the fact that in nature, there is usually only one realisation of each considered process to work with. For that reason, statisticians and data scientists put emphasis on *statistical modelling*.

The process of making a model includes one essential step: analysing the data and choosing what is important and what would be of relevance. This step leads to the loss of information, some of which may be crucial. It is important to stress that the models that will be discussed in this thesis are not the models of this type due to the straightforward fact that we do not have enough data to feed the algorithm with, but the models built on the theory of *random processes*.

Random sets, the primary and potent tool of stochastic geometry, can be used for modelling various phenomena. They have been used over the past few decades, for modelling populations in ecology [1], roots in biology [2], and particles or gaps in material science [3]. Different types of applications can be looked up in books by Illian et al. [4], Baddeley and Jensen [5], and Chiu et al. [6]. In recent years, random sets have found applications in biomedicine for analysing cell patterns and tissues, see [7], [8] and [9].

It is interesting and important to note that the same principle that is used for analysing biological cells can be applied to wireless networks [10].

In this thesis, my goal will be to summarise already existing methods for assessing (dis)similarity of random sets, used primarily in biomedicine, and to suggest and implement a new algorithm that will be verified using simulated data. Consequently, the procedure will be applied to images of two types of mammary tissue in order to test its applicability in practice. In conclusion, the results obtained during writing this thesis will be compared to previous works by other authors.

The presented thesis is organised as follows. In Chapter 2, theoretical background is introduced. In Chapter 3 we provide a review of the most notable papers in the field focusing on the proposed methods and the results obtained. In Chapter 4 we present our approach which focuses on the shape of the components of the random sets. The algorithm is based on evaluating the curvature of the boundary and the ratio of the perimeter and the area of each component. In Chapter 5 we justify the procedure using simulated data and apply it to real data, that is, to two types of mammary tissue. In Chapter 6 we compare the presented method with the methods from Chapter 3, summarising the bases of their approaches, their advantages and disadvantages. In Chapter 7 we review the results and suggest possible topics for future research.

Chapter 2

Theoretical Background

In this chapter, general random set theory and stochastic geometry terms will be introduced in order to build an apparatus for achieving the goals of this work.

2.1 Random Set Theory

All definitions in this section can be found, with slightly different notation, in the book [6], unless stated otherwise.

Definition 2.1.1 (*Metric space*)**.** A *metric space* is an ordered pair $(\mathbf{X}, d)$, where $\mathbf{X}$ is usually $\mathbf{X} \subseteq \mathbb{R}^d$ and d is a mapping $d : \mathbf{X} \times \mathbf{X} \longrightarrow \mathbb{R}$ which satisfies the following conditions:

- $d(x, y) \geq 0$,
- $d(x, y) = 0$ iff $x = y$,
- $d(x, y) = d(y, x)$,
- $d(x, z) \leq d(x, y) + d(y, z)$,

for any $x, y, z \in \mathbf{X}$. The function d is called *metrics* on $\mathbf{X}$ or simply *distance.*

Example. [*Euclidean and Manhattan distance*] Let us consider Euclidean plane with two points $p = [p_1, p_2]$ and $q = [q_1, q_2]$. The *Euclidean distance* between p and q is then defined by

$$d_E = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}. \tag{2.1}$$

For the same points, their *Manhattan distance* is given by

$$d_M = |p_1 - q_1| + |p_2 - q_2|. \tag{2.2}$$

Definition 2.1.2 (*Ball*)**.** Suppose $(\mathbf{X}, d)$ is a metric space and let a be a point in $\mathbf{X}$. For each $r \in \mathbb{R}^+$ we define

- the *closed ball* in $\mathbf{X}$ centered at a with radius r as

$$\mathbf{D}(a, r) = \{x \in \mathbf{X} : d(a, x) \leq r\}, \tag{2.3}$$

- the *open ball* in $\mathbf{X}$ centered at a with radius r as

$$\mathbf{D}^{int}(a, r) = \{x \in \mathbf{X} : d(a, x) < r\}, \tag{2.4}$$

- the *sphere* as the difference between a closed and a concentric open ball

$$\mathbf{D}^{sph}(a, r) = \{x \in \mathbf{X} : d(a, x) = r\}. \tag{2.5}$$

Definition 2.1.3 (*Bounded set*)**.** A set $\mathbf{A} \subset \mathbb{R}^d$ is said to be *bounded* if there exists a ball $\mathbf{D}(x, r) \subset \mathbb{R}^d$, such that $\mathbf{A} \subset \mathbf{D}(x, r)$.

Definition 2.1.4 (*Open and closed sets*)**.** A set $\mathbf{A}$ is said to be *open* if $\forall \mathbf{x} \in \mathbf{A}$ there exists a positive number ε such that $\mathbf{D}(x, \varepsilon) \subset \mathbf{A}$. A set $\mathbf{A}$ is said to be *closed* if its complement $\mathbf{A^c}$ in $\mathbb{R}^d$ is open. The system of all closed subsets of $\mathbb{R}^d$ will be denoted as $\mathbb{F}$.

Definition 2.1.5 (*Interior, closure, and boundary*)**.** The *interior* $\mathbf{A^{int}}$ of the set $\mathbf{A}$ is the union of all open sets contained in $\mathbf{A}$. The *closure* $\mathbf{A^{cl}}$ of the set $\mathbf{A}$ is the intersection of all closed sets containing $\mathbf{A}$. The difference $\partial\mathbf{A} = \mathscr{B}_{\mathbf{A}} = \mathbf{A^{cl}} - \mathbf{A^{int}}$ is called the *boundary* of $\mathbf{A}$.

Definition 2.1.6 (*Compact set*)**.** A set $\mathbf{K} \subset \mathbb{R}^d$ is said to be *compact* if it is both closed and bounded. The system of all compact subsets of $\mathbb{R}^d$ shall be denoted as $\mathbb{K}$.

Definition 2.1.7 (*Topology*)**.** Let $(\mathbf{T}, \mathcal{T})$ be an ordered pair, where $\mathbf{T}$ is a set and $\mathcal{T}$ is a collection of open subsets of $\mathbf{T}$ satisfying:

- $\emptyset \in \mathcal{T}$and $\mathbf{T} \in \mathcal{T}$,
- $\bigcup_i \mathbf{T}_i \in \mathcal{T}$, for any sets $\mathbf{T}_i \in \mathcal{T}, i \in \mathbb{N}$,
- $\bigcap_i \mathbf{T}_i \in \mathcal{T}$, for any sets $\mathbf{T}_i \in \mathcal{T}$where i is finite.

Then the couple $(\mathbf{T}, \mathcal{T})$ is called *topological space* and the collection $\mathcal{T}$ is called the *topology* on $(\mathbf{T}, \mathcal{T})$.

Definition 2.1.8 (*Connected set*)**.** Let $(\mathbf{T}, \mathcal{T})$ be a topological space. A subset $\mathbf{X} \subset \mathbf{T}$ is called a *connected set* if it cannot be separated into two nonempty subsets such that each subset has no common points with the set closure of the other [11].

Definition 2.1.9 (*$\sigma-$algebra, Borel and Effros $\sigma-$algebras*)**.** For each set $\mathbf{X}$, a system $\mathcal{X}$ of its subsets is called σ-algebra if it satisfies the following:

- $\mathbf{X} \in \mathcal{X}$,
- if $\mathbf{A} \in \mathcal{X}$, then $\mathbf{A^c} \in \mathcal{X}$,
- if $\mathbf{A_1}, \mathbf{A}_2, ... \in \mathcal{X}$, then $\bigcup_{i=1}^{\infty} \mathbf{A_i} \in \mathcal{X}$.

The smallest $\sigma-$algebra on $\mathbb{R}^d$ containing all open subsets of $\mathbb{R}^d$ is called *Borel $\sigma-$algebra* and is denoted by $\mathcal{B}$.

The smallest $\sigma-$algebra on $\mathbb{R}^d$ containing all closed subsets of $\mathbb{R}^d$ is called *Effros $\sigma-$algebra* and is denoted by $\mathcal{F}$.

Definition 2.1.10 (*Measurable function*)**.** A function $f : \mathbf{X} \longrightarrow \mathbb{R}$ is said to be $\mathcal{X}-$ measurable if for each Borel set $\mathbf{B} \in \mathcal{B}$ the inverse image $f^{-1}(\mathbf{B})$ belongs to $\sigma-$algebra $\mathcal{X}$ associated with $\mathbf{X}$.

Definition 2.1.11 (*Lebesgue measure*)**.** For $\mathbf{Q} = [u_1, w_1] \times ... \times [u_d, w_d] \subset \mathbb{R}^d$ Lebesgue measure is defined by

$$v_d(\mathbf{Q}) = |\mathbf{Q}| = (u_1 - w_1) \cdot ... \cdot (u_d - w_d), \tag{2.6}$$

i.e. it is characterised by the volume of a d-dimensional hypercube.

Definition 2.1.12 (*Convex set*)**.** A set $\mathbf{K} \subset \mathbb{R}^d$ is said to be *convex* if for every $x, y \in \mathbf{K}$ and every $0 < \alpha < 1$ we have $\alpha x + (1 - \alpha)y \in \mathbf{K}$. Convex sets, which are also compact are called *convex bodies.*

Definition 2.1.13 (*Convex body functional*)**.** A *convex body functional* assigns a real value $h(\mathbf{K})$ for every $\mathbf{K} \in \mathcal{C}$, where $\mathcal{C}$ denotes the system of all convex bodies.

Example. Some of the most important convex body functionals of a set $\mathbf{K} \in \mathcal{C}$ in different dimensions are:

- length of a curve $l(\mathbf{K})$,
- boundary length $L(\mathbf{K})$ and area $A(\mathbf{K})$ of a planar set,
- surface area $S(\mathbf{K})$ and volume $V(\mathbf{K})$ of a 3D body.

Definition 2.1.14 (*Hausdorff metric*)**.** For two convex compact sets $\mathbf{A}, \mathbf{B} \in \mathcal{C}$ we define their *Hausdorff distance* as

$$d_H(\mathbf{A}, \mathbf{B}) = max\{\sup_{x \in \mathbf{A}} \inf_{y \in \mathbf{B}} d_E(x, y), \sup_{y \in \mathbf{B}} \inf_{x \in \mathbf{A}} d_E(x, y)\}. \tag{2.7}$$

Definition 2.1.15 (*Support function*)**.** For every convex $\mathbf{K}$ there is a unique support function defined by

$$s(\mathbf{K}, u) = \sup_{x \in \mathbf{K}} \langle u, x \rangle, \qquad u = \mathbf{D}^{sph}(0, 1), \tag{2.8}$$

where $\mathbf{D}^{sph}(0, 1)$ is a unit sphere in $\mathbb{R}^d$ [12].

Example. Identifying the points $u \in \mathbf{D}^{sph}(0, 1)$ with angles $u \in \langle 0, 2\pi)$, we get support functions as shown in Figure 2.1. Note that for a circle (disc) the support function is constant with the value equal to the radius, while for another set containing the origin, the interpretation of the support function is such that in each direction given by the angle, it is the distance of the origin and the so-called support plane, i.e. the line perpendicular to the direction which is as far as possible having nonempty intersection with the set. In other words, it describes a kind of a reach of the set in all directions.

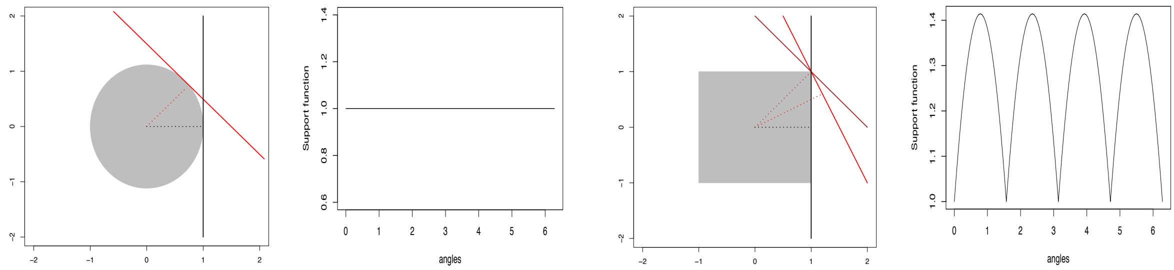


Figure 2.1: Support functions for a disc and a square [13]

Definition 2.1.16 (*Random closed set*)**.** Let (Ω, Σ, P) be a probability space. A measurable mapping $\mathbf{X} : (\Omega, \Sigma, P) \longrightarrow (\mathbb{F}, \mathcal{F})$ is a *random closed set* if for every compact $\mathbf{K} \in \mathbb{K}$ we have $\{\omega \in \Omega : \mathbf{X} \cap \mathbf{K} \neq \emptyset\} \in \Sigma$.

If we replace $\mathbb{K}$ by the system of convex bodies $\mathcal{C}$ in Definition 2.1.16, we get the definition of a *random convex compact set.*

Definition 2.1.17 (*Probability distribution of a random set*)**.** The *probability distribution* $P_{\mathbf{X}}$ of a random set $\mathbf{X}$ is defined by

$$P_{\mathbf{X}}(F) = P(\mathbf{X}^{-1}(\mathbf{F})) = P(\mathbf{X} \in \mathbf{F}), \tag{2.9}$$

for every $\mathbf{F} \in \mathcal{F}$.

Definition 2.1.18 (*Independent random sets*)**.** Two random sets $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if for any $\mathbf{F}_1$ and $\mathbf{F}_2$ in $\mathcal{F}$ we have

$$P(\mathbf{X}^{-1}(\mathbf{F}_1) \cap \mathbf{Y}^{-1}(\mathbf{F}_2)) = P(\mathbf{X}^{-1}(\mathbf{F}_1)) \cdot P(\mathbf{Y}^{-1}(\mathbf{F}_2)). \tag{2.10}$$

We can find this definition in [14].

Definition 2.1.19 (*Stationarity, isotropy*)**.** A random closed set $\mathbf{X}$ is *stationary* if its distribution $P_{\mathbf{X}}(\mathbf{F}) = P(\omega \in \Omega : \mathbf{X}(\omega) \in \mathbf{F})$ for $\mathbf{F} \in \mathcal{F}$ is invariant under translation. A random closed set $\mathbf{X}$ is *isotropic* if its distribution is invariant under rotation. If a random closed set is both stationary and isotropic, it is called *motion invariant.*

Theorem 2.1.1 (*Lavie* [14])**.** Two random convex compact sets $\mathbf{X}_1$ and $\mathbf{X}_2$ are identically distributed if and only if their support functions share identical finite-dimensional distributions.

Definition 2.1.20 (*Neighbourhood*)**.** Consider a finite union of disjoint random sets $\{\mathbf{X}_1, ..., \mathbf{X}_n\}$ within an observation window $\mathbf{W} \subset \mathbb{R}^d$. Every set $\mathbf{X}_i$ generates a *neighbourhood*

$$\mathbf{H}^i_M = \{y \in \mathbf{W} : d_M(\{y\}, \mathbf{X}_i) \leq d_M(\{y\}, \mathbf{X}_j) \text{ for all } i \neq j\}. \tag{2.11}$$

Similarly, we can define neighbourhoods using Hausdorff metric

$$\mathbf{H}^i_H = \{y \in \mathbf{W} : d_H(\{y\}, \mathbf{X}_i) \leq d_H(\{y\}, \mathbf{X}_j) \text{ for all } i \neq j\}. \tag{2.12}$$

2.2 Point Processes

Multidimensional point processes are fundamental entities studied in stochastic geometry. They materialise in nature and our surroundings in the form of collections of cells, particles, spores, trees, or mobile phones, more precisely as a group of characteristic points in these bodies, such as centroids, centres of mass, or geographical locations. They are closely related to and play a role in both the theory and the applications of random sets.

