VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

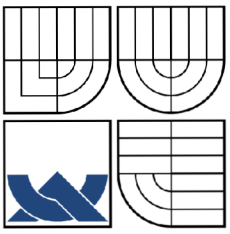
STOCHASTICKÉ MODELOVÁNÍ KOMPOZITNÍCH MATERIÁLŮ

DIZERTAČNÍ PRÁCE
DOCTORAL THESIS

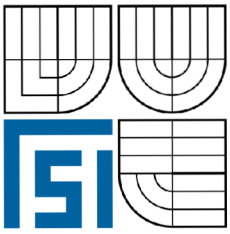
AUTOR PRÁCE
AUTHOR

Ing. TOMÁŠ POSPÍŠIL

BRNO 2010



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY
FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

STOCHASTICKÉ MODELOVÁNÍ KOMPOZITNÍCH MATERIÁLŮ

STOCHASTIC MODELING OF COMPOSITE MATERIALS

DISERTAČNí PRÁCE
DOCTORAL THESIS

AUTOR PRÁCE
AUTHOR

ING. TOMÁŠ POSPÍŠIL

VEDOUCÍ PRÁCE
SUPERVISOR

prof. RNDr. JAN FRANCŮ, CSc.

BRNO 2010

Abstrakt

Předkládaná disertační práce se věnuje generování náhodných struktur dvouvláknových kompozitních materiálů. První část se zabývá známými obecnými principy a zákonitostmi náhodných procesů. Celá úvodní část je směřována k aplikaci náhodných procesů na kompozitní materiály jako je např. anizotropie nebo prostorová korelace. Jsou zde uvedeny základní a nejpoužívanější známé modely pro generování náhodných struktur. Dále je pak diskutována otázka popisu vnitřní struktury kompozitů, zejména pak kompletní prostorové náhodnosti struktur a její detekce různými metodami. Teoretickou část pak uzavírá detailní popis autorem vytvořených čtyř algoritmů pro generování náhodných struktur s nekonstantním průměrem vláken ve vzorku.

Ve druhé, výpočtové části je uvedeno porovnání nasimulovaných vzorků pomocí nových algoritmů navzájem mezi sebou a s reálnými vzorky, které byly k dispozici. Toto porovnání je provedeno metodami deskriptivní statistiky. V neposlední řadě jsou ověřeny předpoklady normality a homogenity rozptylu u jednotlivých vzorků. Tyto předpoklady jsou zpravidla nezbytné pro případné další zpracování dat, např. analýzy rozptylu.

Summary

This thesis is devoted to generating of non-periodic structures of two-fibre composite material. The first part deals with the well-known principles and laws of random processes. The whole introductory part tends to the application of random processes to the composites, e.g. anizotropy or spatial correlation. The most frequently used and well-known algorithms for generating non-periodic patterns are presented here. Next, the description of inner microstructure is discussed together with the methods of detection of complete spatial randomness. The theoretic part ends with detailed description of four algorithms developed by the author for generating random structures with non-constant diameters of fibres.

In the second computational part the comparison of simulated samples obtained by new algorithms and real ones is presented. This comparison is made by mean of techniques of a descriptive statistic. Moreover, the assumptions of normality and homogeneity of samples are checked. These assumptions are usually necessary for contingent next computations, e.g. analysis of variance.

Klíčová slova

Neperiodické struktury, Náhodné procesy, Normalita, Homogenita.

Keywords

Non-periodic structures, Random processes, Normality, Homogeneity.

POSPÍŠIL, T. *Stochastické modelování kompozitních materiálů.* Brno: Vysoké učení technické v Brně, FAKULTA STROJNÍHO INŽENÝRSTVÍ, 2010.

Vedoucí disertační práce prof. RNDr. Jan Franců, CSc.

I declare, that I processed this thesis myself according to the instructions of my advisor and with using a cited literature.

ING. TOMÁŠ POSPÍŠIL

Acknowledgements

First of all, I would like to express my gratitude to my supervisor and good, educate teacher and friend during all my studies from the year 1998. His name is prof. RNDr. Jan Franců, CSc. He was not only the supervisor of my thesis but also a very good teacher and last but not least my friend during the course of all my studies since 1998. I am also thankful to my very good friend and teacher, doc. RNDr. Zdeněk Karpíšek, CSc. for his guidance in mathematical statistics and many fruitful and inspiring discussions on topics concerning not only this thesis.

I would also like to express many thankful words to doc. RNDr. Jaroslav Michálek, CSc. He was the man, who gave me the first guidance from the probability and mathematical statistics in the 4^{th} semester of my basic study at FME BUT. It would be unpardonable not to mention doc. RNDr. Bohumil Maroš, CSc. and RNDr. Pavel Popela, Ph.D., for their never-ending interest in me. Last but not least, I have to express many thanks to prof. RNDr. Miroslav Doupovec, CSc. and prof. RNDr. Petr Dub, CSc. for their optimistic approach to the life.

Now, I would like to thank to prof. RNDr. Alexander Ženíšek, DrSc., who gave me the space to improve my knowledge in functional analysis, and also to my very best teacher, prof. RNDr. Jozef Kačur, DrSc., who gave me the right reason why to study a functional analysis.

Of course, I am grateful to my dear mother and sister for their constant and inexhaustible patience and love during my whole life. They were the people, who always patronized my ideas and helped me to realize everything. Without them this thesis would never come into existence (literally).

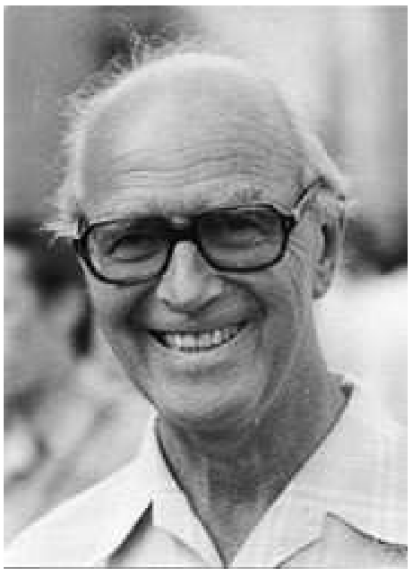
I would also like to express many thanks to ABBA the pop-group for their great and never-dying songs that accompanied me many times during my work on this thesis and gave me also a lot of energy to continue and not to give it up.

Finally, I have to mention my dear father, RNDr. Richard Pospíšil, CSc. (21.8.1935–25.3.1999), who has been surely looking at me from Heaven...

Brno, Czech Republic — Ing. Tomáš Pospíšil
29th August 2010

This research was supported by Grant No. 201/08/0874 of the Grant Agency of the Czech Republic and by Research Project No. 1M06047 of MŠMT of the Czech Republic.

Preface



Motto: Člověk se nenaučí dělat matematiku *poslouchánim* vybroušených výkladů při vyučovacích hodinách, ale zejména *samostatnou prací* s matematickými pojmy.

GUSTAVE CHOQUET

Table of Contents

List of Figures

List of Tables

Notation

$\mathbb{R}$	real numbers
$\mathbf{E}\left[\cdot\right]$	expected (mean) value
$\mathbf{D}\left[\cdot\right]$	dispersion (variation)
$\mathbf{P}\left(\cdot\right)$	probability
$\mathrm{N}(\cdot,\cdot)$	normal distribution
$\left\vert\cdot\right\vert$	absolute value
$\left\vert\left\vert\cdot\right\vert\right\vert$	Euclidean distance
$\{\cdot\}$	sequence of values
$\mathcal{I}^{(r)}(\cdot)$	indicator function for phase r
$\sharp(\cdot)$	counting function
$\langle\cdot\rangle$	ensemble average
ϕ_r	volume fraction of a phase r
$\Phi(\cdot)$	cumulative distribution function
$\hat{F}(\cdot)$	estimation of a function $F(\cdot)$
$\mathbf{AI}-\mathbf{AIV}$	algorithms **AI**–**AIV**

Introduction

The study of composite materials has become a very important subject of research in the materials engineering area. These materials are heterogeneous man-made mixtures of two or more homogeneous phases bonded together. The first phase is called the *matrix* and is usually a metal (e.g. aluminium, steel, titanium) or plastic (e.g. silicon, epoxide), while the other is the reinforcement and is commonly either *particles* or *fibres*. In general, the second-phase substance has much higher stiffness than the matrix.

It is expected that by combining two types of materials one will obtain the best properties of both substances. This material, the so-called *composite material*, has then considerably better mechanical properties and higher performance than any single material from which it is formed.

It is known that mechanical properties (e.g. ductility and fracture toughness) of particular composite materials depend not only on the shape and volume fraction of the composites but also on the spatial and size distributions of the particles or fibres. Moreover, variations in the production process (e.g. rolling, extrusion, centrifuging, temperature of the mother matrix when the particles are added) can affect the mechanical properties of the material.

This evidence shows that quantitative analysis of the microstructure of particulate composite materials is of extreme importance for a better understanding of the relationship between inclusions and mechanical behavior and also for better control of the production of the material.

Material scientists are primarily interested in relating the mechanical properties of the composite to the microstructural features of the second-phase particles such as volume fraction, size, shape and spatial variation. There is only one possible way to achieve this aim when composites are concerned and this is to take a planar section of the material, polish it and then record, with the use of a microscope, the features of interest (e.g. location, size, shape) of each particle that appears in the section. This information is then processed by image analysis techniques, which are described e.g. in [12]. These techniques provide a large amount of data on the reinforced material that must be analyzed by statistical methods. In particular, the currently fast growing area of spatial processes has special relevance to the analysis.

The ultimate aim of investigating the statistical properties of patterns of composite materials is to get some information and additional insight about the underlying mechanisms that rule the way the different materials are formed. Since there are many features relating to the material that can be analyzed, it is appropriate to say that this thesis deals with the statistical description and analysis of spatial distributions of fibres held in planar sections of composite materials.

Mathematical modeling of composite materials leads to solving PDEs with strongly oscillating coefficients. The problem of large number of equations can be solved using

homogenization, that replaces heterogeneous material by an equivalent homogeneous one. This approach assumes periodic structure, which is not true in the reality.

Many methods devoting to the composites and their mathematical description of their physical properties relies on the fact, that the structure is well known. Many materials, meaning two-fibres composites, vary very "widely" than to be imposed to be the sample for mathematical modeling. Such structures can be seen in the following chapters in this thesis.

The predictions of properties of a real random structure of a natural material is a priory very difficult because of the amount of the effects that we are able to hold. And this is the reason, we are still not able to exactly predict a behavior of such material.

One approach how to understand this phenomena is to understand to its inner structure. Once, we have at disposal real samples of a real media in the form of photographs or bitmaps, we are able to simulate very similar patterns to the real ones.