Definition 2.2.1 (*Point process*)**.** Let (Ω, Σ, P) be a probability space. Consider $\mathbb{G}$, the system of locally finite subsets of $\mathbb{R}^d$, with the σ-algebra $\mathcal{G} = \sigma(\{\mathbf{x} \in \mathbb{G} : \#(\mathbf{x} \cap \mathbf{A}) = m\} : \mathbf{A} \in \mathcal{B}, m \in \mathbb{N}_0)$, where $\mathcal{B}$ denotes the system of bounded Borel sets and $\#(\mathbf{x})$ represents the number of points in the configuration $\mathbf{x}$. A point process Φ defined on $\mathbb{R}^d$ is a measurable mapping from (Ω, Σ) to $(\mathbb{G}, \mathcal{G})$.

Definition 2.2.2 (*Distribution of a point process*)**.** The distribution P_Φ of the point process Φ is given by the relation $P_\Phi(\mathbf{G}) = P(\{\omega \in \Omega : \Phi(\omega) \in \mathbf{G}\})$ for $\mathbf{G} \in \mathcal{G}$.

Definition 2.2.3 (*Intensity and homogeneity of a point process*)**.** A measure Λ on $\mathcal{B}$ satisfying $\Lambda(\mathbf{A}) = \ \Phi(\mathbf{A})$ for all $\mathbf{A} \in \mathcal{B}$, where $\Phi(\mathbf{A})$ denotes the number of points in $\mathbf{A}$, is called the *intensity measure.* If there exists a function $\lambda(x)$ for $x \in \mathbb{R}^d$ such that $\Lambda(\mathbf{A}) = \int_{\mathbf{A}} \lambda(x)dx$, then $\lambda(x)$ is called the *intensity function.* If the intensity function $\lambda(x)$ is constant, $\lambda(x) = \lambda$, the point process is called *homogeneous* with the *intensity* λ. Otherwise, it is said to be *inhomogeneous.*

Definition 2.2.4 (*Poisson point process*)**.** Let Λ be a locally-finite non-null measure on $\mathbb{R}^d$. The *Poisson point process* Φ of intensity measure Λ is defined using its finite-dimensional distributions:

$$P(\Phi(\mathbf{A}_1) = m_1, ..., \Phi(\mathbf{A}_k) = m_k) = \prod_{i=1}^{k} e^{-\Lambda(\mathbf{A}_i)} \cdot \frac{\Lambda(\mathbf{A}_i)^{m_i}}{m_i!}, \tag{2.13}$$

for every $k = 1, 2, ...$ and all bounded, disjoint sets $\mathbf{A}_i$, $i = 1, 2, ..., k$, such that $\mathbf{A}_i \subset \mathbb{R}^d$. If $\Lambda(\mathbf{A}_i) = \lambda \cdot |\mathbf{A}_i| = \lambda \cdot v_d(\mathbf{A}_i)$, where λ is a constant, then Φ is called a *homogeneous Poisson point process* [10].

Since we work mainly with homogeneous Poisson point process Φ, we can say, in order to summarise, that it is characterised by:

- Poisson distribution of the number of points in each $\mathbf{A} \in \mathcal{B}$ with the parameter $\Lambda(\mathbf{A})$,
- independent scattering, i.e. the numbers of points in disjoint sets are independent random variables.

One of the operations that are occasionally applied to the Poisson point process is *thinning.* It is characterised by *retention function* p, where $p = p(x)$ denotes the probability that the point $x \in \Phi$ will not be deleted. It means that we can construct a thinned Poisson process Φ^p from Φ by randomly and independently removing points. If the original Poisson process is homogeneous with intensity λ, then the retained process is also homogeneous Poisson process with intensity $\lambda_p = p \cdot \lambda$. This property can be generalised for inhomogeneous Poisson process with intensity λ (the result is known as *Prekopa's theorem*): retained points will form either homogeneous or inhomogeneous Poisson process of intensity $\lambda_p = \lambda \cdot p(x)$. It should be noted that the thinned and the retained process are mutually independent [15].

Example. An example of thinning of a homogeneous Poisson process with retention factor $p = 0.75$ is shown in Figure 2.2. In this case, retained points are represented with blue colour, while deleted points are red.

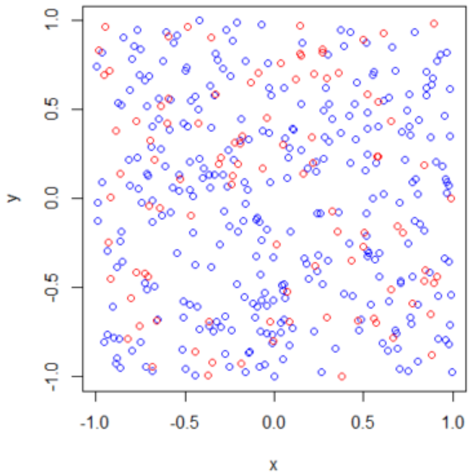


Figure 2.2: Thinning of a homogeneous Poisson process with retention factor $p = 0.75$ [15]

Definition 2.2.5 (*Boolean model*)**.** Let $\mathbf{Y} = \{y_1, y_2, ...\}$ be a stationary Poisson point process in $\mathbb{R}^d$ and $\{\mathbf{B}_1, \mathbf{B}_2, ...\}$ be a sequence of independent identically distributed (i.i.d.) random compact sets in $\mathbb{R}^d$ that are mutually independent and independent of $\mathbf{Y}$. If $\mathbb{E}|\mathbf{B}_1 \oplus \mathbf{K}| < \infty$ for all compact sets $\mathbf{K}$, then the random set

$$\mathbf{B} = \bigcup_{i=1}^{\infty} (y_i + \mathbf{B}_i) \tag{2.14}$$

is called the *Boolean model*.

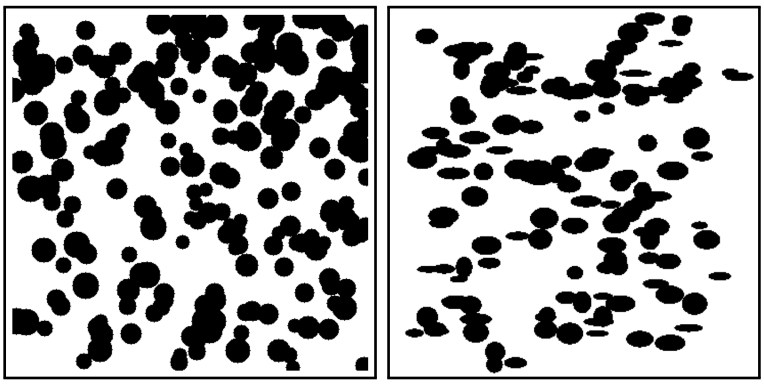


Figure 2.3: Boolean model: random discs (left) and random ellipses (right)

Boolean model is sometimes called Poisson germ-grain model [6]. It can be easily modelled using the Poisson point process with the intensity λ, where around each point of the Poisson process we construct a random geometrical object (e.g. a line segment, a disc, a polygon, a ball etc.). The resulting union is an example of a Boolean model.

The name germ-grain model comes from the point of view that the points of the Poisson process form the *germs*, while the geometrical objects are their corresponding *grains.* The Boolean model is an extremely powerful tool for modelling various natural and artificial phenomena, see [6].

Definition 2.2.6 (*Random disc Quermass-interaction process*)**.** Consider a planar random disc Boolean model. The *random disc 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 is given by

$$f_\theta(\mathbf{D}) = \frac{1}{c_\theta} exp\{\theta_1 A(U_{\mathbf{D}}) + \theta_2 L(U_{\mathbf{D}}) + \theta_3 \chi(U_{\mathbf{D}})\}, \tag{2.15}$$

for each finite disc configuration $\mathbf{D} = \{\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_n\}$, where A, L and χ are, respectively, the area, the perimeter and the Euler–Poincaré characteristic (the number of holes subtracted from the number of connected components) of the union of discs $U_{\mathbf{D}} = \bigcup_{i=1}^{n} \mathbf{D}_i$, $\theta = (\theta_1, \theta_2, \theta_3)$ is a three-dimensional vector of parameters, and c_θ is the normalising constant [16].

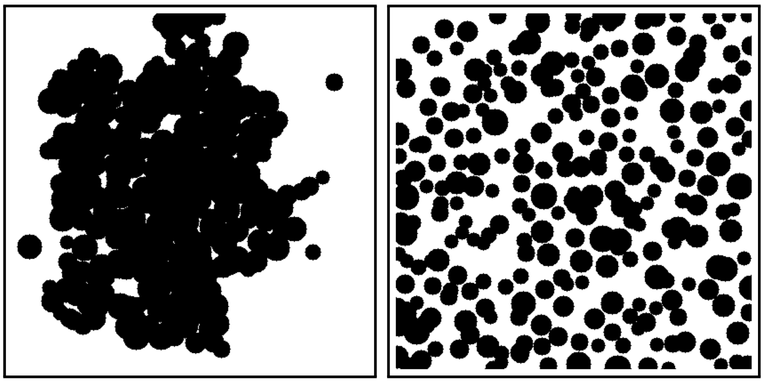


Figure 2.4: Quermass-interaction process: cluster (left) and repulsive (right)

Definition 2.2.7 (*Marked point process*)**.** A *marked point process* is a random sequence $\Psi = \{x_k; m_k\}$ where x_k are points of an unmarked point process (known in the literature as a *ground process* [6]) and m_k are the marks corresponding to the points x_k coming from a given *space of marks* $\mathbb{M}$.

2.3 Random Sets in 2D Image Processing

Sets in Euclidean space can be equipped with following operations:

- Multiplication by real numbers

$$\alpha\mathbf{A} = \{\alpha \cdot x : x \in \mathbf{A}\}. \tag{2.16}$$

If $\alpha = -1$, then $\alpha\mathbf{A}$ is called *reflection* and denoted as $\check{\mathbf{A}}$.

- Translation

$$\mathbf{A}_x = \{x + y : y \in \mathbf{A}\}. \tag{2.17}$$

- Minkowski-addition

$$\mathbf{A} \oplus \mathbf{B} = \{x + y : x \in \mathbf{A}, y \in \mathbf{B}\} = \bigcup_{y \in \mathbf{B}} \mathbf{A}_y. \tag{2.18}$$

- Minkowski-subtraction

$$\mathbf{A} \ominus \mathbf{B} = \bigcap_{y \in \mathbf{B}} \mathbf{A}_y. \tag{2.19}$$

When used for image processing, e.g. for processing data acquired by microscopes, scanners, or tomographs, random set theory is represented by operations used primarily to *modify* the structure of the image. Even though most of them lead to loss of information, they are still quite helpful in determining quantitative and qualitative characteristics of the image. Numerous examples of applications of these operations in medical imaging are introduced in [17].

Definition 2.3.1 (*Operations for image processing*)**.** These definitions come from [17].

- Thresholding
  Thresholding is a process used to convert grey-scale images to binary images, as shown in Fig 2.5. The procedure is very primitive: a constant τ (threshold) is given. If the intensity I_i of a pixel $z_i^\square$ is less than τ, then the pixel will be replaced by a black pixel, and similarly, if the intensity is greater than τ, the pixel will be replaced by a white pixel, see Figure 2.5. The authors of [18] categorise thresholding methods into six groups based on the information they use.

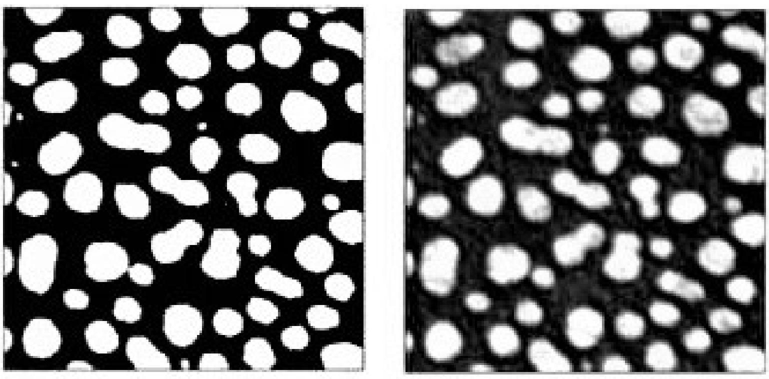


Figure 2.5: Grey-scale image transformed into a binary image using thresholding [19]

- Dilation
  Dilation is the process of enlarging or thickening without changing the shape. It is

a non-linear operation defined by

$$\mathbf{A} \longrightarrow \mathbf{A} \oplus \check{\mathbf{B}} \tag{2.20}$$

If set $\mathbf{B}$ is a ball, then dilation will smooth the resulting image. It is important to mention that dilation is not simple scaling – dilation fills in craters and orifices and connects separated fragments, see Figure 2.6.

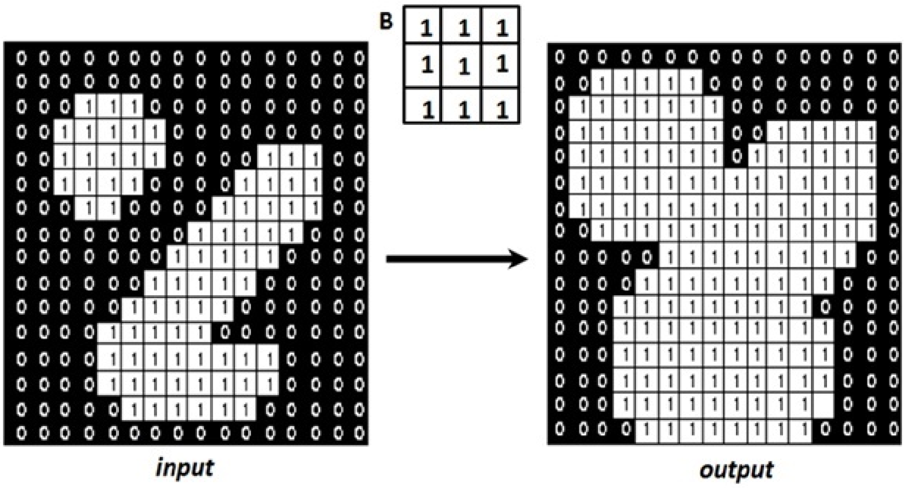


Figure 2.6: Dilation [20]

- Erosion

  Erosion is the process of shrinking or thinning, see Figure 2.7. It is defined by

$$\mathbf{A} \longrightarrow \mathbf{A} \ominus \check{\mathbf{B}} \tag{2.21}$$

  Again, we should mention that it is not mere scaling – connected components can be turned into fragments and some useful information, alongside noise and defects might be lost if set $\mathbf{B}$ is chosen unwisely. However, it is still the most used method for counting particles or other 3D objects which appear to be overlapping when projected to a 2D image.

- Opening

  Opening smooths the boundary of the object, eliminating defects and noise. It is achieved by applying dilation on a previously eroded set

$$\mathbf{A} \longrightarrow \mathbf{A} \circ \mathbf{B} \longrightarrow (\mathbf{A} \ominus \check{\mathbf{B}}) \oplus \mathbf{B} \tag{2.22}$$

  We could say that the opening is achieved by fitting as many sets $\mathbf{B}$ inside $\mathbf{A}$, see Figure 2.8.

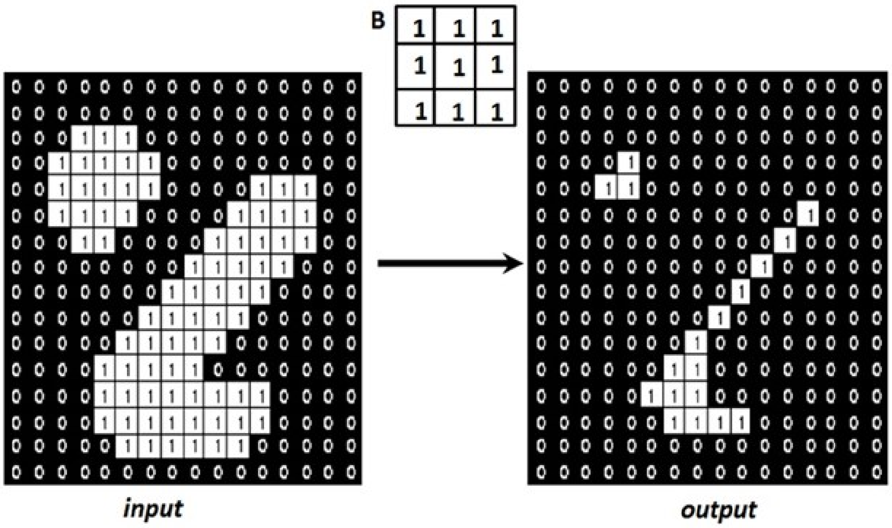


Figure 2.7: Erosion [20]

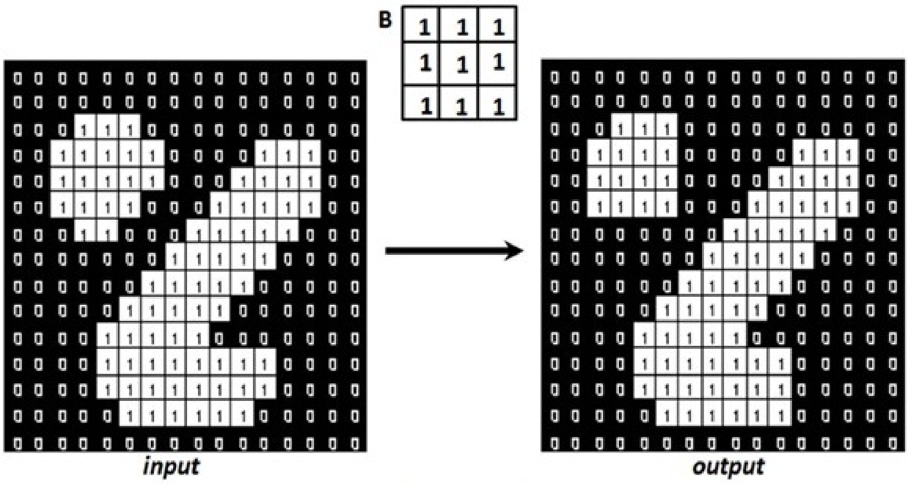


Figure 2.8: Opening [20]

- Closing

  Closing is operation dual to opening: it is dilation followed by erosion,

$$\mathbf{A} \longrightarrow \mathbf{A} \bullet \mathbf{B} \longrightarrow (\mathbf{A} \oplus \check{\mathbf{B}}) \ominus \mathbf{B}) \tag{2.23}$$

  Oppositely to opening, closing is achieved by fitting as many sets **B** outside of borders of **A**, see Figure 2.9.

2.4 Statistical Testing

In Chapters 3, 4 and 5, we use two statistical tests. The tests described in this section are developed on Monte Carlo goodness-of-fit test. The main idea is that if the null hy-

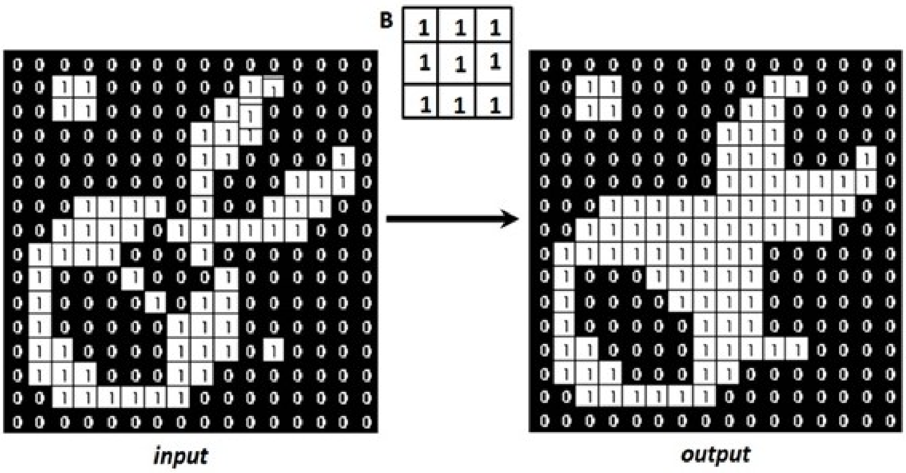


Figure 2.9: Closing [20]

pothesis is true, then the observed (functional) test statistics T_1 and N_{perm} independently simulated test statistics $T_2, ..., T_{N_{perm}+1}$ are identically distributed. That means that the null hypothesis can be rejected with the exact probability α if T_1 is among $\alpha(N_{perm}+1)$ extremal values of $T_i, i = 1, 2, ..., N_{perm}+1$ [21].

2.4.1 Testing Equality in Distribution Based on $\mathcal{N}-$distance of Probability Measures

For more information about the theory of $\mathcal{N}-$distance, see [22].

Definition 2.4.1 (*Negative definite kernel*)**.** Let $\mathbf{X}$ be a nonempty set. A map

$$\mathcal{L} : \mathbf{X} \times \mathbf{X} \to \mathbb{C} \tag{2.24}$$

is called *negative definite kernel* if for any $n \in \mathbb{N}$, arbitrary $c_1, ..., c_n \in \mathbb{C}$ such that $\sum_{i=1}^{n} c_i = 0$ and arbitrary $x_1, ..., x_n \in \mathbf{X}$ it holds

$$\sum_{i}^{n} \sum_{j}^{n} \mathcal{L}(x_i, x_j) c_i \bar{c}_j \leq 0. \tag{2.25}$$

Definition 2.4.2 (*Strongly negative definite kernel*)**.** Let $\mathbf{X}$ be a nonempty set and suppose that the map $\mathcal{L}$ is a real continuous function. The *negative definite kernel*

$$\mathcal{L} : \mathbf{X} \times \mathbf{X} \to \mathbb{R} \tag{2.26}$$

is called *strongly negative definite kernel* if for an arbitrary probability measure μ and an arbitrary real function $f : \mathbf{X} \to \mathbb{R}$ such that $\int_{\mathbf{X}} f(x)d\mu(x) = 0$ holds and

$$\int_{\mathbf{X}} \int_{\mathbf{X}} \mathcal{L}(x,y)f(x)f(y)d\mu(x)d\mu(y) \tag{2.27}$$

exists and is finite, the relation

$$\int_{\mathbf{X}} \int_{\mathbf{X}} \mathcal{L}(x,y)f(x)f(y)d\mu(x)d\mu(y) = 0 \tag{2.28}$$

implies that $f(x) = 0$ μ-almost everywhere.

For a map $\mathcal{L} : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$, denote $M_{\mathcal{L}}$ the set of all measures μ such that

$$\int_{\mathbf{X}} \int_{\mathbf{X}} \mathcal{L}(x,y)d\mu(x)d\mu(y) \tag{2.29}$$

exists.

Theorem 2.4.1 (*Klebanov* [22])**.** Let $\mathcal{L} : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ be a map satisfying $\mathcal{L}(x,y) = \mathcal{L}(y,x)$. Then $\mathcal{N}$-distance of the measures μ and ν is given by equation

$$\begin{aligned}\mathcal{N}(\mu,\nu) =&2\int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{L}(x,y)d\mu(x)d\nu(y) - \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{L}(x,y)d\mu(x)d\mu(y)\\ &- \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{L}(x,y)d\nu(x)d\nu(y) \geq 0\end{aligned} \tag{2.30}$$

which holds for all measures $\mu, \nu \in \mathcal{M}_{\mathcal{L}}$, where $M_{\mathcal{L}}$ denotes the set of all measures μ such that

$$\int_{\mathbf{X}} \int_{\mathbf{X}} \mathcal{L}(x,y)d\mu(x)d\mu(y) \tag{2.31}$$

exists, with equality in the case $\mu = \nu$ if and only if $\mathcal{L}$ is a strongly negative definite kernel.

The process of testing equality of distributions then consists of two parts:

- Estimating $\mathcal{N}$-distance.
- Deriving the p-value of the test.

Suppose that we have m_1 (functional) characteristics from objects $\mathbf{X}_1, ..., \mathbf{X}_{m_1}$ (e.g. m_1 support functions derived from m_1 parts of a realisation of a random convex compact set $\mathbf{X}$) with distribution μ and m_2 (functional) characteristics from objects $\mathbf{Y}_1, ..., \mathbf{Y}_{m_2}$ with distribution ν. We want to test if they come from the same distribution. The null hypothesis is H_0: $\mathbf{X}_1, ..., \mathbf{X}_{m_1}$ and $\mathbf{Y}_1, ..., \mathbf{Y}_{m_2}$ come from the same distribution, against H_A: the distributions are not the same. For calculation of $\mathcal{N}$-distance, an *arbitrary* strongly negative definite kernel can be used. Many examples of kernels for testing one-dimensional characteristics are introduced in [22]. Special kernels used for testing equality of distribution of random functions are derived in [9].

Definition 2.4.3 (*Empirical estimate of $\mathcal{N}$-distance*)**.** Assume we have an observation $\mathbf{X}_1, ..., \mathbf{X}_{m_1}$ from a distribution μ and $\mathbf{Y}_1, ..., \mathbf{Y}_{m_2}$ from a distribution ν. The $\mathcal{N}$-distance of the measures μ and ν is then estimated as

$$\hat{\mathcal{N}_1} = \frac{2}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathcal{L}(\mathbf{X}_i, \mathbf{Y}_j) - \frac{1}{m_1^2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_1} \mathcal{L}(\mathbf{X}_i, \mathbf{X}_j) - \frac{1}{m_2^2} \sum_{i=1}^{m_2} \sum_{j=1}^{m_2} \mathcal{L}(\mathbf{Y}_i, \mathbf{Y}_j). \quad (2.32)$$

This value plays the role of the test characteristic.

Then, a Monte Carlo permutation test [23] is used to make N_{perm} permutations of all observed values $\mathbf{X}_1, ..., \mathbf{X}_{m_1}, \mathbf{Y}_1, ...\mathbf{Y}_{m_2}$. After that, each permutation is divided into two groups of the lengths m_1 and m_2, and, analogously to (2.32), $\hat{\mathcal{N}_i}$ is calculated for the i-th permutation, $i = 2, ..., N_{perm} + 1$.

The p-value of the statistical test based on $\mathcal{N}$-distance is given by

$$p = \frac{\#\{i \in \{2, ..., N_{perm} + 1\} : \hat{\mathcal{N}_i} \geq \hat{\mathcal{N}_1}\} + 1}{N_{perm} + 1}. \quad (2.33)$$

The p-value is the value that can be used as a measure of similarity. However, note that due to the randomness following from random permutations in Monte Carlo method, it can be smaller than 1, even for identical shapes.

2.4.2 Testing Equality in Distribution Using Envelope Test

The envelope test examined here is described in detail in [21].

Definition 2.4.4 (*Extreme rank depth measure*)**.** Consider a group of $N_{perm} + 1$ random geometrical objects, where each object $i = 1, 2, ..., N_{perm} + 1$ is described by a characteristic $T_i(u), u \in I$, I being a finite index set. Let $R_i^{\uparrow}(u)$ and $R_i^{\downarrow}(u)$ denote the ranks of the values $T_i(u)$, ordered from the smallest (with rank 1) to the largest (with rank $N_{perm}+1$), and from the largest (with rank 1) to the smallest (with rank $N_{perm} + 1$), respectively. For each $u \in I$ *point-wise ranks of* $T_i(u)$ are defined by

$$R_i^*(u) = \min(R_i^{\uparrow}(u), R_i^{\downarrow}(u)), i = 1, ..., N_{perm} + 1. \quad (2.34)$$

Extreme rank measure is then defined as

$$R_i = \min R_i^*(u). \quad (2.35)$$

Definition 2.4.5 (*Rank lengths*)**.** *Rank length* and its vector are defined by

$$L_{i,k} = \int_I \mathbf{1}(R_i^*(u) = k)du \text{ and } \mathbf{L}_i = (L_{i,1}, ..., L_{i,\lfloor (N_{perm}+1)/2 \rfloor}). \tag{2.36}$$

For testing hypothesis H_0 that $T_1(u)$ has the same distribution as $T_i(u)$, where $i = 2, ..., N_{perm} + 1$, we have to define p-value.

The p-value is given by

$$p = \frac{1}{N_{perm} + 1}(1 + \sum_{i=1}^{N_{perm}+1} \mathbf{1}(\mathbf{L}_i \prec \mathbf{L}_1)), \tag{2.37}$$

where $\mathbf{L}_i \prec \mathbf{L}_j$ is a reverse lexical ordering of rank length vectors $\mathbf{L}_i$ used for ordering T_i

$$\mathbf{L}_i \prec \mathbf{L}_j \iff \exists n \leq \lfloor (N_{perm} + 1)/2 \rfloor : L_{i,k} = L_{j,k}, \forall k < n, L_{i,n} > L_{j,n} \tag{2.38}$$

In practice, we usually work with binary images, i.e. we observe discrete index set $I = u_1, ..., u_n$, so the definitions above have to be modified because we count ranks at discrete points $u_1, ..., u_n$, see Figure 2.10.

The p-value of the envelope test is defined as

$$p = \frac{1}{N_{perm} + 1}(1 + \sum_{i=1}^{N_{perm}+1} \mathbf{1}(\mathbf{N}_i \prec \mathbf{N}_1)), \tag{2.39}$$

where

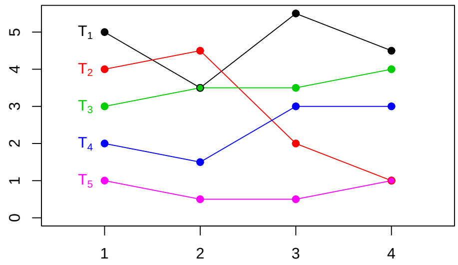
$$N_{i,k} = \sum_{j=1}^{n} \mathbf{1}(R_i^*(u_j) = k), \; \mathbf{N}_i = (N_{i,1}, ..., N_{i,\lfloor (N_{perm}+1)/2 \rfloor}). \tag{2.40}$$

Graphical explanation can be seen in Fig. 2.10.

The test then continues with the Monte Carlo permutation test, which is used to make N_{perm} permutations of testing functions $t_1^1(u), ..., t_{m_1}^1(u)$ from the first observation and $t_1^2(u), ..., t_{m_2}^2(u)$ from the second observation. The testing characteristic is then given (see [24]) by the normalised difference of their means

$$T_1(u) = \frac{\bar{t}^1(u) - \bar{t}^2(u)}{\sqrt{\text{var}t^1(u) + \text{var}t^2(u)}}. \tag{2.41}$$

The characteristics $T_2, ..., T_{N_{perm}+1}$ are calculated analogously with respect to the permutations in the Monte Carlo test.



$$
\begin{array}{ccccc}
R_1(1) = 1 & R_2(1) = 2 & R_3(1) = 3 & R_4(1) = 2 & R_5(1) = 1 \\
R_1(2) = 2 & R_2(2) = 1 & R_3(2) = 2 & R_4(2) = 2 & R_5(2) = 1 \\
R_1(3) = 1 & R_2(3) = 2 & R_3(3) = 2 & R_4(3) = 3 & R_5(3) = 1 \\
R_1(4) = 1 & R_2(4) = 1 & R_3(4) = 2 & R_4(4) = 3 & R_5(4) = 1
\end{array}
$$

$$\mathbf{N}_1 = (3,1,0),\ \mathbf{N}_2 = (2,2,0),\ \mathbf{N}_3 = (0,3,1),\ \mathbf{N}_4 = (0,2,2),\ \mathbf{N}_5 = (4,0,0)$$

$$\Rightarrow \mathbf{N}_5 \prec \mathbf{N}_1 \prec \mathbf{N}_2 \prec \mathbf{N}_3 \prec \mathbf{N}_4$$

Figure 2.10: Example of five functional characteristics, their point-wise ranks and ordering

Chapter 3

State of the Art

In this chapter, our main goal will be to summarise some of the already existing results in the field and to adumbrate paths taken by other authors with the aim of distinguishing between two realisations of random sets.