In this thesis we give a brief summary of standard methods dealing with describing and comparing of various random patterns. We introduce here a collection of methods for describing a various random material. In literature we can find many algorithms and methods for generating random patterns (meaning cross-sections of two-phase fibre composites), see e.g. [8] or [9], but the only disadvantage of this is the fact, that they operate only with the constant diameters of the fibres. Applying these methods we admittedly obtain random samples, but such patterns do not correspond to the real ones because of the diameters. In this theses we summarize the basic descriptors for random samples and introduce four algorithms for generating non-periodic structures with non-constant diameters. Such obtained samples will be consequently compared with real samples by means of descriptive statistic techniques and standard microstructural descriptors.

The content of the thesis will be as follows: The theses consists of three parts. The first part, called *Theoretical Aspects*, is divided into four sections. The first section, *Models for Random Multi-phase Materials*, is devoted to the common description of random spatial processes, their definition and properties. Moreover, the basic models for generating random patterns are presented here. The second one gives a brief summarization about complete spatial randomness of samples and their description using second order Ripley's K function. In the third section the microstructural descriptors of composites are presented and in the four one the new algorithms **AI**–**AIV** generating non-periodic patterns with non-constant fibre diameters are introduced.

The second part of the thesis, called *Statistical Computations*, contains the computations dealing with developed algorithms **AI**–**AIV** and their comparing with the real samples. At the beginning the real samples are described, then the methods of descriptive statistics are applied on these simulated samples and the question of anizotropy is discussed. At the end of this part the basic assumptions for further analysis, such as normality and homogeneity, are presented.

Part I

Theoretical Aspects

The winner takes it all...

1 Models for Random Multi-phase Materials

1.1 Introduction

The study of micromechanics in composite materials has been performed assuming periodicity in the distribution of the fibres. This approach provides simplifications which lead to the possibility of analytical solutions or in the case of computational methods it reduces its time. However, this approach represents an idealized material which may be useful for computing effective elastic properties, but differs from the real one in some aspects, see [20].

It has been shown that avoiding real, i.e. non-periodic distribution of fibres have not so large negative effects to effective properties of material, but local (e.g. mechanical) properties vary very intensive, see [28].

1.2 Random process – properties and definitions

The use of term *random heterogeneous material* or simply *random medium* rests on the assumption that any sample of the medium is a realization of a specific random or stochastic process (or random field). An *ensemble* is a collection of all the possible realizations of a random medium generated by a specific stochastic process, see [37]. We let $(\Omega, \mathcal{F}, \mathbf{P})$ be some fixed *probability space,* where Ω is a sample space, $\mathcal{F}$ is a σ-algebra of subsets of Ω(set of events), and $\mathbf{P}$ is a probability measure.

1.2.1 Random process

When we analyze real materials in a microscopic scale, there exist many variables which should be considered random, and which depend on spatial distributions of phases.

Let $\boldsymbol{x} \in \mathbb{R}^d$ be a spatial location in a $d-$dimensional space and let us assume $Z(\boldsymbol{x})$ is a random variable. If we let $\boldsymbol{x}$ vary over a fixed set $D \subset \mathbb{R}^d$, we can express the *random process* $Z(\boldsymbol{x})$ as, see [8], [20]:

$$\{Z(\boldsymbol{x}) : \boldsymbol{x} \in D\}.$$

To emphasize the source of randomness, the previous notation is sometimes written as

$$\{Z(\boldsymbol{x}, \omega) : \boldsymbol{x} \in D;\ \omega \in \Omega\},$$

where $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space. If $\boldsymbol{x} \in \mathbb{R}$ (i.e., the variable is function of one spatial dimension), the term *random process* or *stochastic process* is often used instead of *random field* or *random function* as in the case of $d > 1$.

1.2.2 Types of random processes

According to [8], usually D is assumed to be a fixed, i.e. nonrandom subset of $\mathbb{R}^d$, but we shall assume more generally that D is a *random set*. Roughly formally speaking, we shall assume that D as well as Z may vary from realization to realization, giving another source of randomness to the problem, see [8] for details. Generally, depending on the nature of the set D, four types of random processes can be defined:

- *Time-space processes* – are processes which variation is given in space D and time interval $\langle 0; T\rangle$. This can be written:

$$\{Z(\boldsymbol{x}; t) : \boldsymbol{x} \in D,\ t \in \langle 0; T\rangle\}.$$

  The special case of time-space processes are the so called *time series*, in which the set D is the temporal dimension. Usually the fatique behaviour of composite materials and mechanical properties are modelled using time series.

- *Geostatistical data* – when the spatial variable $\boldsymbol{x}$ varies continuously within D, which is a subset of $\mathbb{R}^d$ and $Z(\boldsymbol{x})$ is a random vector at location $\boldsymbol{x} \in D$. Here, measurements are taken at a fixed number of chosen locations. Most of the physical properties can be seen as geostatistical data.

- *Lattice data* – when D is a fixed(regular or irregular) collection of countable many points of $\mathbb{R}^d$ and $Z(\boldsymbol{x})$ is a random vector at location $\boldsymbol{x} \in D$. Here, measurements are taken at a lattice and at each point on this lattice a measurement is collected. Sometimes, measured properties are computed as mean values. Sometimes, like it happens in finite element meshes, the same value of the property is considered for a subdomain (the element of the lattice). In this case we are working with lattice data, see figure 1.1.

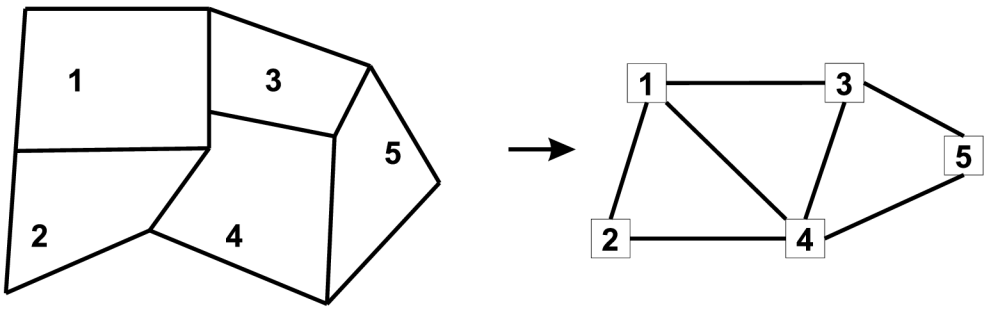


Figure 1.1: *An averaging example of the lattice data*

- *Point patterns* – data in the form of a set of points, regularly or irregularly distributed within a region. Each item of data consists of the location of an event. The random position of e.g. carbon or glass fibres in a fibre reinforced composite is a good example of a point pattern. Point pattern analysis is concerned with the location of events, and with answering questions about the distribution of those locations, specifically whether they are clustered, randomly or regularly distributed.

Point pattern analysis is very sensitive to the definition of the study area, since a regularly distributed pattern can be made to seem clustered by including large margins within the study area.

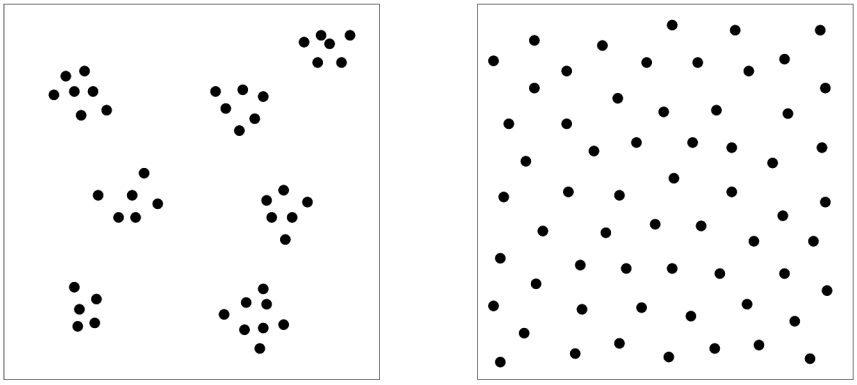


Figure 1.2: *Clustered point pattern on the left and regular point pattern on the right*

1.2.3 Characteristics of random processes

Among basic numeric characteristics of a random process belong **expected(mean) value** $\mathbf{E}(Z)$ and **dispersion(variance)** $\mathbf{D}(Z)$ of the stochastic process $Z(\boldsymbol{x})$. But they are so known terms in the theory of probability and statistics, that it is meaningless to mention them here. Among next important characteristics belong next ones

Autocovariance function $C(\boldsymbol{x},\boldsymbol{y})$ of $Z(\boldsymbol{x})$, *where* $(\boldsymbol{x},\boldsymbol{y}) \in D \times D$ is defined as

$$C(\boldsymbol{x},\boldsymbol{y}) = \operatorname{cov}[Z(\boldsymbol{x}), Z(\boldsymbol{y})] = \mathbf{E}\left\{[Z(\boldsymbol{x}) - \mathbf{E}(Z(\boldsymbol{x}))][Z(\boldsymbol{y}) - \mathbf{E}(Z(\boldsymbol{y}))]\right\}$$

and has these properties

$$C(\boldsymbol{x},\boldsymbol{y}) = C(\boldsymbol{y},\boldsymbol{x}) \qquad \text{and} \qquad C(\boldsymbol{x},\boldsymbol{x}) = \mathbf{D}(Z(\boldsymbol{x})).$$

Autocorrelation function $\rho(\boldsymbol{x},\boldsymbol{y})$ is defined for $(\boldsymbol{x},\boldsymbol{y}) \in D \times D$ as

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \frac{C(\boldsymbol{x},\boldsymbol{y})}{\mathbf{D}(Z(\boldsymbol{x}))\mathbf{D}(Z(\boldsymbol{y}))}, \qquad (\boldsymbol{x},\boldsymbol{y}) \in D \times D$$

with properties

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \rho(\boldsymbol{y},\boldsymbol{x}) \qquad \text{and} \qquad \rho(\boldsymbol{x},\boldsymbol{x}) = 1.$$

Semivariance function $\gamma(\boldsymbol{x},\boldsymbol{y})$ is defined for $(\boldsymbol{x},\boldsymbol{y}) \in D \times D$ as

$$\gamma(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\mathbf{D}[Z(\boldsymbol{y}) - Z(\boldsymbol{x})], \qquad (\boldsymbol{x},\boldsymbol{y}) \in D \times D$$

which has properties

$$\gamma(\boldsymbol{x},\boldsymbol{y}) = \gamma(\boldsymbol{y},\boldsymbol{x}), \quad \gamma(\boldsymbol{x},\boldsymbol{y}) \geq 0 \quad \text{and} \quad \gamma(\boldsymbol{x},\boldsymbol{x}) = 0.$$

To the point-estimations of functions mentioned above, we should generally know a big number of realizations of the process $Z(\boldsymbol{x})$, $\boldsymbol{x} \in D$. However, these realizations in reality we do not have and the estimations we often compute only from a one realization of $Z(\boldsymbol{x})$. The obtained results are then independent to the position of the point $\boldsymbol{x}$ and about a behavior of the process they give us no information. The easier situation is, when the process is *stationary* in some kind of sense. It then leads to the terms *variogram* and *covariogram*.