3.1 Convex Compact Approximations

This section reviews the paper [13]. For a detailed description of the procedure, the reader is referred to that paper. In Chapter 6, we will refer to this method using the abbreviation RC, which stands for random covering, which is the dominant tool in this method.

3.1.1 Algorithm

The starting point for the algorithm is a binary image $\mathbf{X}$ containing a digitised planar set $\mathbf{S}$. Digitisation is done in two steps:

1. Thresholding, as defined in Definition 2.3.1,

2. Masking - the pixel $z_i^{\square} \in \mathbf{X}$ is black if and only if its centre lies in $\mathbf{S}$.

The second step after getting the digital approximation $\mathbf{M}$ of $\mathbf{S}$ is to try to approximate the shape of $\mathbf{S}$ by covering $\mathbf{M}$ with 2D convex compact sets (the authors chose discs and consequent construction of Voronoi tessellations on their union), because they have desirable support functions, see Figure 2.1.

In order to cover $\mathbf{M}$ with (digitised) discs $\mathbf{D}(x_i, r)$, whose centres come from the *maximal Poisson-disc sampling*, we have to construct a point process $\mathbf{\Phi} = \{x_i, ..., x_n\}$ satisfying:

- For every $x_i, x_j \in \mathbf{\Phi}$, such that $x_i \neq x_j$ we have $|x_i - x_j| \geq r$,

- If $\mathbf{M}_{i-1} = \mathbf{M} - \bigcup_{j=1}^{i-1} \mathbf{D}(x_j, r)$, then for each $x_i \in \mathbf{\Phi}$ and each $\mathbf{A} \subset \mathbf{M}_{i-1}$

$$P(x_i \in \mathbf{A}) = \frac{|\mathbf{A}|}{|\mathbf{M}_{i-1}|}. \tag{3.1}$$

However, this process is problematic because it is strongly dependent on choosing the right r - this deviation can increase some of the body functionals, namely area and boundary length. The authors tried to avoid this flaw by firstly eroding the set $\mathbf{M}$ using $\mathbf{D}(o, r)$, where o represents the origin, and then covering eroded set $\mathbf{M}_r = \mathbf{M} \ominus \mathbf{D}(o, r)$ using maximal Poisson-disc sampling [25] and thus gaining the covered set $\mathbf{M}_c$. This new method was firstly applied to border pixels in order to preserve the shape, but it led to the loss of some inner pixels in the resulting set $\mathbf{M}_c$. In order to cover some of them, to some predefined threshold τ, the authors introduced pixel difference measure.

Definition 3.1.1 (*Pixel difference measure*)**.** Let $\mathbf{X}$ be the original digital picture. *Pixel difference measure* is defined by

$$PD(r) = \frac{\#\widehat{BW}(r) + \#\widehat{WB}(r)}{\#B_{\mathbf{X}}^{orig}}, \tag{3.2}$$

where $\#B_{\mathbf{X}}^{orig}$ denotes the number of black pixels in $\mathbf{X}$, i.e. $\#B_{\mathbf{X}}^{orig} = A(\mathbf{M})$, and $\widehat{BW} = B \longrightarrow W$ (similarly, $\widehat{WB} = W \longrightarrow B$) denotes the number of pixels that changed from black in $\mathbf{X}$ to white in $\mathbf{M}_c$ (respectively from white to black).

This led to another inaccuracy because choosing different τ can lead to different results of covering algorithm for previously selected r.

After covering some inner pixels, the authors constructed Voronoi tessellation on the union of discs $\mathbf{D}$ covering the set $\mathbf{M}$ which served as a basis for similarity measurement.

Definition 3.1.2 (*Voronoi tessellation*)**.** Let a finite disc configuration $\{\mathbf{D}_1, ..., \mathbf{D}_n\}$ of discs $\mathbf{D}_i(c_i, r)$ be given. The system $\mathcal{V}$ of all sets $\mathbf{V}_i$ such that

$$\mathbf{V}_i = \{y \in \mathbf{D}_i : |y - c_i| \leq |y - c_j| \text{for all } j \neq i\} \tag{3.3}$$

is called *Voronoi tessellation on the union* $\cup_{i=1}^{n}\mathbf{D}_i$ *of the discs* $\{\boldsymbol{D}_1, ..., \boldsymbol{D}_n\}$.

The workflow of the algorithm is shown in Figure 3.1.

3.1.2 Results

In the simulation study, authors heuristically derived optimal values for r and τ for each of the studied models: random-disc Boolean model, as shown in Figure 2.3, a cluster model

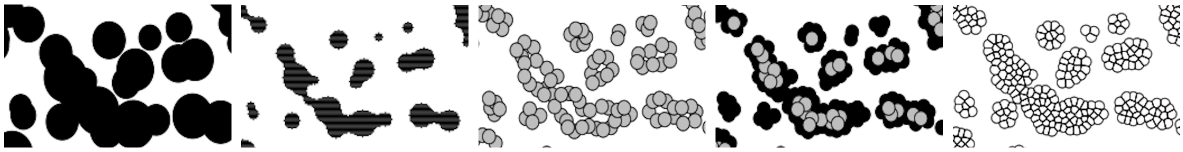


Figure 3.1: Covering of a planar set by discs of identical radii using adjusted Poisson disc sampling, and consequent construction of Voronoi tessellation on their union: digital approximation, reduced set, covering of border pixels, covering of inner pixels, construction of Voronoi tessellation, respectively [13]

and a repulsive model (Quermass-interaction processes with suitably chosen parameters), as shown in Figure 2.4. For more information about simulating algorithms, see [26]. For each of 200 realisations organised in 100 pairs, 100 randomly sampled non-neighbouring Voronoi cells were chosen. Consequently, their support functions were calculated and aligned using agglomerative hierarchical clustering. Support functions obtained in this way were used as testing functions for the envelope test [21] and described in Section 2.4.2. Histograms of p-values that were acquired in this manner show that the proposed algorithm is able to distinguish between different processes (p-values are close to 0) while it was slightly weaker when comparing the Boolean and the repulsive model, see the first row in Figure 6.1 (E stands for envelope test).

The main disadvantage of this approach is the obligation to choose 'free' parameters - size (radius) of covering discs, and pixel difference level. Due to the inability of the envelope test to capture the similarity of differently oriented cells of the same shape, the method is sensitive to rotation. Also, edge-effects may play a significant role.

3.1.3 Improvement by Using $\mathcal{N}$-distance Test

In [9], the authors used the same algorithm for the approximation of components by convex compact sets, but instead of the envelope test, they used the test based on $\mathcal{N}$-distances, see Section 2.4.1. They constructed some special negative definite kernels for testing the equality in distribution for random functions. Such tests have much higher power than the envelope test since they are not as sensitive to rotations of the tested convex compact sets. For comparison, see Figure 6.1.

3.2 Skeletons

This section is the shortened description of the method explained in detail in the paper [24]. For more information about the algorithm, a reader is referred to that paper. In Chapter 6, we will refer to this method using the abbreviation S.

3.2.1 Algorithm

The starting point for this algorithm is a binary image $\mathbf{X}$. Note that Definition 3.2.1 has to be slightly corrected: discs considered in this section are open and $\mathbf{A}$ is an open set in $\mathbb{R}^2$. Let further $\mathbf{D}_{max}(\mathbf{A})$ be the set of all maximal discs with respect to $\mathbf{A}$, see below.

Definition 3.2.1 (*Maximal disc*)**.** A disc $\mathbf{D}(x, r)$ is called *maximal* with respect to the set $\mathbf{A}$ if there is no other disc **D'** included in $\mathbf{A}$ and containing $\mathbf{D}(x, r)$.

Definition 3.2.2 (*Skeleton*)**.** The *skeleton* SK($\mathbf{A}$) of the set $\mathbf{A}$ is defined by

$$SK(\mathbf{A}) = \{z : \mathbf{D}(z, r) \in \mathbf{D}_{max}(\mathbf{A}), r > 0\}. \tag{3.4}$$

For $r > 0$, the r-th skeleton subset is defined as

$$S_r(\mathbf{A}) = \{z \in SK(\mathbf{A}) : \mathbf{D}(z, r) \in \mathbf{D}_{max}(A)\}. \tag{3.5}$$

It is easily deduced that

$$SK(\mathbf{A}) = \bigcup_{r>0} S_r(\mathbf{A}). \tag{3.6}$$

Using the language and notation of morphological transformations defined in Definition 2.3.1, we can write that

$$SK(A) = \bigcup_{r>0} \bigcap_{s>0} \{(\mathbf{A} \ominus \check{\mathbf{D}}(z, r)) - \{[(\mathbf{A} \ominus \check{\mathbf{D}}(z, r)) \ominus \check{\mathbf{D}}(z, s)] \oplus \mathbf{D}(z, s)\}\}, \tag{3.7}$$

where $\mathbf{D}(z, r), \mathbf{D}(z, s) \in \mathbf{D}_{max}(\mathbf{A})$. The original set can be easily reconstructed (as shown in Figure 3.2):

$$\mathbf{A} = \bigcup_{r>0}^{r_{max}} \{S_r(\mathbf{A}) \oplus \check{\mathbf{D}}(z, r)\} \tag{3.8}$$

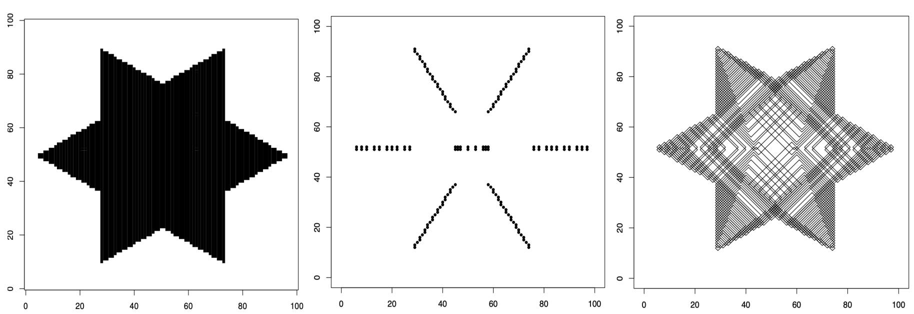


Figure 3.2: Example of a binary image of a set, its skeleton and reconstruction of the set using the skeleton and corresponding maximal discs [24]

Since we work with binary pictures, which are matrices of pixels $z^\square$ centered at points z, and since the number of pixels is a discrete value, instead of Euclidean metrics, the authors used Manhattan metrics as set in Definition 2.1, so the discs $\mathbf{D}(z,r)$ were transformed into *discrete discs* $\mathbf{D}(z,r) = \{\bigcup_i z_i^\square : d_M(z,z_i) \leq r\}$.

In order to assess the similarity of random sets, for each binary image $\mathbf{X}$ the authors defined the random testing function

$$t^{\mathbf{X}}(u) = \sum_{r=1}^{\infty} \sum_{y \in \mathbf{W}} r\mathbf{1}(|x-y|) < u)\mathbf{1}(\mathbf{D}(y,r) \in \mathbf{D}_{max}(\mathbf{X})), \quad u \in \mathbb{N}, \tag{3.9}$$

where x is a randomly chosen point from $SK(\mathbf{X})$, and $\mathbf{W}$ is a bounded observation window. Since the skeletons are formed from centres of maximal discs, the testing functions will be similar for points that are close to each other. In order to avoid misleading results, they introduced minimal distance d_{min} of points in which testing functions are calculated.

Let us consider a binary image containing a stationary random set $\mathbf{X}$ and its skeleton $SK(\mathbf{X})$. For each point $z_i \in SK(\mathbf{X})$ denote r^i_{max} the radius of the maximal disc centred in z_i. Then, the testing function at the point z_i can be approximated by

$$t_i(u) = \sum_{j \neq i} r^j_{max} \mathbf{1}(|z_i - z_j| < u), \tag{3.10}$$

where $u = 1, 2, ..., U_{max}$ and $U_{max} \in \mathbb{N}$.

After choosing the minimal distance d_{min}, which is strongly dependent on the chosen U_{max} (both values affect overlapping of discs $\mathbf{D}(z_j, r^j_{max})$ which leads to dependency between respective t_j) a subset $\mathbf{M}$ of $SK(\mathbf{X})$ is constructed using Mátern thinning method [27], where for each point $z_j \in SK(\mathbf{X})$ corresponding r^j_{max} are used as marks. Its points are then used as starting points for calculating t_j.

3.2.2 Results

In the simulation study, the authors compared realisations of the same models as introduced in Section 3.1.2. For all of 200 realisations of each model organised in 100 pairs, testing functions were calculated in 20 values, i.e. $U_{max} = 20$, with $d_{min} = U_{max}$. Histograms of p-values that were acquired show that the algorithm based on skeletons can distinguish between different processes (p-values are less than 0.05, see Figure 6.1, where E stands for envelope, while N stands for $\mathcal{N}$-distance test) but it is sensitive to the choice of parameters when we compare two realisations of the same processes and assess them as similar. The authors provided a discussion about the choice of the parameters, their advantages, disadvantages, possible complications and instructions on how to avoid them.

The main disadvantage of this procedure is voluntarism when choosing testing parameters and extreme sensitivity to small changes in shape, for example, change of width or the number of internal holes. The second disadvantage is critical when working with real data (i.e. images of tissues, see Section 5.2) because the presence of optical noise can create small holes in resulting images.

3.3 Symmetric Differences of Components and Neighbourhoods

This method comes from [16]. For detailed description, a reader is referred to the paper. In Chapter 6, we will refer to this method using the abbreviation T, which stands for tessellation.

3.3.1 Algorithm

The starting point for the algorithm is, as in the previous cases, a binary image $\mathbf{X}$ containing a realisation of a random planar set $\mathbf{S}$ consisting of components $\mathbf{M}_i, i \in \mathbf{I}$ (which are, in fact, discretised versions of disjoint random sets).

The first step in the algorithm is to mark all connected components of the studied realisations. For each pair of components, the authors defined their *symmetric difference*.

Definition 3.3.1 (*Symmetric difference between components*)**.** For each pair of components (M_i, M_j), $M_i, M_j \subset \mathbf{X}$ we define their symmetric difference as

$$\Delta(\mathbf{M}_i, \mathbf{M}_j) = \{x \in \mathbf{X} : (x \in \mathbf{M}_i \wedge x \notin \mathbf{M}_j) \vee (x \in \mathbf{M}_j \wedge x \notin \mathbf{M}_i)\}. \tag{3.11}$$

Definition 3.3.2 (*Frobenius matrix norm*)**.** Let $\mathbf{A}$ be a $m \times n$ matrix. Then its *Frobenius norm* is defined as

$$||\mathbf{A}||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{i,j}|^2}. \tag{3.12}$$

The second step, after the components are determined, is to construct their respective *neighbourhood tessellations*.

Definition 3.3.3 (*Neighbourhood tessellation*)**.** Let $\mathbf{H}_H^i$ denote Hausdorff neighbourhoods of components $\mathbf{M}_i, i \in \mathbf{I}$, see Definition 2.1.20. Then the system $\mathcal{H}$ of all sets $\mathbf{H}_H^i, i \in \mathbf{I}$ is called the *neighbourhood tessellation* on the union of the sets $\mathbf{M}_i$.

Examples of the neighbourhood tessellations are shown in Figure 3.3. Note that in this paper, the author used a cluster model simulated using slightly different parameters in order to obtain more components in realisations.