1.2.4 Variogram

Basic terms

1. The process $Z(\boldsymbol{x})$ is said to be *stationary to its mean* if its mean is constant for every $\boldsymbol{x} \in D$.

2. Whether the autocovariance function is dependent only on the difference of arguments, we say, that the process is *stationary to its autocovariance*, i.e. for $\forall \boldsymbol{x} \in D$ and $\forall \boldsymbol{h} = (h_1, \ldots, h_d)$ provided $\boldsymbol{x} + \boldsymbol{h} \in D$ it holds

$$C(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{h}) = C(\boldsymbol{h}).$$

3. If the semivariance function is dependent only on the difference of arguments, we say, that the process is *stationary to its semivariance*, i.e. for $\forall \boldsymbol{x} \in D$ and $\forall \boldsymbol{h} = (h_1, \ldots, h_d)$ provided $\boldsymbol{x} + \boldsymbol{h} \in D$ it holds

$$\gamma(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{h}) = \gamma(\boldsymbol{h}).$$

In this case, the covariance, resp. semivariance function is said to be a **covariogram,** resp. **variogram(semivariogram)**. We denote these functions by the same letter as covariance, resp. semivariance function, even they are different.

The random process $Z(\boldsymbol{x})$ is *weakly stationary* if it fulfils conditions (1) and (2). Next, the process is said to be *intrinsically stationary(stationary)*, if it satisfy to conditions (1) and (3). So, it means that from the weak stationarity it follows intrinsically stationarity, but not conversely. Finally, one can say, that if the process is intrinsically stationary, then we obtain

$$\gamma(\boldsymbol{h}) = \frac{1}{2}\mathbf{E}\left\{[Z(\boldsymbol{x} + \boldsymbol{h}) - Z(\boldsymbol{x})]^2\right\}.$$

According to this formula, a variogram is computed and estimated. In other words, the variogram gives us "amount of dissimilarity". Similarly, covariogram $C(\boldsymbol{h})$ measures correlation dependency. Finally, it states

$$C(\mathbf{0}) = \gamma(\boldsymbol{h}) + C(\boldsymbol{h}).$$

So, we can say, that total spatial variability expressed by variance, we can divide into two parts - regular, described by covariogram, and random, described by variogram.

Variogram	Covariogram
$\gamma(\mathbf{0}) = 0$	$C(\mathbf{0}) = C(\boldsymbol{x}, \boldsymbol{x}) = \mathbf{D}(Z(\boldsymbol{x})) > 0$
$\gamma(-\boldsymbol{h}) = \gamma(\boldsymbol{h})$	$C(-\boldsymbol{h}) = C(\boldsymbol{h})$
$\gamma(\boldsymbol{h}) \geq 0$	$\lvert C(\boldsymbol{h})\rvert \leq C(\mathbf{0})$

Table 1.1: *Main properties of variogram and covariogram.*

In the case $D \subset \mathbb{R}^2$, both $C(\boldsymbol{h})$ and $\gamma(\boldsymbol{h})$ are function of $\boldsymbol{h} = (h_1, h_2)$ or of a direction α and the length $h = ||\boldsymbol{h}|| = \sqrt{h_1^2 + h_2^2}$. If we consider $C(\boldsymbol{h})$, resp. $\gamma(\boldsymbol{h})$ in the direction $\boldsymbol{h}$, consequently both $C(\boldsymbol{h})$ and $\gamma(\boldsymbol{h})$ are functions only of $\boldsymbol{h}$, i.e. distance of $\boldsymbol{x} + \boldsymbol{h}$ and $\boldsymbol{x}$. Obtained functions we caption as **directional covariogram**, resp. **directional variogram** and we write $C(\boldsymbol{h}) = C(h)$, resp. $\gamma(\boldsymbol{h}) = \gamma(h)$. From the properties introduced in the Table 1.1 we can see they are even, so it is sufficient to compute the values only for $h > 0$.

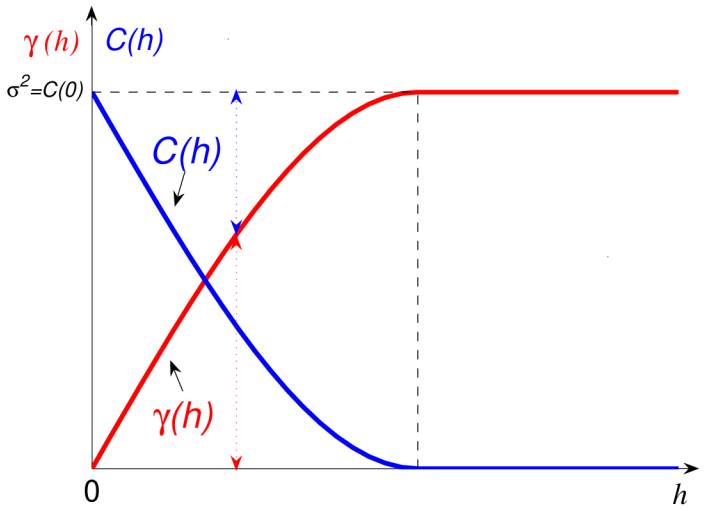


Figure 1.3: *The relation between variogram and covariogram.*

Whether the variogram depends only on the distance h of the points $\boldsymbol{x} + \boldsymbol{h}$ and $\boldsymbol{x}$ and not also on the angle of the vector $\boldsymbol{h}$, i.e. all the directional variograms are the same, then the process is *isotropic*, otherwise *anisotropic*.

Experimental variogram

From the obtained values of a measurement we compute point-estimation $\hat{\gamma}(h)$ of the variogram $\gamma(h)$ and we get so called *experimental variogram*. In a plane we estimate so-called **omnidirectional variogram** if we are sure about isotropic process, otherwise we estimate **directional variogram**. In the second case we choose several directions(e.g. horizontal, vertical and diagonals) during computations. According to [24] and references therein, the omnidirectional variogram we obtain by averaging of appropriate directional variograms. In the case, when the measurements are distributed regularly, e.g. rectangular grid, which will be our case, then we compute directional variograms in horizontal and

vertical directions and in the directions of both diagonals. For non-regular distributed measurements, see literature, e.g. [8] and others.

The behavior of a variogram for "large" h

It is possible to prove, see [8] for details, that for the weak stationary process its variogram is a top-bounded function. Next, it holds for the weak stationary and intrinsically stationary process

$$\lim_{h\to\infty} \frac{\gamma(h)}{h^2} = 0.$$

These properties allow us to decide, whether the process is weakly stationary, intrinsically stationary or non-stationary. If the variogram $\gamma(h)$ is non-bounded, then it is either intrinsically stationary or non-stationary process. Next, if the limit mentioned above is greater than zero, then the process is non-stationary, otherwise it is stationary one.

So, as the values of the variogram $\gamma(h)$ for $h \to \infty$ have finite limit, then the appropriate random process is weakly stationary, otherwise is intrinsically stationary. In the first case, the variogram have to achieve its limit value in a finite distance, say a - so called **range**, which denote the so called **zone of an effect**. The more a is bigger, the more a zone of an effect is also bigger. Then, the graph of $\gamma(h)$ is for $h \geq a$ equidistant with the $h-$axes and this part is called as a **sill**. It means, that for $h \geq a$, the values $Z(\boldsymbol{x} + \boldsymbol{h})$ and $Z(\boldsymbol{x})$ are uncorrelated and it holds $\gamma(h) = C(0)$ for $\forall h \geq a$. In the following figure you can see the situation described above.

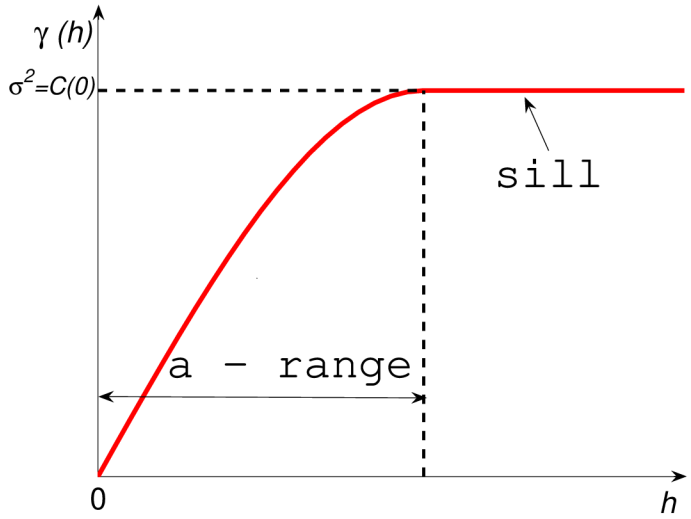


Figure 1.4: *A typical variogram - explanation of the terms.*

The behavior of a variogram for "small" h (near an origin)

Now, we will study the behavior of a variogram near the origin, because it is important for continuity and regularity of random process. In literature we can find four types of behavior of a variogram near origin:

1. **Quadratic shape.** If $\gamma(h) \leq Ah^2$, then the process is differentiable and non-similarity grows very fast.

2. **Linear shape.** If $\gamma(h)$ is linear near an origin, then $\lim_{h\to 0}\gamma(h) = 0$. It is less regular than in quadratic shape.

3. **Discontinuity at origin.** If $\lim_{h\to 0}\gamma(h) \neq 0$, then the process is neither regular nor continuous at origin. It means, that the process is variable in short distances. Non-continuity at origin is called as **nugget effect**. It indicates a variability of small-scale-distances and usually it is caused by the factors such as a microstructure, which is not measurable by given scale of sampling(short distance between two points leads to large difference of measured values), see [8] for details.

4. **Flat shape.** In this case, the process is fully random. All values $Z(\boldsymbol{x}+\boldsymbol{h})$ and $Z(\boldsymbol{x})$ are uncorrelated for $\forall h > 0$. It is a limit case of total absence of a structure.