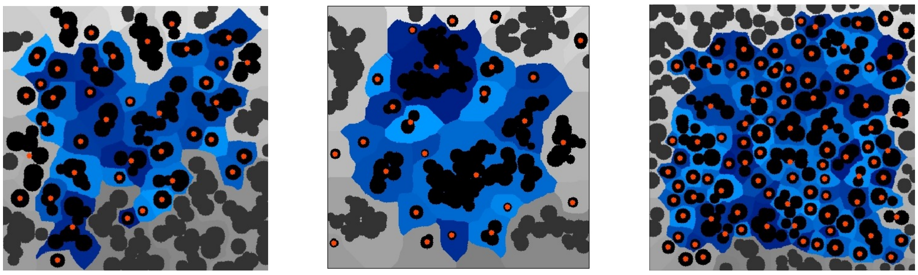


Figure 3.3: Components of the realisations of Boolean (left) cluster (middle) and repulsive (right) model used for simulation study together with their centroids and neighbourhoods [16]

The final step, after the neighbourhood tessellations are constructed, is to construct the similarity measure, for which the author used the approach based on $\mathcal{N}$-distances, as described in Section 2.4.1, with appropriate negative definite kernels.

For each pair of components M_i, M_j we can consider the negative kernel (used later for the construction of $\mathcal{N}-$distance)

$$\mathcal{L}(\mathbf{M}_i, \mathbf{M}_j) = \mu^{r/2}(\Delta(\mathbf{M}_i, \mathbf{M}_j)), \tag{3.13}$$

where $0 < r \leq 2$ and $\mu(\mathbf{A})$ is the area of the set $\mathbf{A}$.

Since for a binary image $\mathbf{X}$ the square of its Frobenius norm is equal to its area (measured in pixels), i.e. $||\mathbf{X}||_F = \mu^{1/2}(\mathbf{X})$, we can write $\mathcal{L}(\mathbf{M}_i, \mathbf{M}_j) = ||\Delta(\mathbf{M}_i, \mathbf{M}_j)||_F$. This kernel is used when we consider only the shapes of the corresponding components (this is denoted as CC kernel in the sequel), or we use $\mathcal{L}(\mathbf{H}_H^i, \mathbf{H}_H^j) = \mu^{1/2}(\Delta(\mathbf{H}_H^i, \mathbf{H}_H^j)) = ||\Delta(H_H^i, H_H^j)||_F$ when working with their neigbourhoods (this is denoted as N kernel in the sequel).

For each combination of ordered pairs $((M_i, H_H^i), (M_j, H_H^j))$ of components and their corresponding neighbourhoods we can consider the negative kernel

$$\begin{aligned} \mathcal{L}((M_i, H_H^i), (M_j, H_H^j)) &= \sqrt{\mu(\Delta(\mathbf{M}_i, \mathbf{M}_j)) + \mu(\Delta(\mathbf{H}_H^i, \mathbf{H}_H^j))} \\ &= \sqrt{||\Delta(M_i, M_j)||_F^2 + ||\Delta(H_H^i, H_H^j)||_F^2}. \end{aligned} \tag{3.14}$$

Note that this kernel should be used when we want to consider both shape and the position of a component inside realisation $\mathbf{X}$. It will be denoted as CCN kernel from now on.

3.3.2 Results

In the simulation study, the author studied the same models that were already introduced in Section 3.1.2, with the exception of the cluster model, which was slightly modified (see

Figure 5.3 in Section 5.1). For all possible pairs of models, 100 pairs of realisations were studied, i.e. their connected components were isolated and respective neighbourhood tessellations were constructed. Histograms of p-values obtained in this way were approximately uniformly distributed when comparing pairs of the same model (thus rejecting dissimilarity), and close to 0 when comparing pairs coming from different processes (thus rejecting similarity), see the third (CC kernel used), the fourth (N kernel used) and the fifth row (CCN kernel used) in Figure 6.1. When using the kernel which takes into account the shape of the neighbourhoods, the obtained p-values are even smaller when comparing different models.

The method was also applied to real data, i.e. to the images of two types of mammary tissue (thoroughly described in Section 5.2). In the first step, the multiply connected components (i.e. with many white holes) had to be decomposed into simply connected components so that each component has only one white hole and black pixels that are closer to that hole than to any other hole within that realisation. After that, the respective neighbourhood is constructed around the centroid of each component. Finally, 50 components from each image are sampled. The p-values of the pairs of samples that were obtained after testing the similarity show that the method is able to distinguish between the two types of tissue.

The main disadvantage of this method is its sensitivity to edge effects and its dependence on the distribution of the components and their neighbourhoods. For that reason, the method poorly distinguishes between cluster and other models.

Chapter 4

New Method for Assessing Similarity of Random Sets

As we have seen in Chapter 3, the method based upon morphological skeletons gives the best results in the simulation study. However, it is highly dependent on the placement of the components inside the image and small changes in shape. For that reason, we will examine individual components inside the picture. More precisely, we will focus on finding an algorithm that will be able to describe each component in a way as unique as possible for our purposes, focusing on the shape of the component. Simultaneously, the algorithm has to take into account the influence of components that are neighbouring the examined component (assumption of independence). Furthermore, it should minimise the effect of small changes in shape. In this chapter an algorithm will be proposed and presented, together with the theory it was built upon.

4.1 Curvature of a Planar Curve

The following two definitions come from [28]

Definition 4.1.1 (*Curvature of a curve*)**.** Let $\mathcal{C}$ be a smooth twice differentiable 2D curve that is properly parameterised by a parameter $s \in [0, s_{max}] \subset \mathbb{R}$, i.e. $\mathcal{C}(s) = (x(s), y(s))$. The *curvature* of the curve $\mathcal{C}$ in the point $\mathcal{C}(s)$ is then defined by

$$\kappa(\mathcal{C}(s)) = \frac{x'(s)y''(s) - x''(s)y'(s)}{(x'^2(s) + y'^2(s))^{3/2}}, \tag{4.1}$$

where $'$ denotes derivative with respect to s.

In other words, if $R(s)$ is the radius of the osculating circle touching the curve in the point $[x(s), y(s)]$, then the curvature is given by $\kappa(s) = \pm 1/R(s)$, where the choice between "+" and "−" is dictated by the local convexity convention.

Let $\mathcal{C}$ be a continuous, closed (i.e. $\mathcal{C}(0) = \mathcal{C}(s_{max})$), and non-self-intersecting curve (i.e. if $\mathcal{C}(s_1) = \mathcal{C}(s_2)$ then $s_1 = s_2$). Suppose that $\mathbf{S}$ is a planar (connected) set whose boundary is determined by $\mathcal{C}$ (with appropriately chosen orientation in order to ensure the right sign +/-). Curvature $\kappa(z)$, at the point $z \in \mathcal{C}$ and for r small enough, is then given by

$$\kappa(z) \approx \frac{3A^*(\mathbf{D}(z,r))}{r^3} - \frac{3\pi}{2r} = \frac{3\pi}{r}\left(\frac{A^*(\mathbf{D}(z,r))}{A(\mathbf{D}(z,r))} - \frac{1}{2}\right), \tag{4.2}$$

where $A(\mathbf{D}(z,r))$ is the area of the disc $\mathbf{D}(z,r)$ centred at z and $A^*(\mathbf{D}(z,r))$ is the area of $\mathbf{D}(z,r) \cap \mathbf{S}$ [28].

4.2 Implementation

The starting point for our algorithm is a binary image $\mathbf{X}$ containing a digital approximation $\mathbf{M}$ of a planar set $\mathbf{S}$, such that there are n black disjoint components $\mathbf{M}_k$, $k = 1, 2, ..., n$ inside $\mathbf{M}$, see Section 3.1.1. As it was already mentioned at the beginning of this chapter, we will focus on individual components. First, we have to (re)define a few terms in order to adapt them for working with binary pictures. Note that in the rest of the text, the terms *point* and *centre of a pixel* will have a synonymous meaning, while a *pixel* $z^\square$ will be interpreted as a square of the unit area centred at point $z = [x, y]$.

Since we are working with the binary image $\mathbf{X}$ of the set $\mathbf{S}$, we have to discretise the function (4.2) in such a way that area $A(\mathbf{D}(z,r))$ represents the number of pixels inside the disc $\mathbf{D}(z,r)$ centred at the boundary of $\mathbf{X}$, and $A^*(\mathbf{D}(z,r))$ is the number of pixels of $\mathbf{D}(z,r)$ inside $\mathbf{X}$.

Definition 4.2.1 (*4-neighbourhood*)**.** Let z be a point in a binary image $\mathbf{X}$.
4-neighbourhood of the pixel $z^\square$ is then defined as

$$\mathbf{H}_4 = \{\bigcup_i z_i^\square \in \mathbf{X} : d_M(z, z_i) \leq 1\}. \tag{4.3}$$

Definition 4.2.2 (*Boundary pixel, boundary*)**.** Let $\mathbf{M}_k$ be a connected random set consisting of black pixels $z_i^\square$, $\mathbf{M}_k \subset \mathbf{X}$. A pixel $z^\square \in \mathbf{M}_k$ is called a *boundary pixel* if and only if at least one of its neighbouring pixel in a 4-neighbourhood $\mathbf{H}_4$ is white. Union of all boundary pixels of the same component is called *boundary* and denoted by $\mathcal{B}_{\mathbf{M}_k}$.

Let $\mathbf{M}_k$ be a connected random set with boundary $\mathcal{B}_{\mathbf{M}_k}$. Then

- the boundary length $L(\mathcal{B}_{\mathbf{M}_k})$ is called *perimeter* and calculated by

$$L(\mathcal{B}_{\mathbf{M}_k}) = \#\{z_i^\square : z_i^\square \in \mathcal{B}_{\mathbf{M}_k}\}, \tag{4.4}$$

- the area $A(\mathbf{M}_k)$ is calculated as

$$A(\mathbf{M}_k) = \#\{z_i^\square : z_i^\square \in \mathbf{M}_k\}, \tag{4.5}$$

- for each component, we define a *ratio of its perimeter and area* as

$$R_{\mathbf{M}_k} = \frac{L(\mathcal{B}_{\mathbf{M}_k})}{A(\mathbf{M}_k)} = \frac{\#\{z_i^\square : z_i^\square \in \mathcal{B}_{\mathbf{M}_k}\}}{\#\{z_i^\square : z_i^\square \in \mathbf{M}_k\}}. \tag{4.6}$$

Example. An illustration of the algorithm is shown in Figure 4.1 where ellipse-shaped component $\mathbf{X}$ is given and a disc $\mathbf{D}$ with centre on the boundary point of the set $\mathbf{X}$ has been constructed. The resulting estimate of the curvature will be $\frac{\#C}{\#C+\#D} = \frac{5}{5+8} = \frac{5}{13}$, while the respective ratio of the perimeter and area will be $\frac{\#B}{\#E} = \frac{12}{19}$.

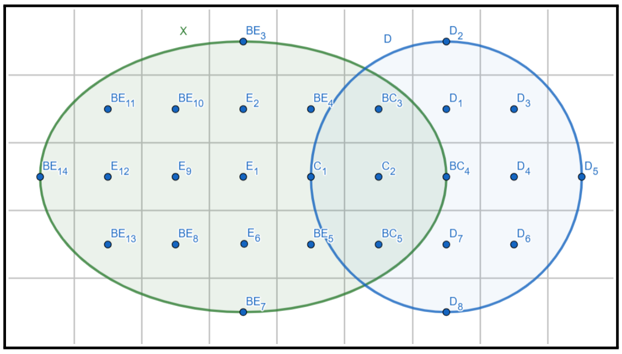


Figure 4.1: Estimating the curvature and the ratio of the perimeter and the area

4.3 Statistical Test

Once we have marked all border points of the connected random sets $\mathbf{M}_k$, we have to calculate curvature $\kappa_k(z)$, at each point $z \in \mathcal{B}_{\mathbf{M}_k}$. From equation (4.2), we can see that $\kappa_k(z)$ is proportional to

$$\kappa_k(z) \simeq \frac{A^*(\mathbf{D}(z,r))}{A(\mathbf{D}(z,r))} = O_{k,\mathbf{D}(z,r)}, \tag{4.7}$$

for appropriately chosen r. This fact will be used as a guideline for devising a testing characteristic.

Definition 4.3.1 (*Distribution of curvature*)**.** Let $O_{k,\mathbf{D}(z,r)}$ be the ratio as defined by equation (4.7). Define by

$$\tilde{\kappa}_{\mathbf{M}_k,\mathbf{D}(.,r)}(u) = \frac{1}{L(\mathcal{B}_{\mathbf{M}_k})} \int_{B_{\mathbf{M}_k}} \mathbf{1}\{O_{k,\mathbf{D}(z,r)} \leq u\} dz, \quad u \in \langle 0, 1 \rangle. \tag{4.8}$$

It is an analogy of the distribution function of the curvature at points on the boundary, with the difference that we work with highly dependent values here. From this function, an analogy to the density function can be defined as

$$t_{\mathbf{M}_k,\mathbf{D}(.,r)}(u) = \tilde{\kappa}'_{\mathbf{M}_k,\mathbf{D}(.,r)}(u) \tag{4.9}$$

which will be used as a testing function.

Definition 4.3.2 (*Similarity*)**.** Two connected random sets $\mathbf{X}$ and $\mathbf{Y}$ are considered to be similar if the distributions of $\lim_{r\to 0} t_{\mathbf{X},\mathbf{D}(.,r)}$ and $\lim_{r\to 0} t_{\mathbf{Y},\mathbf{D}(.,r)}$ and the distributions of $R_{\mathbf{X}}$ and $R_{\mathbf{Y}}$ defined by (4.6) are equal.

Since we are working with binary pictures, i.e. with discrete values, we have to approximate the distribution function of curvature.

Let $\mathbf{X}$ be a binary image containing a digitised realisation of a connected random set $\mathbf{M}_k$. For each boundary pixel z_i and a fixed radius $r \in \mathbb{N}$ we approximate

$$K(z_i) = \frac{\#\{z_j^\square \in \mathbf{X} : z_j^\square \in \mathbf{D}(z_i,r) \cap \mathbf{M}_k\}}{\#\{z_j^\square \in \mathbf{X} : z_j^\square \in \mathbf{D}(z_i,r)\}}. \tag{4.10}$$

Using this approximation, we can further set

$$t(u) = \frac{\#\{i \in \{1,\dots,n\} : K(z_i) \in [u - 1/l, u)\}}{n}, \quad u = \frac{1}{l}, \frac{2}{l}, \dots, 1, \tag{4.11}$$

which will be used as a testing function.

For testing the equality in distribution of random functions, we use the test based on $\mathcal{N}-$distances described in [9], as well as in Section 2.4.1. A kernel constructed especially for random functions can also be found in [9] and is defined by

$$\mathcal{L}(t^1,t^2) = \sum_{m=1}^{d} \sum_{\{k_1,\dots,k_m\}\subseteq\{1,\dots,n\}} \left(\sum_{l=1}^{m} \left(t^1(u_{k_l}) - t^2(u_{k_l}) \right) \right)^2 \Bigg)^{1/2}, \tag{4.12}$$

or an appropriately chosen d ($d = 3$ in our case, see [9]), so the estimate of $\mathcal{N}-$distance is of the form

$$\hat{\mathcal{N}_1} = \frac{2}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathcal{L}(t_i^{(1)}, t_j^{(2)}) - \frac{1}{m_1^2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_1} \mathcal{L}(t_i^{(1)}, t_j^{(1)}) - \frac{1}{m_2^2} \sum_{i=1}^{m_2} \sum_{j=1}^{m_2} \mathcal{L}(t_i^{(2)}, t_j^{(2)}). \tag{4.13}$$

When testing equality of the distributions of two realisations $\mathbf{X}_1, ..., \mathbf{X}_{m_1}$ and $\mathbf{Y}_1, ..., \mathbf{Y}_{m_2}$, of connected random sets $\mathbf{X}$ and $\mathbf{Y}$, we evaluate ratios $R_{\mathbf{X}_1}, ..., R_{\mathbf{X}_{m_1}}$, $R_{\mathbf{Y}_1}, ..., R_{\mathbf{Y}_{m_2}}$ (using

(4.6)) and testing functions $t_{\mathbf{X}_1}(u), ..., t_{\mathbf{X}_{m_1}}(u)$, $t_{\mathbf{Y}_1}(u), ..., t_{\mathbf{Y}_{m_2}}(u)$ (using (4.11)). In the next step we estimate corresponding $\mathcal{N}$-distances: $\hat{\mathcal{N}}_1^R$ (calculated by (2.32) where $\mathbf{X}_i$, $\mathbf{Y}_i$ are replaced by corresponding $R_{\mathbf{X}_i}$, $R_{\mathbf{Y}_i}$ respectively) and $\hat{\mathcal{N}}_1^t$ (calculated by (4.13)). The pair $(\hat{\mathcal{N}}_1^R, \hat{\mathcal{N}}_1^t)$ is the test statistic.