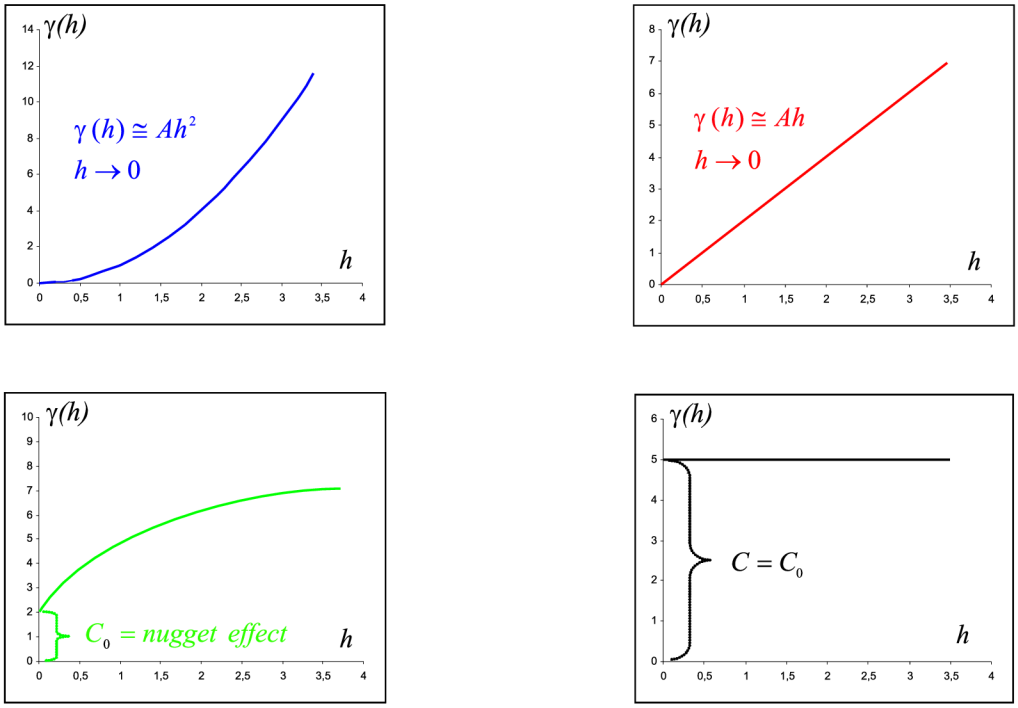


Figure 1.5: *The behavior of variograms near origin: Quadratic shape, Linear shape, Discontinuity in a origin-nugget effect and* ***Flat shape***

1.2.5 Anisotropy

About isotropy, resp. anisotropy we can decide according directional variograms, i.e variograms estimated in different directions. In the case, that these estimates are of the same or similar shape and roughly the same parameters, then we consider the process to be isotropy, otherwise anisotropy. In principle we distinguish **geometric(affine) anisotropy** and **zonal anisotropy.**

Whether the estimates of directional variograms of the similar or the same shape differ only in ranges, while the sill remains constant, then we have geometric anisotropy. But, if the directional variograms differ in more parametres than only in ranges, then it is the case of zonal anisotropy, see the following figures:

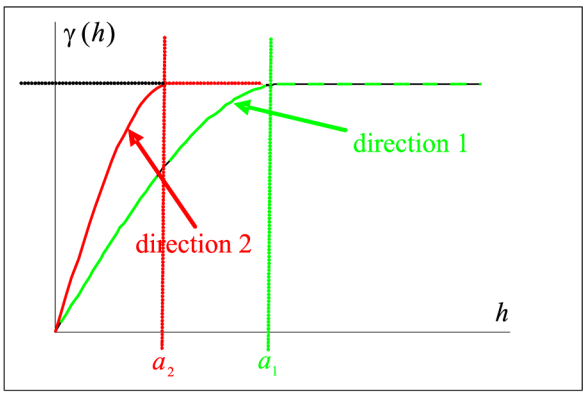


Figure 1.6: *Geometric anisotropy*

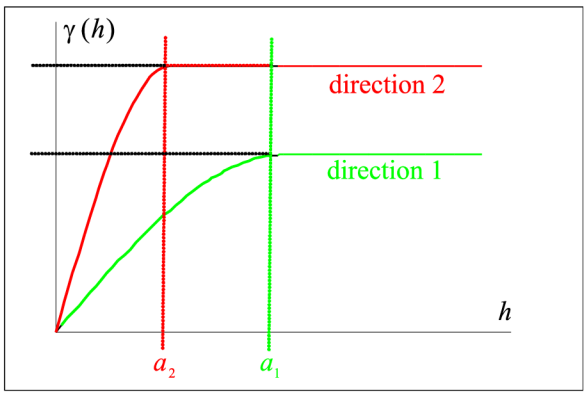


Figure 1.7: *Zonal anisotropy*

During detecting of the anisotropy in a plane, it is necessary to estimate variograms at least at four different directions to get rid of the doubt, that the anisotropy will not be detected.

In the next figures the representation of the isotropy resp. anisotropy by means of the ranges of variograms is displayed.

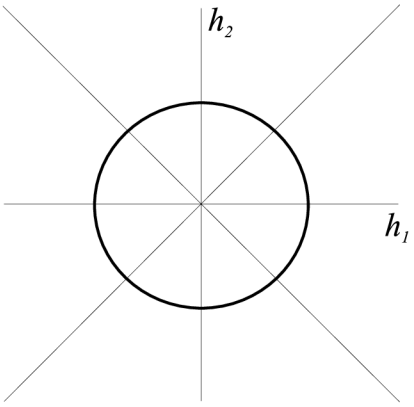


Figure 1.8: *Isotropy*

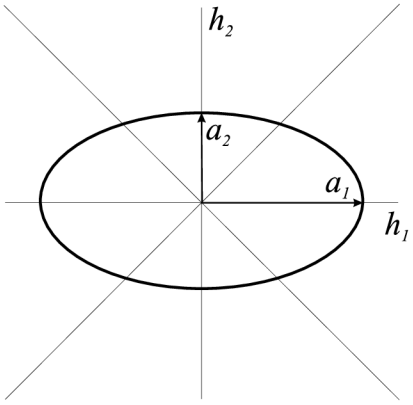


Figure 1.9: *Geometric anisotropy*

In Appendix there are presented theoretical models of variograms, which are necessary to next computations and estimations.

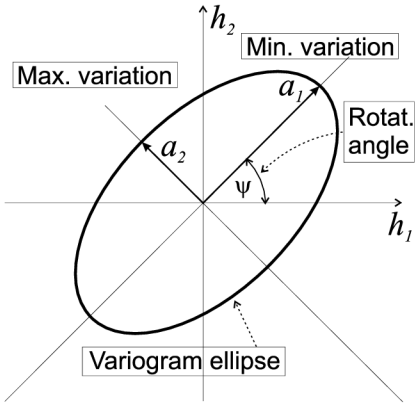


Figure 1.10: *Geometric anisotropy*

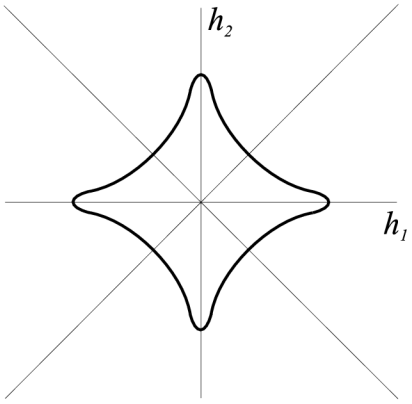


Figure 1.11: *Zonal anisotropy*

1.2.6 Spatial autocorrelation

Autocorrelation literally means that a variable is correlated with itself. The simplest definition of autocorrelation states that pairs of subjects that are close to each other are more likely to have values that are more similar, and pairs of subjects far apart from each other are more likely to have values that are less similar. The spatial structure of the data refers to any patterns that may exist. Clusters are examples of spatial structures that are positively correlated, whereas negative correlation may be exhibited in a checkerboard pattern where subjects appear to repulse each other. When data are spatially autocorrelated, it is possible to predict the value at one location based on the value sampled from a nearby location when data using interpolation methods. The absence of autocorrelation implies data are independent.

Moran's I and *Geary's c* are well known tests for spatial autocorrelation. They represent two special cases of the general cross-product statistic that measures spatial autocorrelation. Moran's I is produced by standardizing the spatial autocovariance by the variance of the data. Geary's c uses the sum of the squared differences between pairs of data values as its measure of covariation. Both of these statistics depend on a spatial structural specification such as a spatial weights matrix or a distance related decline function.

The expected value of Moran's I is $-1/(n-1)$. Values of I that exceed $-1/(n-1)$ indicate positive spatial autocorrelation, in which similar values, either high values or low values are spatially clustered. Values of I below $-1/(n-1)$ indicate negative spatial autocorrelation, in which neighboring values are dissimilar.

The theoretical expected value for Geary's c is 1. A value of Geary's c less than 1 indicates positive spatial autocorrelation, while a value larger than 1 points to negative spatial autocorrelation. The appropriate formulas for computations are:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w(i,j)} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{w(i,j)\left(Z(\boldsymbol{x}_i)-\overline{Z(\boldsymbol{x})}\right)\left(Z(\boldsymbol{x}_j)-\overline{Z(\boldsymbol{x})}\right)\right\}}{\sum_{i=1}^{n}\left(Z(\boldsymbol{x}_i)-\overline{Z(\boldsymbol{x})}\right)^2}$$

and

$$c = \frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w(i,j)} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \{w(i,j)\,(Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_j))\}}{\sum_{i=1}^{n}\left(Z(\boldsymbol{x}_i) - \overline{Z(\boldsymbol{x})}\right)^2},$$

where $\overline{Z(\boldsymbol{x})} = \frac{1}{n}\sum_{i=1}^{n} Z(\boldsymbol{x}_i)$ and $w(i,j)$ is the connectivity spatial weight between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. More information about this contiguity and probabilistic relations, see Appendix.

1.3 Models for Spatial Point Patterns

The homogeneous Poisson process provides the natural starting point for a statistical investigation of an observed point pattern. Rejection of the complete spatial randomness hypothesis does not come as a great surprise in many applications and we are naturally confronted with the question "What kind of pattern is it?" If the complete spatial randomness test suggests a clustered pattern, one may want to compare another characteristics, e.g. second-order moments (Ripley's K function), see later.

We can only skim some point processes models here. A large number of models have been developed and described for clustered and regular alternatives. Details can be found e.g. in [9], [8], etc. The models presented here were chosen for their representativeness and for their importance in theoretical and applied statistics.

1.3.1 Poisson process

The Poisson point process is the simplest yet the most important random point pattern. The reasons for this importance are, firstly, that typically the Poisson model is the "null model" implying complete lack of structure or external influence on the pattern, so departures from this will reflect some practical feature in the production of the patterns. The second aspect of the role of a Poisson process is that many more complex models have the Poisson process as a constituent part.