Then, the test continues as already described in Section 2.4.1: a Monte Carlo permutation test makes N_{perm} permutations of $\mathbf{X}_1, ..., \mathbf{X}_{m_1}$ and $\mathbf{Y}_1, ..., \mathbf{Y}_{m_2}$, and separate them into two groups of sizes m_1 and m_2, respectively. This way we obtain $\hat{\mathcal{N}}_i$, $i = 2, ..., N_{perm} + 1$, and evaluate p-value as

$$p = \frac{\#\{i \in \{2, \dots, N_{perm} + 1\} : \hat{\mathcal{N}}_i^R \geq \hat{\mathcal{N}}_1^R \wedge \hat{\mathcal{N}}_i^t \geq \hat{\mathcal{N}}_1^t\} + 1}{N_{perm} + 1}. \tag{4.14}$$

Chapter 5

Simulation Study and Application to Real Data

Our main goal in this chapter will be to show on simulated data how the two-step method for assessing similarity proposed in Chapter 4 is able to determine whether two processes are similar or not. In the second part of this chapter, we will apply the method to the real data - two different types of mammary tissue.

5.1 Simulated Data

Once we have defined appropriate test statistics (by (4.12) and (4.13)) and p-value (by (4.14)) of the $\mathcal{N}$-distance test, we should apply the procedure to the simulated data. The first step that is required is choosing the right value for the radius r of the disc that is used for calculating the curvature at the boundary point (by (4.10)). The area (measured in pixels) of the disc with radius r is given [29] by

$$A(\mathbf{D}(.,r)) = 1 + 4 \cdot \sum_{j \geq 0} \left(\left\lfloor \frac{r^2}{4j+1} \right\rfloor - \left\lfloor \frac{r^2}{4j+3} \right\rfloor \right). \tag{5.1}$$

A list with values for $r = 1, ..., 10000$ can be found in [30]. For our study, we will use $r = 3$ and $r = 5$, because choosing too large radius will lead to great mistake because the disc will not be able to recognise local changes in curvature while choosing too small radius will also lead to a mistake because the disc will not be able to detect curvature due to discretisation.

The next step is to simulate the data that will be compared. For all models that are taken into consideration, we simulate 200 realisations and compare 100 vs 100 realisations of the same models. Consequently, we compare 100 vs 100 realisations of different models. In the first place, for proper illustration of the strategy of the method, we will use the models shown in Figure 5.1, especially simulated for this purpose. The first two illustrating models are the *Boolean model* and *reduced Boolean model* (simulated from the Boolean model where deletion probability of each component is set to 0.5). The third model is the *square model* consisting of disjoint squares whose perimeter to area ratio comes from the same (empirically obtained) distribution as that of the ratio of the Boolean model. Finally, the fourth model is the *rectangle model* containing disjoint rectangles whose one side is fixed to a length of 4 pixels and whose perimeter comes from the same distribution as the perimeter of the squares. The simulated data of the Boolean model were provided by the authors of [13] and [16].

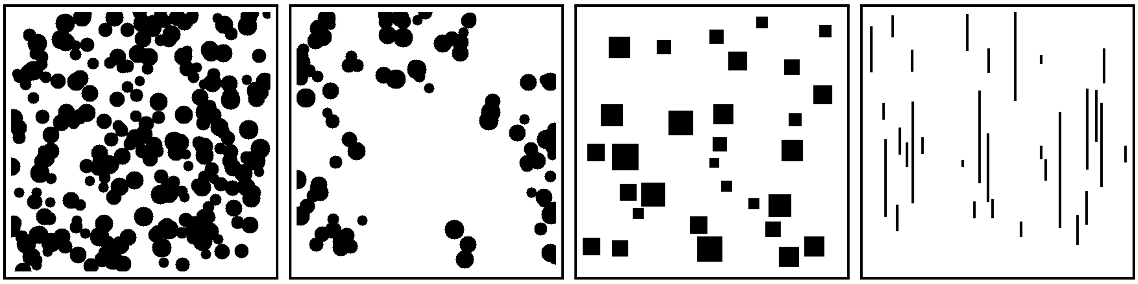


Figure 5.1: Examples of realisations of the Boolean, the reduced Boolean, the square and the rectangle models, respectively

Once we have the input data, we estimate the ratios and the curvatures. Consequently, we run the similarity test as described in the previous chapter. Firstly, we compare Boolean and reduced Boolean to show that the method does not distinguish between them because it is based on the similarity of components. Secondly, Boolean and square models are compared to show that the algorithm distinguishes between them due to the difference in the distribution of the curvature. Finally, squares and rectangles are compared to show that the procedure distinguishes them due to the difference in the ratios of the perimeter and the area of the components.

After running the test, we obtain 100 p-values (one for each tested pair). The histograms of the p-values are shown in Figure 5.2, where in the first column we have the histograms obtained by testing the equality of the ratios and the curvature, while in the second and the third column are the histograms of the p-values obtained by testing the equality of the ratios alone and the curvatures alone, respectively. The p-value should be interpreted as follows: if the p-value is close to zero, then the equality of the distributions is rejected. That means that if we test the similarity of realisations, then the p-value is uniformly distributed on $[0, 1]$ if the realisations come from the same model, while for different models p-value is close to zero.

The histograms of p-values shown in the first row in Figure 5.2 display approximately uniform distribution of p-values that we got comparing the Boolean vs the reduced Boolean model, so we can conclude that the method works as expected, i.e. it marks models that

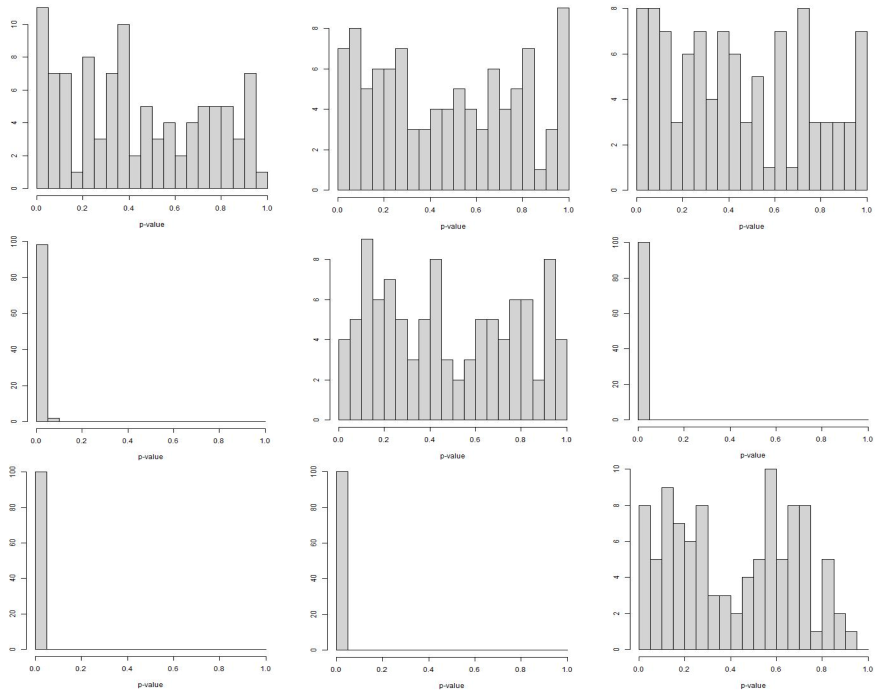


Figure 5.2: Histograms of p-values obtained by testing the Boolean model vs the reduced Boolean model (the first row), the Boolean model vs the square model (the second row) and the square morel vs the rectangle model (the third row), where in the first column are the results obtained using both the ratios (of the perimeter and the area) and the curvatures (of the boundary), in the second column using only ratios and in the third column using only curvatures

have components of the same shape as similar. However, when comparing the Boolean vs the square model, and the square vs the rectangle model, we can see that using only one characteristic, e.g. using the ratio for Boolean and square models, or using the curvature for square and rectangle models, we obtain erroneous results, i.e. histograms that suggest similarity. However, when comparing the curvatures of the first pair and the ratios of the second pair of models, we can see that p-values are close to zero. Thus, we can conclude that when comparing models that contain differently shaped components, we have to consider both characteristics.

In the second part of the simulation study, we compare the models that have already been studied by different authors in order to compare our results with previous work. The models studied are shown in Figure 5.3. The first model is the Boolean model, which is widely studied, the second and the third model are the cluster and the repulsive model (both are simulated using Quermass-interaction process as described in Definition 2.2.6 with suitably chosen parameters) studied in [16] and [9], [13], [16], [24], respectively. The last model is the Boolean ellipse model studied just in [16]. The simulated data were provided by the authors of [13] and [16].

Since we are working with models that significantly differ by the number of components

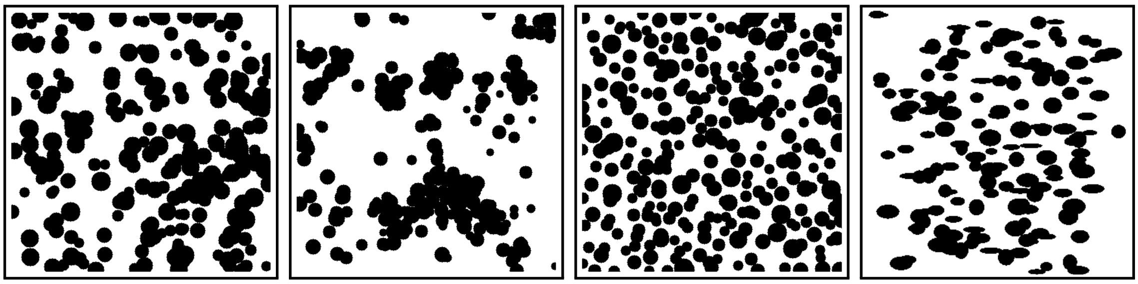


Figure 5.3: Previously studied models: the Boolean, the cluster, the repulsive and the ellipse model, respectively

in their respective realisations, we had to empirically determine the optimal number of components used for testing similarity. To obtain this information, we compared 100 pairs of the Boolean model using samples of different size. Histograms of p-values obtained using samples of 10, 20, 30 and 50 components are shown in Figure 5.4. We can see that p-values are approximately uniformly distributed for samples of size 10 and 20, because between sparsely sampled components we have a smaller correlation. Thus, we will use samples of size 10 and 20 for testing the aforementioned models.

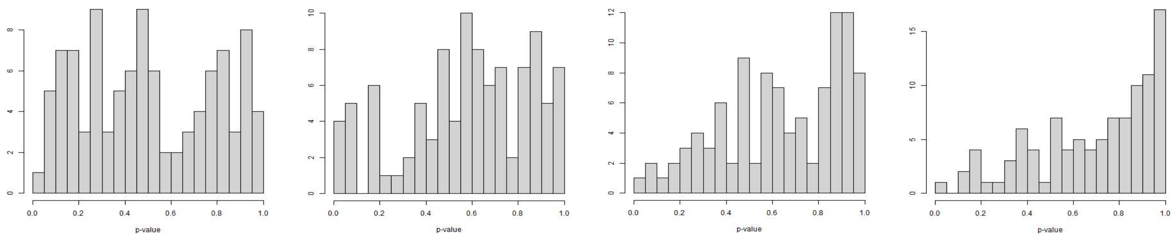


Figure 5.4: Histograms of p-values when testing pairs of the Boolean model using samples of 10, 20, 30 and 50 components from each realisation, respectively

For each model and for each pair of different models we test 100 pairs of realisations, as we already described above when we were testing illustrative models. The histograms of p-values for pairs coming from the same model (except for the Boolean model, which is shown in Figure 5.4) are shown in Figure 5.5.

From histograms in Figure 5.5 we can see that p-values are uniformly distributed for all cases except when comparing the pairs of samples of size 20 of the cluster model because in this case the sample is not scarce enough. Thus, we will use samples of size 10 for comparing different models.

The histograms of p-values obtained when testing different models with sample size 10 are shown in Figure 5.6. We can see that the rejection of similarity is not very convincing. Assuming that this error comes from the small number of components tested, we apply the bootstrap method, i.e. we randomly choose 100 components from the mix of all realisations of each model and consequently, we compare them. This way, we avoid dependence between components. After comparing 100 pairs obtained in this way, we

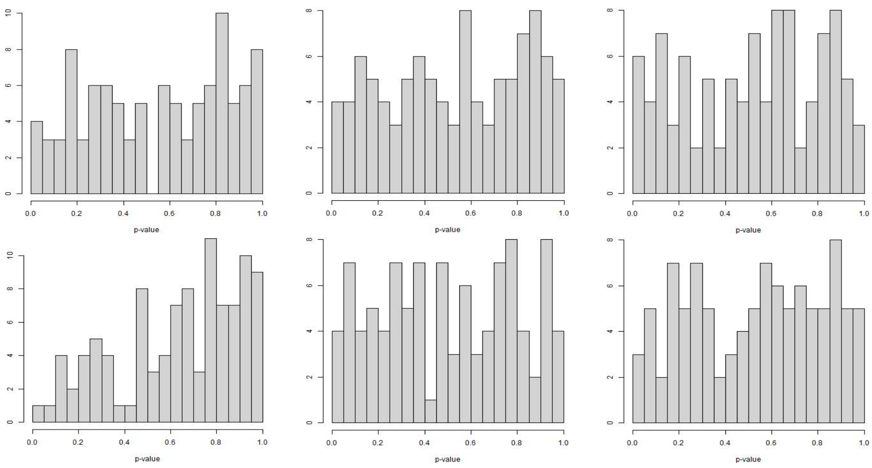


Figure 5.5: Histograms of p-values when testing pairs of realisations that come from the same model: cluster, repulsive, and ellipse, respectively with sample size 10 in the first, and 20 in the second row

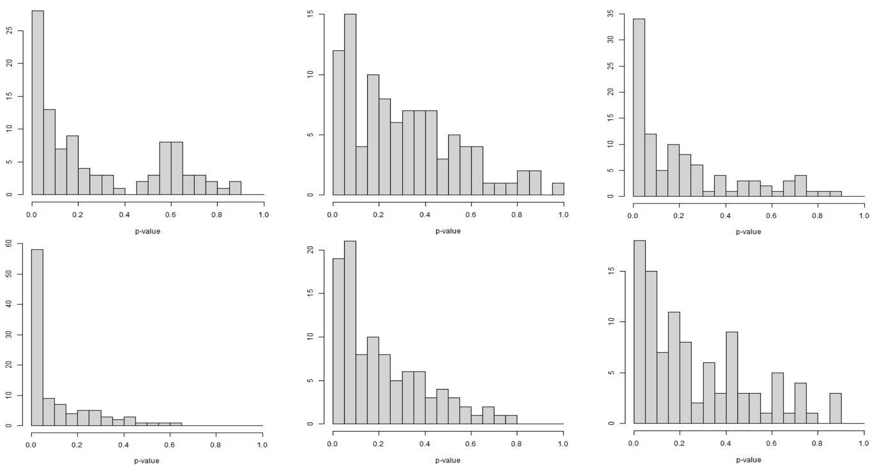


Figure 5.6: Histograms of p-values obtained when testing similarity of the Boolean model vs the repulsive model (upper left), the Boolean model vs the cluster model (upper middle), the repulsive model vs the cluster model (upper right), the ellipse model vs the Boolean model (lower left), the ellipse model vs repulsive model (lower middle) and the ellipse model vs the cluster model (lower right) using the samples of 10 components

construct histograms of p-values that are shown in 5.7. We can see that except for cluster and repulsive models, all p-values are smaller than 0.05. We assume that the reason behind bigger p-values for those models is the fact that both of them contain components that are made from isolated discs that come from the same distribution.