A given point pattern may exhibit various kinds of interaction between its constituent points. Thus, the points may occur in clusters or may exhibit great regularity. There may be a threshold distance (also called hard-core distance) which is a minimal inter point distance. These extreme features may even occur together in the same pattern. The aim of point process statistics is to detect and to quantify such interactions. If none of the above interactions is present, the point pattern can be thought of as completely random, that is, its points are randomly distributed in the space, they form a Poisson process.

Models from the theory of point processes can be used both in comparison to the original point pattern and also in representation of it. Clark and Evans (1954) describe a random distribution as being a set of points on a given area, where it is assumed that any point has had the same chance of occurring on any sub-area as any other point, that any sub-area of specified size has had the same chance of receiving a point as any other

sub-area of that size, and that the placement of each point has not been influenced by that of any other point. Thus, randomness is dependent upon the boundaries of the space chosen by the investigator. A set of points may be random with respect to a specified area but decidedly non-random with respect to a larger space which includes the specified area. In order to get meaningful results, the areas selected for the investigation should be chosen with care.

Definition of the Poisson Process

The Poisson process is a formalization of the concept of randomness and is defined by the following postulates.

- For some $\lambda > 0$, and any finite region A, $N(A)$ has a Poisson distribution with mean $\lambda|A|$.
- Given $N(A) = n$, the n events in A form an independent random sample from the uniform distribution on A.
- For any two disjoint regions A and B, $N(A)$ and $N(B)$ are independent.

A spatial point pattern satisfying these criteria is also said to exhibit complete spatial randomness, abbreviated to CSR. According to the first item, CSR therefore implies that the intensity of events does not vary over the region (and consequently it explains the reason why a random distribution of points in space may be referred to as a spatial Poisson process). According to the second, CSR also implies that there are no interactions amongst the events.

Note: The generalization of the Poisson process is the so called *inhomogeneous Poisson process*. It differs from the homogeneous one in the fact, that the intensity λ is not constant overall the domain, but varies spatially. It usually leads to clustering and we refer to [9] for more information.

Regular Processes

1.3.2 Hard-Core Models

A hard-core point process is a point process in which the constituent points are forbidden to lie closer together than a certain minimum distance, denoted throughout as τ, resulting in an even or regular spatial distribution of points.

Regular patterns arise most naturally by the imposition of a minimum permissible distance, τ say, between any two points. This may simply reflect the physical size of the entities whose locations define the point pattern. Matèrn was the first(1960), who described formally the hard-core models.

Processes of this sort, which incorporate no further departure from complete spatial randomness, are also commonly called simple inhibition processes. Monograph [8] provides a detailed descriptions of simple inhibition (or Hard-Core) point processes including Matérn's Models whose **some** definitions will be stated in the following two subsections. Basically, these models describe two possible ways to obtain inhibited patterns from a Poisson process.

Note: The theory of the *Hard-Core models* is so broad, that it can not be held everything in this thesis. We will try, only, introduce only the basics of it.

Matérn's Model I

Consider a Poisson process N_0 on $\mathbb{R}^d$ with intensity ρ. Model I is formed by deleting all pairs of points of the Poisson process that are separated by a distance of less than τ whether or not either point of that pair had already been deleted. The remaining points form a (more regular) process.

Matérn's Model II

Let N_0 be a homogeneous Poisson process on $\mathbb{R}^d$ with intensity ρ. Independently mark the events s of N_0 with numbers $Z(s)$ from any absolutely continuous distribution function F. An event s of N_0 with mark $Z(s)$ is deleted if there exists another event u with $||s-u|| < \tau$ and $Z(u) < Z(s)$. The retained events form the (more regular) spatial point process.

Simple Sequential Inhibition Point Process

A simple sequential inhibition point process (SSI) is defined as the output of an algorithm that repeatedly introduces particles at random into a bounded window A, discarding those that would overlap a previously introduced particle, until some stopping criterion is satisfied. It can be imagined in this way: Again, consider a Poisson process N_0 on a domain A with intensity ρ. We place a disc of a radius, say δ at random in a region A. Then we determine the remaining points in A, for which we can place a disc of radius δ that do not overlap with the first disc. Then we select the center point at which the next disc at random from a uniform distribution of these points. We continue in this fashion, choosing at each stage the disc center at random from the points at which the next disc does not overlap with any of the previous discs. The process stops when a pre-specified number of discs have been placed or no additional disc can be placed without overlapping previously placed discs.

Summarizing

In [8] detailed information about hard-core models are presented. After the deeper studium of spatial processes, one can see, that it is meaningless to delay with another

types of spatial processes because of their non-similarity to fibre composites. From the previous mentioned, it is clear, that fibre composites can not have CSR character, because of their nonzero fibre diameters. So, it means, that for their modeling it is required to use regular spatial processes. Later we show, the tests of CSR indicating regularity of fibre reinforced composite materials.

Cluster Processes

1.3.3 Soft-Core (Cluster Point) Models

In contrast to hard-core models, soft-core models are those, where the number of neighbors within some critical distance δ is smaller than expected under CSR, but the number is not zero. These processes are sometimes called as *cluster-point processes with spherically shaped neighborhood.* The construction of this processes is very simple, see [25] for details and appropriate algorithms inside.

1.3.4 Cox Process

If the point intensity varies from sub-region to sub-region, thereby implying that some sub-regions are more likely to contain points than others, then the resulting point distribution will take on a "patchy" appearance. This is what is called a Cox process (also named a doubly stochastic Poisson process). The latter comes from the idea that such a process can be thought of as arising from a two-step random mechanism.

Note: A generalization of a Poisson process is made by supposing that the intensity measure is itself random, with the point process being Poisson conditional on the realization of the intensity. In the simple homogeneous Poisson process, the intensity is the same everywhere.

2 Spatial Data Analysis

2.1 Introduction

As we said in the previous chapter, data in the form of a set of points, irregularly distributed within a region of space creates the so called spatial point pattern. In figure 1.2 we can see an example of clustered and regularized point pattern. Our next example of a point pattern, shown in figure 2.1 introduces the idea of a *multivariate point pattern*. In this example, the points represent cells of two different types (hence bivariate), e.g. three-phase composite material reinforced by fibres made of two types of materials.

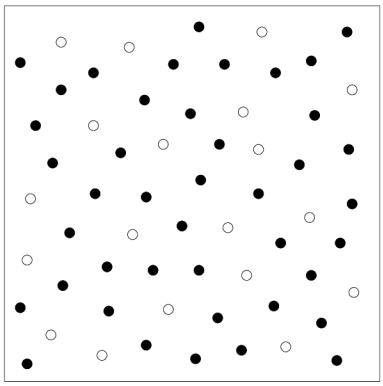


Figure 2.1: *Multivariate point process – three-phase composite material*

Further, edge effects play very important role in spatial statistics, see [9]. *Edge effects* arise in spatial point pattern analysis when, as is often in practice, the region, say A, on which the pattern is observed is part of a larger region on which the underlying process operates. The essential difficulty is that unobserved events outside A may interact with observed events within A.

In many publications, e.g. see [9], [32] or [30], many techniques how to avoid mistakes by not including these edge effects are described, but in our accounts we will not consider these effects from the reason of their exigence from the computational time point of view.

2.2 Complete spatial randomness (CSR)

Complete spatial randomness (CSR) data describes a point process whereby points are placed within a volume in a completely uncorrelated, i.e. random fashion. Such a process requires only one parameter, i.e. the density of points, λ within a volume. This model is, that points are derived from a spatial Poisson process, see 1.3.1.

The study of such a point process is essential for the comparison of point data from experimental sources to examine data sources for statistical correlations. As a statistical testing method, the CSR distribution finds applications in areas.

For any finite region of space, the average number of points located within the volume will be given by the density of the data multiplied by the volume of the region. However,

for each individual sampling of the data, the number of points in the volume is governed by a Poisson distribution.

According to [9], the hypothesis of CSR for a spatial point pattern asserts that

1. the number of events in any planar region A with area $|A|$ follows the Poisson distribution with mean $\lambda|A|$,

2. given n events x_i in a region A, the x_i are independent random sample from the uniform distribution on $|A|$.

For more information see [30]. The constant λ is the so called *intensity*, or mean number of events per unit area. According to the first item, CSR therefore implies that the intensity of events does not vary over the plane. According to the second item, CSR also implies that there are no interactions amongst the events.

Our interest in CSR is that it represents an idealized standard which, if strictly unattainable in practice, may nevertheless be tenable as a convenient first approximation. Most analyzes begin with a test of CSR and there are several reasons for this: Firstly, a pattern for which CSR is not rejected scarcely merits any further formal statistical analysis. Secondly, test are used as a means of exploring a set of data, because rejection of CSR is of intrinsic interest. Thirdly, CSR acts as a dividing hypothesis to distinguish between patterns which are broadly classifiable as a regular or clustered. Another use of CSR is as a building block in the construction of more complex models.

2.3 Tests of complete spatial randomness

Although CSR is of limited scientific interest in itself, there are several good reasons why we might begin an analysis with a test of CSR: rejection of CSR is a minimal prerequisite to any serious attempt to model an observed pattern; tests are used to explore a set of data and to assist in the formulation of plausible alternatives to CSR and of course CSR operates as a dividing hypothesis between regular and clustered patterns.

Several different approaches will be taken to quantify types of spatial point pattern. The general goal in the following subsections is to reduce the spatial data to informative descriptives statistics that can help elucidate models that might be fitted to the real point pattern.

Randomness tests are based on the following three methods:

- Quadrat tests
- Second-order methods
- Distance methods

Methods of the first type are the most appropriate in preliminary studies and they should always be backed up by other tests. Problems of edge correction are avoided here for the sake of simplicity.

2.3.1 Quadrat methods

One type of descriptive statistic is based on quadrats (i.e. well defined areas, often rectangular in a region of interest A). According to [8], usually, quadrats of random location and orientation are sampled, the number of events in the quadrats are counted (here the events are fibres) and statistics derived from the counts are computed. As well as a count of fibres, the percent of area covered by the fibres in the quadrats might also be recorded.

Random quadrats

We shall demonstrate using of random quadrats on the sample of fibre composite. Figure 2.2 depicts the positions of $m = 36$ squared quadrats in the extended study area. Note, that no two quadrats overlap.

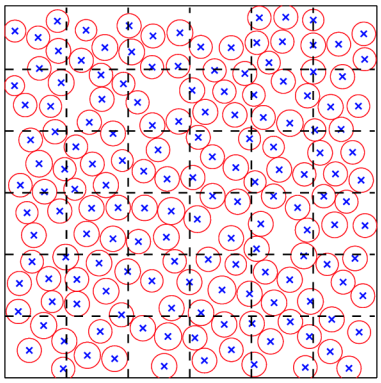


Figure 2.2: *The sample of real composite showing 160 fibres placed in squared quadrats.*

In computation, first of all, the fibres in each quadrant are enumerated. Table 2.1 gives the frequency distribution of the number of fibres per quadrat. Under CSR, the number of fibres in a quadrat, say A_1, of area $|A_1|$, has a Poisson distribution with mean $\lambda|A_1|$, where λ is the intensity of the Poisson process. Table 2.1 also gives the expected frequency distribution of number of fibres per quadrat under a Poisson distribution with estimated mean λ (here $\lambda = \frac{160}{6.6} \approx 4,44$). According to [8] or [9], one test for CSR is Pearson's χ^2 goodness-of-fit test.

Note, that in reality, the distribution of fibres in a quadrats is driven by binomial distribution, but from the computational point of view it is replaced by Poisson one, which also states in definition in CSR. So, if we denote by n the number of points in a sample, $A = \bigcup_{i=1}^{m} A_i$ the explored area and by n/m the expected number of fibres(their centers) in each guadrat, then we can write

$$\hat{\lambda} = \frac{n}{|A|}$$

and the *chi-square statistic*

$$Q = \sum_{i=1}^{m} \frac{\left(n_i - \frac{n}{m}\right)^2}{\frac{n}{m}}$$

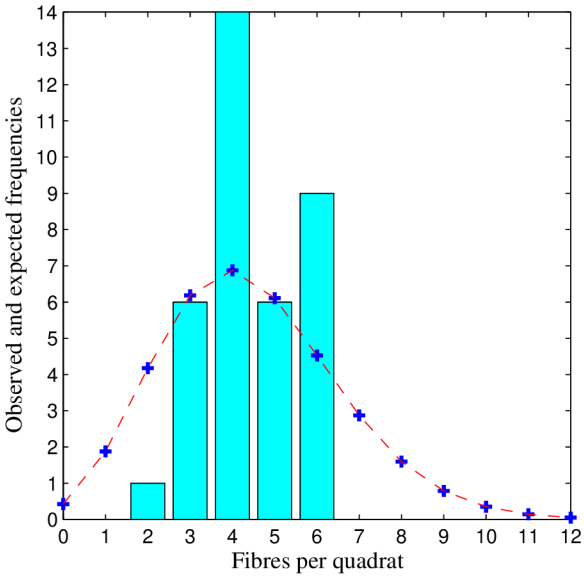


Figure 2.3: *Graph of frequencies*

Fibres per quadrat	Observed frequency	Expected frequency
0	0	0,42
1	0	1,88
2	1	4,18
3	6	6,19
4	14	6,87
5	6	6,11
6	9	4,53
7	0	2,87
8	0	1,60
9	0	0,79
10	0	0,35
11	0	0,14
12	0	0,05

Table 2.1: *Frequency distribution of number of fibres per quadrat.*

is known to be asymptotically chi-squared distributed with $m-1$ degrees of freedom, under CSR hypothesis. But since n/m is simply the *sample mean*, i.e.

$$\frac{n}{m} = \frac{1}{m}\sum_{i=1}^{m} n_i = \overline{n},$$

this statistic can also be written as

$$Q = \sum_{i=1}^{m} \frac{(n_i - \overline{n})^2}{\overline{n}} = (m-1)\frac{S^2}{\overline{n}},$$

where $S^2 = \frac{1}{m-1}\sum_{i=1}^{m}(n_i - \overline{n})^2$ is the *sample variance*. For more detail see e.g. [3], [2], [22], [21], [16] or [14]. In our example Pearson's test statistic $Q = 10,10 < \chi^2_{35}(0,975) = 20,57$ indicates significant departure from a Poisson distribution, i.e. CSR. So, the next question is about regularity or clustering.

Regularity and clustering

Once CSR hypothesis is rejected, the next step in a spatial analysis may be to measure the departure from CSR. According to [8], in table 2.2 we can see some characteristics for identifying clustering or regularity.

Here, in the Table 2.2, $\overline{X} = 4,444$ is the sample mean of the quadrat counts and $S^2 = 1,284$ is the sample variance.

The relative variance index I and the clumping index ICS were obviously motivated by the equality of mean and variance of Poisson quadrat counts (mean-to-variance ratio).

Description	Index	Estimator	Realization
Relative variance index	I	$\frac{S^2}{\overline{X}}$	0,289
Clumping index	ICS	$\frac{S^2}{\overline{X}} - 1$	-0,711
Cluster frequency index	ICF	$\frac{\overline{X}^2}{S^2 - \overline{X}}$	-6,247
Mean event index	$\overset{*}{X}$	$\overline{X} + \frac{S^2}{\overline{X}} - 1$	3,733
Mean crowding index	IP	$\frac{\overset{*}{X}}{\overline{X}}$	0,840
Morisita's index	I_δ	$\frac{\sum_{i=1}^{m} X_i(X_i - 1)}{\overline{X}(m\overline{X} - 1)}$	0,843

Table 2.2: *Indices for quadrats count data, see [8], [30].*

It is clear, that the expected value of ICS is zero and value of I equals to one for Poisson quadrat counts. Values of I greater than 1 and ICS greater than 0 would indicate that the fibres are clustered. If ICS (our case) is less than 0 and I is less than 1, then the fibres indicate a tendency for regular spacing. In [8] and [30] you can get more information about a relation between ICS and ICF. Index ICF is meaningful for samples without CSR. The mean event index $\overset{*}{X}$ indicates an average number of events sharing a quadrat. Mean crowding Index IP is often called as an index of patchiness. If IP is equal to 1, then the distribution is random, regular for if $IP > 1$ and clustered if $IP < 1$. Morisita's index I_δ comes from the idea, that the point process consists of patches of differing intensities and it measures variability between patches. The previous results we can see collected in the Table 2.3 and according to thie methods we can say, that our sample is regular.

Index	Random	Regular	Clustered
I	$= 1$	< 1	> 1
ICS	$= 0$	< 0	> 0
IP	$= 1$	< 1	> 1
I_δ	$= 1$	< 1	> 1

Table 2.3: *The values of indexes for different types of patterns.*

Note: Of course, the natural question is: what constitutes "big" or "small"? To answer this question, the behavior (i.e. sampling distribution) of I needs to be known when the null hypothesis is true. If I is standardized as

$$T = \frac{I - 1}{\sqrt{\frac{2}{n-1}}},$$

then T follows a t-distribution on $n - 1$ degrees of freedom approximately under the hypothesis of complete randomness.

2.3.2 Second Order Methods

These tests are designed to detect deviation from randomness and consist of the use of Monte-Carlo tests which are backed up by a graphical procedure.

Tests Based on Ripley's K Function

Monte-Carlo statistics measure the discrepancies between the estimated function, i.e. the empirical distribution function obtained from the pattern, $\widehat{K}(t)$, and the expected function that would be obtained in the case of randomness, $\mathbf{E}\left[\widehat{K}(t)\right]$. This measure of the discrepancy is calculated over a specific range of distances t.

Three statistics that measure possible discrepancies are:

$$K_M = \max_{t_0 \le t \le t_n} \left| \widehat{K}(t) - \mathbf{E}\left[\widehat{K}(t)\right] \right|, \quad L_M = \max_{t_0 \le t \le t_n} \left| \sqrt{\widehat{K}(t)} - \sqrt{\mathbf{E}\left[\widehat{K}(t)\right]} \right|,$$

$$L_I = \int_{t_0}^{t_n} \left(\sqrt{\widehat{K}(t)} - \sqrt{\mathbf{E}\left[\widehat{K}(t)\right]} \right)^2 \mathrm{d}t.$$

The square-root transformation used in the latter two statistics was suggested as a variance stabiliser, see [32].

For two-dimensional patterns, it holds $\mathbf{E}\left[\widehat{K}(t)\right] = \pi t^2$, which is the expression $K(t)$ for a Poisson process. K_M and L_M measure the maximum discrepancy between observed and expected values of $K(t)$ over the range t_0 to t_n. These limits are chosen according to the window size and also to the range of distances t, between the events, one is interested in studying.

L_I measures the integrated squared distance between $\sqrt{\widehat{K}(t)}$ and $\sqrt{\mathbf{E}\left[\widehat{K}(t)\right]}$ over the t_0 to t_n range and is thus an aggregated measure of discrepancy. The Monte Carlo tests assess only deviation from randomness. The null hypothesis of randomness will be rejected in the presence of either a clustered or an inhibited pattern. When that happens the only way of finding out whether the pattern shows evidence of clustering or regularity is by the use of a graphical procedure.

As recommended in [30] and [9], a graphical procedure consists of comparing the $\widehat{K}(t)$ and $\mathbf{E}\left[\widehat{K}(t)\right]$ if known with the upper $U(t)$ and lower $L(t)$ simulation envelopes defined by

$$U(t) = \max_i \widehat{K}_i(t) \quad \text{and} \quad L(t) = \min_i \widehat{K}_i(t),$$

where the empirical distribution functions $\widehat{K}_i(t)$ are obtained from each independent simulations. The simulated envelopes provide the acceptance region for a further nonparametric test of the hypothesis that the process is Poisson.

Therefore, if in the plot $\widehat{K}(t)$ lies entirely between $U(t)$ and $L(t)$ throughout its range (i.e. $\widehat{K}(t)$ lies between the simulated envelopes), there is no evidence to suggest any departure from a CSR model. If $\widehat{K}(t)$ lies entirely below $L(t)$ it means that for the values of t considered there were few points which were within distance t suggesting that there must exist some sort of inhibition that keeps the points at a certain distance apart. As a result, there are strong reasons to believe the events on the patterns to be regularly distributed.

If the opposite happens, i.e. if $\widehat{K}(t)$ lies entirely above both envelopes, it means that for every value of t there are many points at most a distance t from each other. This suggests that the points must be somehow clumped together and so giving strong evidence of clustering in the pattern.

The less clear-cut case is when $\widehat{K}(t)$ lies outside the envelopes for just part of the range but inside them for other parts. This problem leads to the empirical study and it is more detailed described in [32].

Tests Based on the F Function

The Monte Carlo statistics based on the F function are calculated by the following expressions:

$$F_M = \max_{t_0 \le t \le t_n} \left| \widehat{F}(t) - \mathbf{E}\left[\widehat{F}(t)\right] \right| \quad \text{and} \quad F_I = \int_{t_0}^{t_n} \left(\widehat{F}(t) - \mathbf{E}\left[\widehat{F}(t)\right] \right)^2 \mathrm{d}t.$$

From the first postulate of a Poisson process (that for some $\lambda > 0$, and any finite planar region A, the number of points in A has a Poisson distribution with mean $\lambda|A|$), we deduce that, for two dimensional spaces: $F(t)$=**P**(there is at least one event in the circle centered at x_0 with radius t)$=1-\exp(-\lambda\pi t^2)$. If we undertake a graphical procedure, we arrive at the following result: Here, the plot's interpretation is different from that of the Ripley's K function.