5.2 Real Data

Once we have shown that the procedure is able to measure (dis)similarity of random processes, we will apply it to the real data. The morphology of the tissue between the lactiferous duct system and mammary glands can indicate various types of benign or ma-

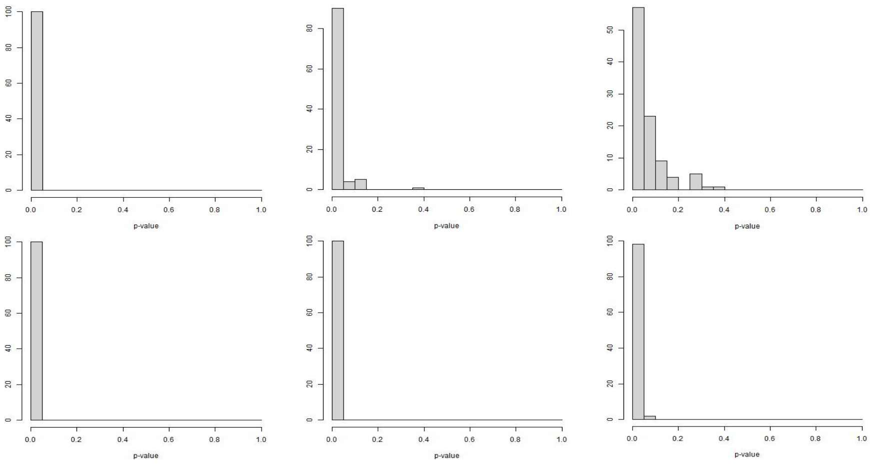


Figure 5.7: Histograms of p-values obtained when testing similarity of the Boolean model vs the repulsive model (upper left), the Boolean model vs the cluster model (upper middle), the repulsive model vs the cluster model (upper right), the ellipse model vs the Boolean model (lower left), the ellipse model vs repulsive model (lower middle) and the ellipse model vs the cluster model (lower right) using the bootstrap method and the samples of 100 components

lignant changes. In our study, we will consider two types of mammary tissue - mastopathic (referred to as Masto or "m" only from now on) and mammary cancer tissue (referred to as Mamca or "c" only). Note that this data has already been studied in [7] and [16]. Samples (in the form of binary images containing 10 sub-samples of size 512×512 representing cross-sections of the duct system), which are used in our study, are shown in Figure 5.8 and Figure 5.9, with black areas representing the aforementioned tissue. The data of mammary cancer and mastopathic tissue were kindly provided by the authors of [7] and edited in [16].

In order to test the similarity, we will apply the procedure in the same way as we applied it to the simulated data, i.e. we will first mark the components in the usual way and then evaluate the respective curvatures and ratios for both values of r. In the next step, we evaluate the p-values for all pairs of tissue samples (including the image with itself), where we consider samples of size 10 and 20 components.

The numbers of p-values below 0.05 obtained in this way for $R = 5$ and samples of size 20 are represented in Table 5.2. We can observe that the number of p-values below 0.05 is significantly lower when comparing pairs of the same type of tissue. Similarly, mean p-values represented in Table 5.2 are significantly lower for different types of tissue, which indicates that they are less similar than the pairs formed of the same type of tissue.

	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

Table 5.1: The number of p-values below .05 when comparing the corresponding samples 100 times. The values related to couples of different types of tissue are marked with italic font

	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 5.2: 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 marked with italic font

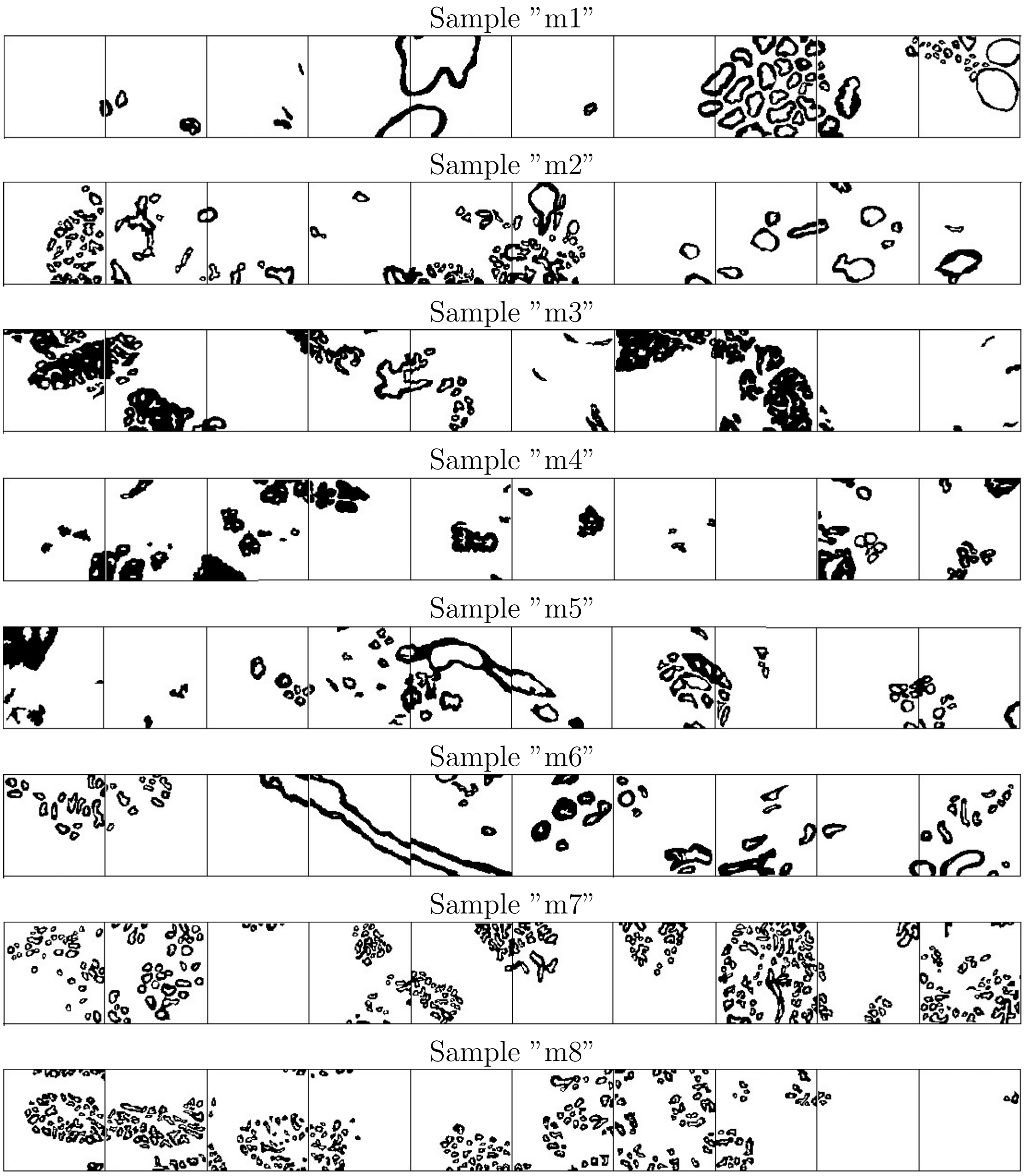


Figure 5.8: Samples of mastopathic breast tissue [7], [16]

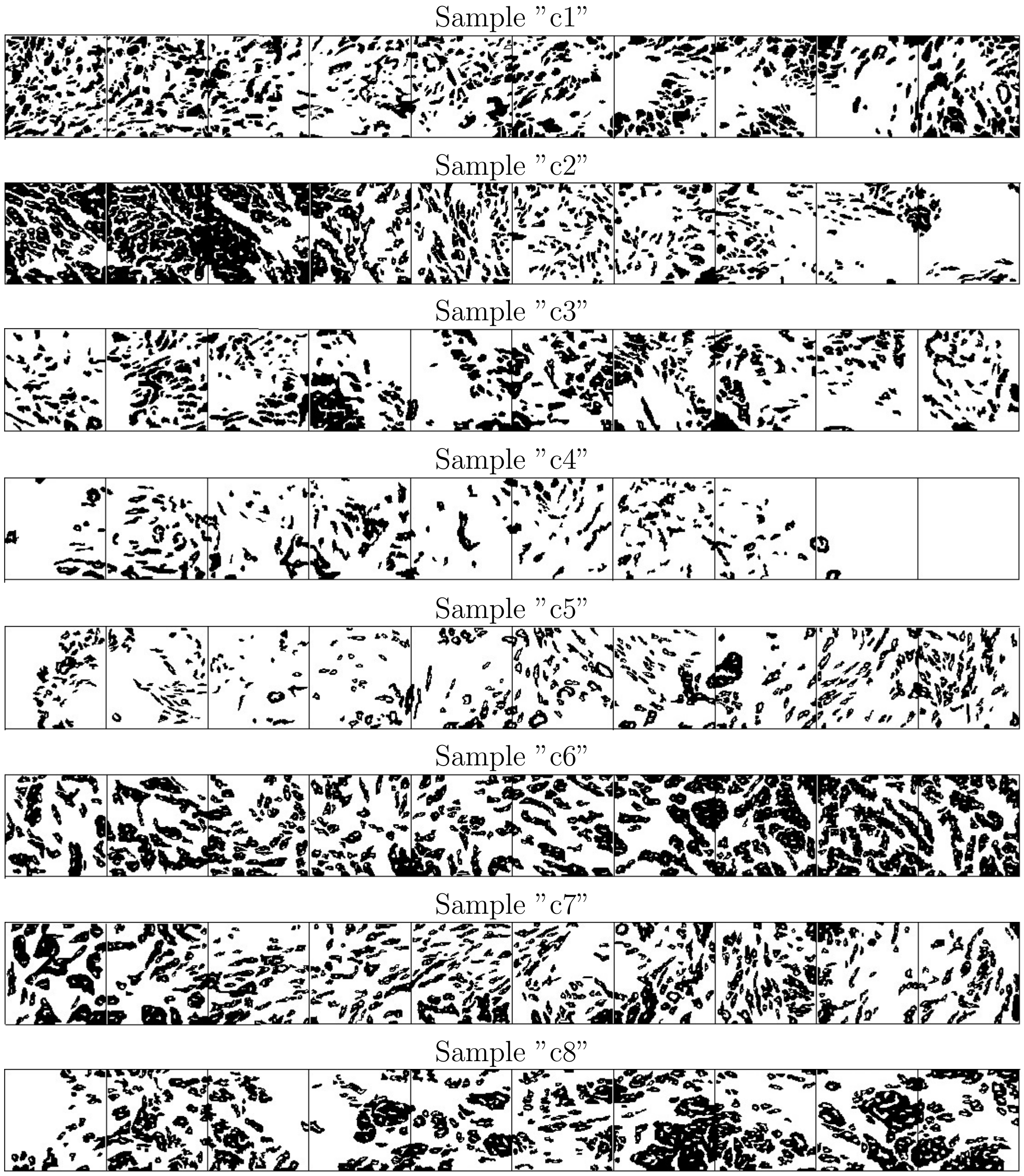


Figure 5.9: Samples of mammary cancer [7], [16]

Chapter 6

Comparison of Methods for Assessing Similarity of Random Sets

In this chapter, our main goal will be to compare the methods described in Chapter 3 with the method proposed in Chapter 4 and tested in Chapter 5. All the methods mentioned in this work are based on deriving a sample of values and/or functions for every realisation, which describes its characteristic features. Consequently, the equality of probability distributions of these functions is tested using tests described in Section 2.4. Note that both tests denote the tested realisations as similar if the testing functions come from the same distribution.

Many advantages and disadvantages of the considered methods have already been mentioned in previous chapters. However, for better orientation, we will summarise and compare them once more, this time putting stress on key similarities and differences. We will focus on the following questions:

- What does the given method distinguish between?
- What are the advantages of the given method?
- What are the disadvantages of the given method?
- How accurate is the given method?

Recall that some methods focus on the shape of the component, while others also consider their mutual positions. Both approaches are applicable in different situations, based on what do we want to achieve and what kind of data are we considering. The answer to the first abovementioned question is given in Table 6.1.

The advantages of the given methods are summarised in Table 6.2. Separate column is dedicated to advantages of different versions of the method (where they exist), based on the version of the test that was used.

<table>
<tr><th>Method</th><th>Test/version</th><th>Distinguishes between</th></tr>
<tr><td rowspan="2">Approximation by random convex compact covering</td><td>envelope (RCE)</td><td rowspan="2">shapes of components</td></tr>
<tr><td>$\mathcal{N}$-distance (RCN)</td></tr>
<tr><td rowspan="2">Skeletons</td><td>envelope (SE)</td><td rowspan="2">positions and shapes of components</td></tr>
<tr><td>$\mathcal{N}$-distance (SN)</td></tr>
<tr><td rowspan="3">Connected components and neighbourhoods tessellations</td><td>$\mathcal{N}$-distance; connected components (TCC)</td><td>shapes of components</td></tr>
<tr><td>$\mathcal{N}$-distance; neighbourhoods (TN)</td><td rowspan="2">positions and shapes of components</td></tr>
<tr><td>$\mathcal{N}$-distance; both (TCCN)</td></tr>
<tr><td>Two-step procedure</td><td>$\mathcal{N}$-distance (2S)</td><td>shapes of components</td></tr>
</table>

Table 6.1: Distinctive aspects of the considered methods

<table>
<tr><th>Method</th><th>Test/version</th><th colspan="2">Advantages</th></tr>
<tr><td rowspan="2">Approximation by random convex compact covering</td><td>envelope (RCE)</td><td rowspan="2">simple interpretation of testing functions</td><td>low time-consumption (w.r.t. RCN)</td></tr>
<tr><td>$\mathcal{N}$-distance (RCN)</td><td>accuracy (w.r.t. RCE)</td></tr>
<tr><td rowspan="2">Skeletons</td><td>envelope (SE)</td><td rowspan="2">high accuracy, no random approximation</td><td>low time-consumption</td></tr>
<tr><td>$\mathcal{N}$-distance (SN)</td><td>the highest accuracy in general</td></tr>
<tr><td rowspan="3">Connected components and neighbourhoods tessellations</td><td>$\mathcal{N}$-distance; connected components (TCC)</td><td rowspan="3">simple interpretation, flexibility (specific objects of interest can be considered), no random approximation</td><td>low time-consumption (no neighbourhood construction needed)</td></tr>
<tr><td>$\mathcal{N}$-distance; neighbourhoods (TN)</td><td rowspan="2">neighbourhoods of components taken into account</td></tr>
<tr><td>$\mathcal{N}$-distance; both (TCCN)</td></tr>
<tr><td>Two-step procedure</td><td>$\mathcal{N}$-distance (2S)</td><td colspan="2">simple and straightforward interpretation, no random approximation, flexibility (specific objects of interest can be considered), high accuracy</td></tr>
</table>

Table 6.2: Advantages of the considered methods

Method also has disadvantages, which were addressed in greater detail in previous Chapters. They are, similarly to disadvantages, summarised in Table 6.3. A separate column is again dedicated to the advantages of different versions of the method (where they exist), based on the version of the test that was used.