In the presence of a clustered pattern, there will be a smaller number of point-to-object distances than would be the case in a Poisson process and so the estimated EDF, $\widehat{F}(t)$, takes smaller values than the theoretical function, $F(t) = 1 - \exp(-\lambda\pi t^2)$ for all (or at least for most) of the t range of distances considered. However, for a regular alternative there will be a greater number of point-to-object distances than would be the case for a random process (i.e. a Poisson process) and $\widehat{F}(t)$ would be much greater than $F(t)$.

Plotting $\widehat{F}(t)$ as the ordinate against t as the abscissa together with upper $U(t)$ and lower $L(t)$ envelopes, helps identify the type of model appropriate for the spatial distribution of the particles. The envelopes are obtained similarly to those for the nearest-neighbor and K functions, however, their interpretation is different (essentially "reversed").

If the plot of $\widehat{F}(t)$ lies between $U(t)$ and $L(t)$ throughout its range it indicates no evidence to suggest any deviation from a CSR model. If $\widehat{F}(t)$ lies beneath $L(t)$, it means that for the values of t considered there are very few points whose distance to their nearest neighbor is at least t. This indicates that the particles in the pattern might somehow be clumped together.

If $\widehat{F}(t)$ lies above both envelopes, it means that for every value of t there are many points whose distance to any of the m fixed points are at least t. This suggests that the points must somehow be restricted to a minimum distance apart, giving rise to regularly distributed patterns.

An entirely similar procedure can be employed to determine CSR using J function, [32].

2.3.3 Distance methods

Distance methods, also known as plotless sampling techniques, were introduced because of the practical difficulties sometimes raised by quadrat sampling. Whereas quadrat methods lend themselves to field sampling, some of the more powerful distance rely on having a good map of all events. Distance methods make use of precise information on the locations of events and have the advantage of not depending on arbitrary choices of quadrat size or shape.

Nearest-neighbor methods

Here, event-to-event or point-to-event distances are computed and summarized. The following Figure 2.4 illustrates various possibilities. Distances may be measured between events and nearest-neighboring events (W) or between sample points and nearest events (X). Sometimes it is used the second nearest event X_2. Sample points usually are located randomly in the study area, but may be placed systematically. The distribution theory for W and X under CSR is well known, see [8] for details. In $\mathbb{R}^2$, the density of the positive random variable W is

$$g(w) = 2\pi\lambda w \mathrm{e}^{-\pi\lambda w^2}, \quad w > 0.$$

The distance from a randomly placed sample point to the nearest event X, has the same distribution as W.

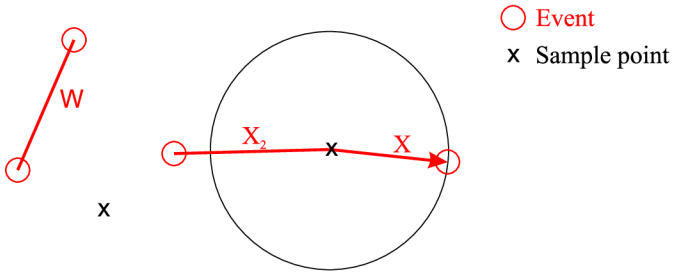


Figure 2.4: *Types of nearest-neighbor distances* X, X_2, W.

Test statistics. Many statistics have been proposed for testing CSR, usually based on random sample of n points or a random sample of n events. A summary of test statistics and their asymptotic distributions under CSR is presented in the following Table, see [8]. Distribution theory for those tests is based on independence of n nearest-neighbor measurements randomly sampled from a region A. Some comments to the statistics presented in the table below, you can see in the Appendix and in [8].

Measurement	**Test statistic**	**Distribution**	**Author**
W	$S_1 = \frac{1}{m}\sum_{i=1}^{m} W_i$	$\mathrm{N}(\frac{1}{2\sqrt{\lambda}}, \frac{4-\pi}{4\lambda\pi m})$	Clark & Evans
W	$S_2 = 2\pi\lambda\sum_{i=1}^{m} W_i^2$	χ^2_{2m}	Skellam
X	$S_3 = \pi\lambda\frac{1}{m}\sum_{i=1}^{m} X_i^2$	$\mathrm{N}\left(1, \frac{1}{m}\right)$	Pielou
X	$S_4 = m\frac{\sum_{i=1}^{m} X_i^2}{\left(\sum_{i=1}^{m} X_i\right)^2}$	By simulation	Eberhardt
X	$S_5 = 12\frac{m^2 \log\frac{\sum_{i=1}^{m} X_i^2}{m} - \sum_{i=1}^{m} \log X_i^2}{7m+1}$	χ^2_{N-1}	Pollard
$X,\ X_2$	$S_6 = \frac{\sum_{i=1}^{m} \frac{X_i^2}{X_{2,i}^2}}{m}$	$\mathrm{N}\left(\frac{1}{2}, \frac{1}{12}m\right)$	Holgate
$X,\ X_2$	$S_7 = \frac{\sum_{i=1}^{m} X_i^2}{\sum_{i=1}^{m} X_{2,i}^2}$	$\beta(m,\, m)$	Holgate
$X,\ W$	$S_8 = \frac{\sum_{i=1}^{m} \frac{X_i^2}{X_i^2+W_i^2}}{m}$	$\mathrm{N}\left(\frac{1}{2}, \frac{1}{12}m\right)$	Byth & Ripley
$X,\ W$	$S_9 = \frac{\sum_{i=1}^{m} X_i^2}{\sum_{i=1}^{m} W_i^2}$	$F_{2m,\,2m}$	Hopkins

Table 2.4: *Nearest-neighbor statistics and their asymptotic distribution under CSR*

To reduction of complex point patterns to a one-dimensional nearest-neighbor summary statistic results in a considerable loss of information. Nearest-neighbor statistics indicate only departure from the CSR. Little is known about the behavior of these statistics when CSR does not hold, see [8].

More information about mentioned statistics together with its detailed description you can find namely in [8] and the first two of them is described below and the second one in the Appendix.

Two-tailed test of CSR — Clark-Evans test

This two-tailed test of CSR is in literature very known as the so called *Clark-Evans test of CSR*. To construct a test of the CSR hypothesis based on the Clark & Evens statistic, suppose that one starts with a sample pattern $S_n = \{s_i : i = 1, \ldots, n\}$ and constructs the *nn-distance* (nearest-neighbour) for each point $s_i \in S_n$. Then it would seem most natural to use all these distances $\{d_1, \ldots, d_n\}$ to construct the sample-mean statistic in 2.3.3. However, this would violate the assumed *independence* of nn-distances on which this theory is based. To see this, it is enough to observe that if s_i and s_j are mutual nearest neighbors, so that $d_i \equiv d_j$, then these are obviously not independent. More generally, if s_j is the nearest neighbor of s_i, then again d_i and d_j must be dependent. However, if one

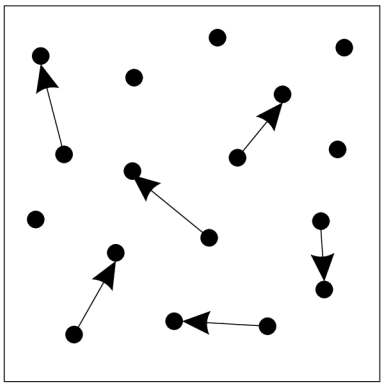


Figure 2.5: *Cell of radius d*

selects a subset of nn-distance values that contained no common points, such as those shown in Figure 2.5, then this problem could be in principle avoided. The question is how to choose *independent* pairs. Now we simply assume, that some "independent" subset $(W_1, \ldots, W_m)$ of these distance values has been selected (with $m < n$). Widely, it is for computations the following rule

$$m = \frac{n}{2}.$$

By $\overline{W}_m$ we denote the sample-mean value

$$\overline{W}_m = \frac{1}{m} \sum_{i=1}^{m} W_i. \tag{2.3.1}$$

By differentiating we obtain the probability density f_W of W as

$$f_W(w) = F'_W(w) = 2\pi\lambda w \mathrm{e}^{-\lambda\pi w^2}. \tag{2.3.2}$$

It can be shown, see [34], that mean and variance of this distribution are given respectively by

$$\mathbf{E}[W] = \frac{1}{2\sqrt{\lambda}}, \qquad \mathbf{D}[W] = \frac{4-\pi}{4\lambda\pi}. \tag{2.3.3}$$

Next we observe from the properties of *iid* random samples that for the sample mean $\overline{W}_m$ in 2.3.1 it holds

$$\mathbf{E}\left[\bar{W}_m\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbf{E}\left[W_i\right] = \frac{1}{m}\left(m\mathbf{E}\left[W_1\right]\right) = \mathbf{E}\left[W_1\right] = \frac{1}{2\sqrt{\lambda}} \tag{2.3.4}$$

and similarly

$$\mathbf{D}\left[\bar{W}_m\right] = \left(\frac{1}{m}\right)^2\sum_{i=1}^{m}\mathbf{D}\left[W_i\right] = \frac{1}{m^2}\left(m\mathbf{E}\left[W_1\right]\right) = \frac{4-\pi}{m(4\lambda\pi)}. \tag{2.3.5}$$

From the *central limit theorem* we obtain

$$\overline{W}_m \sim \mathrm{N}\left(\frac{1}{2\sqrt{\lambda}}, \frac{4-\pi}{4\lambda\pi m}\right) \tag{2.3.6}$$

and after standardization we can write

$$Z_m = \frac{\overline{W}_m - \mathbf{E}(\overline{W}_m)}{\sqrt{\mathbf{D}\left[\overline{W}_m\right]}} \sim \mathrm{N}(0;\,1). \tag{2.3.7}$$

and use it to construct tests of CSR. The standard test of CSR in most software is a two-tailed test in which both the possibility of "significantly small" values of $\overline{w}_m$ (*clustering*) and "significantly large" values of $\overline{w}_m$ (*regularity*) are considered. First, recall the notion of *upper-tail* points, z_α, for the standard normal distribution as defined by $\mathbf{P}(Z \geq z_\alpha) = \alpha$ for $Z \sim \mathrm{N}(0,1)$. In these terms, it follows that for the standardized mean in 2.3.6

$$\mathbf{P}\left(|Z_m| \geq z_{\alpha/2}\right) = \mathbf{P}\left[(Z_m \leq -z_{\alpha/2}) \vee (z_{\alpha/2} \leq Z_m)\right] = \alpha \tag{2.3.8}$$

under CSR hypothesis. If we write the estimates of the mean and standard deviation under CSR by

$$\hat{\mu} = \frac{1}{\sqrt{2\hat{\lambda}}}, \qquad \hat{\sigma}_m = \sqrt{\frac{4-\pi}{4\pi\hat{\lambda}m}}, \tag{2.3.9}$$

then one can test the CSR hypothesis by constructing the following standardized sample mean:

$$z_m = \frac{\overline{w}_m - \hat{\mu}}{\hat{\sigma}_m}. \tag{2.3.10}$$

If the CSR hypothesis is true, then z_m should be a sample from $\mathrm{N}(0,1)$. Hence a test of CSR at the $\alpha-$level of significance is then given by the rule:

Two-tailed CSR test: *Reject the CSR hypothesis if and only if* $|z_m| > z_{\alpha/2}$.