<table>
<tr><th>Method</th><th>Test/version</th><th colspan="2">Disadvantages</th></tr>
<tr><td rowspan="2">Approximation by random convex compact covering</td><td>envelope (RCE)</td><td rowspan="2">random covering, choice of optimal radius, influence of rotation, heuristic approach</td><td>very low accuracy in general</td></tr>
<tr><td>$\mathcal{N}$-distance (RCN)</td><td>high time-consumption in general</td></tr>
<tr><td rowspan="2">Skeletons</td><td>envelope (SE)</td><td rowspan="2">choice of input parameters</td><td>slightly lower accuracy (w.r.t. SN)</td></tr>
<tr><td>$\mathcal{N}$-distance (SN)</td><td>slightly higher time-consumption (w.r.t. SE)</td></tr>
<tr><td rowspan="3">Connected components and neighbourhoods tessellations</td><td>$\mathcal{N}$-distance; connected components (TCC)</td><td rowspan="3">request of many components</td><td>very low accuracy</td></tr>
<tr><td>$\mathcal{N}$-distance; neighbourhoods (TN)</td><td rowspan="2">omission of some con.comp. from TCC (neighbourhood edge effects)</td></tr>
<tr><td>$\mathcal{N}$-distance; both (TCCN)</td></tr>
<tr><td>Two-step procedure</td><td>$\mathcal{N}$-distance (2S)</td><td colspan="2">possible dependence of components</td></tr>
</table>

Table 6.3: Disadvantages of the considered methods

From results of simulation studies, which were summarised in Chapters 3 and 5, we can conclude that all the methods exhibit the same accuracy when realisations of the same model are compared, i.e. respective p-values obtained in this way are close to 0. However, inaccuracies appear when comparing realisations from different models. These inaccuracies are caused by different factors, which were studied in the abovementioned chapters and are closely related to the disadvantages listed in Table 6.3. For example, histograms of p-values that were obtained when comparing different models are presented in Figure 6.1. We can see that the highest accuracy is acquired using the second method.

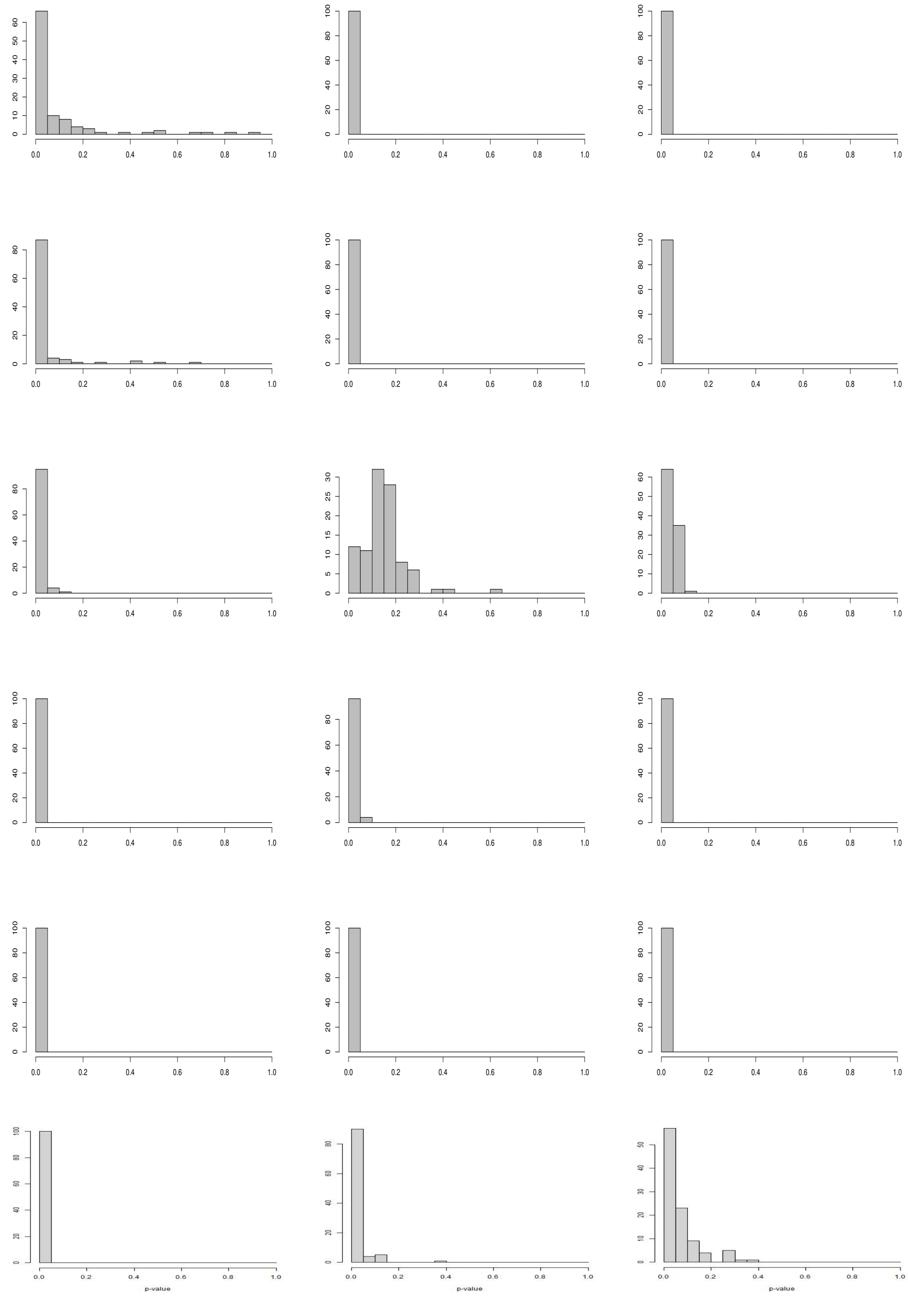


Figure 6.1: Histograms of p-values when comparing Boolean vs repulsive (left), Boolean vs cluster (middle) and repulsive vs cluster (right) models using RCE (the first row), RCN (the second row), TCC (the third row), TN (the fourth row), TCCN, SE and SN (identical histograms in the fifth row) and 2S (the sixth row)

Chapter 7

Conclusion

The first goal of this thesis was to summarise already existing results in the field of testing similarity of random sets. The main focus of this work was on methods devised in the last five years, because if we exclude traditional methods of comparing random sets, the field is relatively new and still developing. For better understanding of the described algorithms, an introduction to *stochastic geometry* had to be made in Chapter 2. The most notable papers, namely [13], [24] and [16] were reviewed in Chapter 3. The approaches presented in these papers are either not applicable in all cases or they are too sensitive to small changes in certain parameters. For that reason we drew a conclusion that new methods have to be proposed, which was our second goal.

In Chapter 4 we constructed a new two-step method for assessing similarity of random sets. The method is based on evaluating the curvature measure at the points of the boundary and evaluating the ratio of the perimeter and the area of components. Components that are examined are isolated for the purpose of minimising the weak points mentioned above, so the first step in our algorithm is to isolate the components. The next step is to evaluate chosen features, namely boundary curvature (at every point of the boundary) and ratio of perimeter and area for each component. After that we construct appropriate kernel that is used for estimating the $\mathcal{N}$-distance, which is later used as test statistic. Finally, we run the Monte Carlo permutation test.

In Chapter 5, we validated the procedure by applying it to simulated data. From the histograms of p-values, we conclude that the method is working in the expected way. Of course, the method exhibits some limitations. The first one is dependence on density of the components, because densely packed components can affect shapes of some surrounding components. This behaviour is existent because we ignored the correlation between the individual components, which is clearly present. In many cases this mistake can be eliminated by randomly choosing a sample of components. We took this fact in consideration when testing the similarity of models that were studied by different authors and in the second part of our simulation study. The histograms of p-values showed that our method is able to tell if the realisations have similar components in the sense of Definition 4.3.2. However, due to the small sample of components, there were problems that the similarity of different models is not rejected as often as we would like. This error can be compensated using bootstrapping, i.e. taking a random sample of components from the mix of all realisations of examined model. In this way the correlation between components can be significantly reduced. Similarly to other methods, one of the disadvantages is also obligation to choose the size (and shape) of the circle that is used for evaluating curvature.

As a final step, we applied the procedure to the samples representing different types of mammary tissue (mastopathic and mammary cancer tissue). The results are satisfactory when we take into account all the difficulties coming from the variation in shapes and size of components for the same type of tissue, and problems with identifying characteristic features for different types of tissues.

In the sixth chapter, we compared our method with the previous methods in the key aspects. We conclude that the new method has numerous advantages, namely flexibility, straightforwardness and high accuracy.

Note that the presented method and results have been already uploaded to arxiv.org, see [31].

In the future, when concerning its application in practise, especially on data which contains multiply connected components, the method could be improved by finding a way to differentiate between inner and outer boundaries. Furthermore, the algorithm for estimating curvature could also be improved by deriving the optimal value of the radius using machine learning. The method also shows the potential to be used as a tool for classification of realisations of random sets.

Bibliography

1. Diggle, P. J. & Milne, R. K. Bivariate Cox Processes: Some Models for Bivariate Spatial Point Patterns. *Journal of the Royal Statistical Society. Series B (Methodological)* **45,** 11–21. ISSN: 00359246. http://www.jstor.org/stable/2345617 (1983).
2. Grabarnik, P., Pagès, L. & Bengough, A. Geometrical properties of simulated maize root systems: Consequences for length density and intersection density. *Plant and Soil* **200,** 157–167 (Mar. 1998).
3. Kadashevich, I., Schneider, H.-J. & Stoyan, D. Statistical modeling of the geometrical structure of the system of artificial air pores in autoclaved aerated concrete. *Cement and Concrete Research* **35,** 1495–1502 (Aug. 2005).
4. Illian, J., Penttinen, A., Stoyan, H. & Stoyan, D. *Statistical Analysis and Modelling of Spatial Point Patterns* English. ISBN: 978-0-470-01491-2 (John Wiley and Sons, Chichester, United Kingdom, 2008).
5. Baddeley, A. & Jensen, E. *Stereology for Statisticians* ISBN: 9780203496817. https://books.google.cz/books?id=il0fXb%5C_GSowC (CRC Press, 2004).
6. Chiu, S., Stoyan, D., Kendall, W. & Mecke, J. *Stochastic Geometry and Its Applications* ISBN: 9780470664810 (John Wiley and Sons, Chichester, United Kingdom, Sept. 2013).
7. Mrkvička, T. & Mattfeldt, T. Testing histological images of mammary tissues on compatibility with the Boolean model of random sets. *Image Analysis and Stereology* **30,** 101–108 (Mar. 2011).
8. Hermann, P. *et al.* Fractal and stochastic geometry inference for breast cancer: a case study with random fractal models and Quermass-interaction process. *Statistics in medicine* **34,** 2636–2661 (Apr. 2015).
9. Gotovac, V. & Helisová, K. Testing Equality of Distributions of Random Convex Compact Sets via Theory of N-Distances. *Methodology and Computing in Applied Probability.* https://doi.org/10.1007/s11009-019-09747-z (2021+).

10. Baccelli, F. & Blaszczyszyn, B. Stochastic Geometry and Wireless Networks: Volume I Theory. *Foundations and Trends in Networking* **3,** 249–449 (Jan. 2009).

11. Muscat, J. & Buhagiar, D. Connective spaces. *Series B: Mathematical Science* **39,** 1–13 (Jan. 2006).

12. Barbati, A. & Hess, C. The Largest Class of Closed Convex Valued Multifunctions for which Effros Measurability and Scalar Measurability Coincide. *Set-Valued Analysis* **6,** 209–236 (Jan. 1998).

13. Gotovac, V., Helisová, K. & Ugrina, I. Assessing Dissimilarity of Random Sets Through Convex Compact Approximations, Support Functions and Envlope Tests. *Image Analysis & Stereology* **35,** 181–193 (Dec. 2016).

14. Lavie, M. Characteristic function for Random Sets and Convergence of Sums of Independent Random Sets. *Acta Mathematica Vietnamica* **25,** 87–99 (Jan. 2000).

15. Keeler, P. *Thinning point processes* [Accessed on 28.03.2021]. Nov. 2020. `https://hpaulkeeler.com/thinning-point-processes/`.

16. Gotovac, V. Similarity Between Random Sets Consisting of Many Components. *Image Analysis & Stereology* **38,** 185–199 (July 2019).

17. Gonzalez, R. C. & Woods, R. E. *Digital Image Processing* 2nd. ISBN: 0201180758 (Addison-Wesley Longman Publishing Co., Inc., USA, 2001).

18. Sezgin, M. & Sankur, B. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging* **13,** 146–168 (Jan. 2004).

19. Bankman, I. N. in *Handbook of Medical Image Processing and Analysis (Second Edition)* (ed BANKMAN, I. N.) Second Edition, 71–72 (Academic Press, Burlington, 2009). ISBN: 978-0-12-373904-9. `https://www.sciencedirect.com/science/article/pii/B978012373904950012X`.

20. *Virtual LAB in Image Processing* [Accessed on 20.3.2021]. Initiative by Ministry of Education of India [Online], 2018. `https://cse19-iiith.vlabs.ac.in/theory.php?exp=morph`.

21. Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H. & Hahn, U. Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79,** 381–404. ISSN: 1369-7412. `http://dx.doi.org/10.1111/rssb.12172` (Mar. 2016).

22. Klebanov, L. *N-distances and Their Applications* ISBN: 80-246-1152-X (Karolinum Press, Charles University, Prague, Jan. 2005).

23. Davison, A. & Hinkley, D. Bootstrap Methods and Their Application. *Journal of the American Statistical Association* **94** (Jan. 1997).

24. Debayle, J., Gotovac, V., Helisová, K., Staněk, J. & Zikmundová, M. Assessing Similarity of Random sets via Skeletons. *Methodology and Computing in Applied Probability.* `https://doi.org/10.1007/s11009-020-09785-y` (Mar. 2020).

25. Ebeida, M. S. *et al. Efficient Maximal Poisson-Disk Sampling* in *ACM SIGGRAPH 2011 Papers* (Association for Computing Machinery, Vancouver, British Columbia, Canada, 2011). ISBN: 9781450309431. `https://doi.org/10.1145/1964921.1964944`.

26. Møller, J. & Helisová, K. Power diagrams and interaction process for unions of discs. *Advances in Applied Probability* **40,** 321–347 (June 2008).

27. Teichmann, J., Ballani, F. & van den Boogaart, K. Generalizations of Matérn's hard-core point processes. *Spatial Statistics* **3,** 33–53. ISSN: 2211-6753. `https://www.sciencedirect.com/science/article/pii/S2211675313000043` (2013).

28. Bullard, J., Garboczi, E., Carter, W. & Fuller, E. Numerical methods for computing interfacial mean curvature. *Computational Materials Science* **4,** 103–116. ISSN: 0927-0256. `https://www.sciencedirect.com/science/article/pii/092702569500014H` (1995).

29. Fraser, W. & Gotlieb, C. C. A calculation of the number of lattice points in the circle and sphere. *Mathematics of Computation* **16,** 282–282. `https://doi.org/10.1090/s0025-5718-1962-0155788-9` (Sept. 1962).

30. Sloane, N. J. A. *The On-Line Encyclopedia of Integer Sequences® (OEIS®)* [Accessed on 05.03.2021]. Apr. 1991. `https://oeis.org/A000328/b000328.txt`.

31. Radović, B. *et al. Two-step method for assessing dissimilarity of random sets* 2021. arXiv: `2105.05952 [stat.ME]`.

Contents of Enclosed CD

- readme.txt ... the file with CD contents description
- Inputs ... the directory with simulated input data
  - readme.txt ... the file with contents description
- Outputs .. the directory with output files, i.e. calculated ratios of perimeter and area and respective curvatures* used for testing similarity
  - readme.txt ... the file with contents description
- Programs ... the directory with source codes
  - readme.txt ... the file with contents description
- Results ... the directory with files containing p-values obtained from testing
  - readme.txt ... the file with contents description
- Thesis ... the directory of LaTeX source codes of the thesis
- thesis.pdf ... the thesis text in PDF format