The significance level α is also called the *size* of the test. Example results of this testing procedure for a test of size α are illustrated in Figure 2.6. Here the two samples, z_m, in the tails of the distribution are seen to yield strong evidence against the CSR hypothesis, while the sample in between does not.

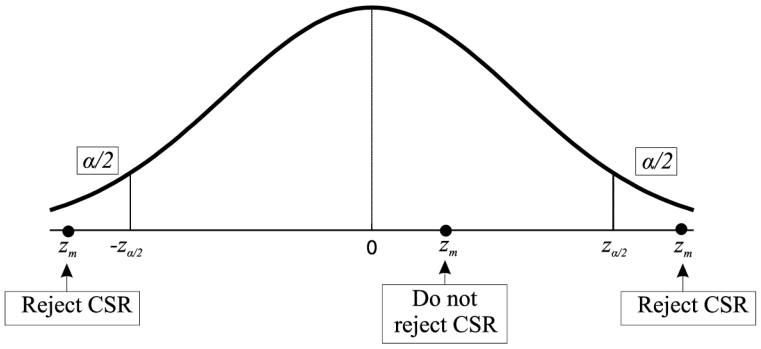


Figure 2.6: *Two-tailed test of CSR.*

One-tailed tests of clustering and regularity

As already noted, values of w_m (and hence z_m) that are too low to be plausible under CSR are indicative of pattern more regular than random. Similarly, values too large are indicative of patterns more clustered than random. In many cases, one of these alternatives is more relevant than the other. So the key question here is whether our pattern is *significantly more clustered than random.* Similarly, one can ask whether the pattern is significantly more regular than random. Such questions lead naturally to one-tailed versions of the test above. First, a test of *clustering* versus CSR hypothesis at the $\alpha-$level of significance is given by the rule:

Clustering versus CSR test: *Conclude significant clustering if and only if* $z_m < z_\alpha$.

Example results of this testing procedure for a test of size α are illustrated in Figure 2.7 below. Here the standardized sample mean z_m to the right is sufficiently low to conclude the presence of clustering (at the $\alpha-$level of significance), and the sample toward the middle is not. In a similar manner, one can construct a test of *regularity* versus CSR

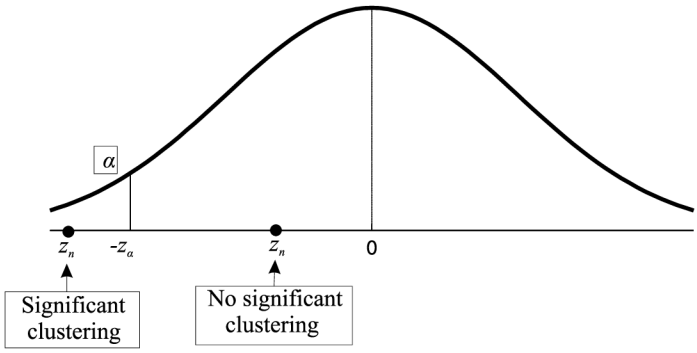


Figure 2.7: *One-tailed test of clustering.*

hypothesis at the $\alpha-$level of significance using the rule:

Regularity versus CSR test: *Conclude significant clustering if and only if* $z_m > z_\alpha$.

Example results for a test of size α are illustrated in Figure 2.8 below, where the sample z_m to the left is sufficiently high to conclude the presence of regularity (at the $\alpha-$level of

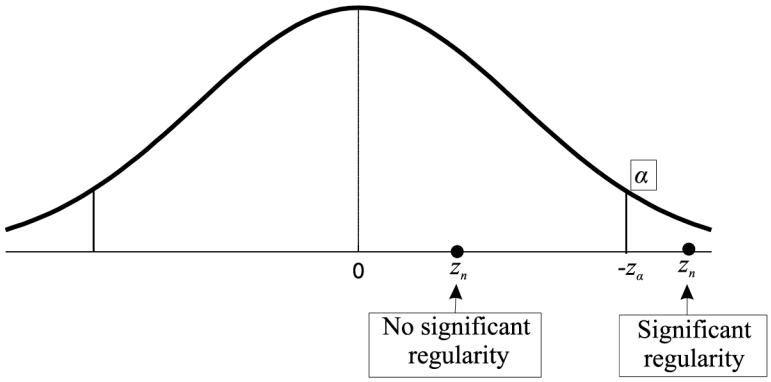


Figure 2.8: *One-tailed test of regularity.*

significance) and the sample toward the middle is not. While such tests are standard in literature, it is important to emphasize that there is no "best" choice of α. The typical values given by most statistical tests are listed in tables below.

Significance	$\boldsymbol{\alpha}$	$z_{\boldsymbol{\alpha/2}}$
"Strong"	0,01	2,58
"Standard"	0,05	1,96
"Weak"	0,10	1,65

Table 2.5: *Two-tailed significance*

Significance	$\boldsymbol{\alpha}$	$z_{\boldsymbol{\alpha}}$
'Strong"	0,01	2,33
'Standard"	0,05	1,65
'Weak"	0,10	1,28

Table 2.6: *One-tailed significance*

However, since these distinctions are admittedly arbitrary, another approach is often adopted in evaluating test results. The question is easily answered by simply calculating the probability of a sample value as z_m for the standard normal distribution $\mathrm{N}(0,1)$. If the cumulative distribution function for the normal distribution is denoted by

$$\Phi(z) = \mathbf{P}(Z < z), \tag{2.3.11}$$

then this probability, called *p-value* of the test, is given by

$$\mathbf{P}(Z \leq z_m) = \Phi(z_m). \tag{2.3.12}$$

Notice that unlike the significance level α above, the p-value for a test depends on the realized sample value z_m and hence is itself a random variable that changes from sample to sample. More generally, the p-value can be defined as the largest level of significance (smallest value of α) at which CSR would be rejected in favor of clustering based on the given sample value z_m.

Similarly, one can define p-value for a test of *regularity* in the same way. Hence, the p-value in this case is

$$\mathbf{P}(Z \geq z_m) = \mathbf{P}(Z > z_m) = 1 - \mathbf{P}(Z \leq z_m) = 1 - \Phi(z_m), \tag{2.3.13}$$

where the first equality follows from the fact that $\mathbf{P}(Z = z_m) = 0$ for continuous distributions.

Finally, the corresponding p-value for the general two-tailed test is given by

$$\mathbf{P}(|Z| \geq z_m) = 2\Phi(-|z_m|). \tag{2.3.14}$$

Now we briefly present the computations mentioned above on a real sample of composite material.

Following the previous statistics, our real example is really regular distributed. Of course, it is clear, because of the nonzero diameters of fibres, which centers we are investigating. These results follow from the p-values of all statistics mentioned in Table 2.3.3.

Summary of nearest-neighbor methods

The reduction of point patterns to a one-dimensional nearest-neighbor summary statistics results in a considerable loss of information. Information on individual nearest-neighbor distances is lost. Because distances are measured only to the closest events, only the smallest scales of pattern are considered, and information on larger scales of pattern is unavailable. Nearest-neighbor statistics indicate only the direction of departure from CSR. Little is known about the behavior of these statistics, when CSR does not hold.

Unlike quadrates methods, these statistics do not depend on some arbitrary choice of quadrat size. In conclusion, because much of the spatial information is lost, and because for non-CSR models it is debatable what these statistics are measuring, so nearest-neighbor statistics for mapped data can not be recommended.

2.4 Ripley's K function

Ripley's $K(t)$ function is a tool for analyzing a completely mapped spatial point processes data, i.e. data on the locations of events. Here we describe $K(t)$ for two-dimensional spatial data. Completely mapped data include the locations of all events in a predefined study area. Ripley's $K(t)$ function can be used to summarize a point pattern, estimate parameters and fit models.

The K function is

$$K(t) = \lambda^{-1}\mathbf{E}[\text{ number of events within distance}t \text{ of a randomly chosen event }],$$

where λ is the density (number of fibres per unit area) of events. So, $K(t)$ describes characteristics of the point process at many distances scales. As we have said before, another alternative summaries do not have these property.

$K(t)$ does not uniquely define the point process in the sense that the two different processes can have the same $K(t)$ function. Also, processes with the same $K(t)$ function may have different nearest-neighbor distribution function. Nevertheless, the K function is the basis of routine tools (for descriptive and testing purposes) widely used in the analysis of spatial processes.

For many point processes the expectation in the numerator of the $K(t)$ function can be analytically evaluated, so the $K(t)$ function can be written in a close form. The simplest and most commonly used, is $K(t)$ for a homogeneous Poisson process (CSR):

$$K(t) = \pi t^2.$$

Values of $K(t)$ for a process are often compared with those for the Poisson process. Values larger or smaller than πt^2 respectively indicate a more clustered or more regular process than the Poisson process. In [10], $K(t)$ functions for various types of process are presented in details.

2.4.1 Estimating $K(t)$

Given the locations of all events within a defined study area, $K(t)$ is a ratio of a numerator and the density of events λ. The density can be estimated as $\hat{\lambda} = n/A$, where n is the observed number of points and A is the area of the study region. If edge effects are ignored, then the numerator can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} I(d_{ij} < t),$$

where d_{ij} is the distance between the ith and jth points, and $I(x)$ is the indicator function with the value 1 if x is true and 0 otherwise. **Edge effects** arise because points outside the boundary are not counted in the numerator, even if they are within distance t of a point in the study area. Ignoring edge effects biases the estimator $\hat{K}(t)$, especially at large values of t. A variety of edge-corrected estimators have been proposed, see e.g. [32], [9], [8] or [30]. The most commonly used estimator is

$$\hat{K}(t) = \hat{\lambda}^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} w(l_i, l_j)^{-1}\frac{I(d_{ij} < t)}{n} = \frac{|A|}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} w(l_i, l_j)^{-1} I(d_{ij} < t).$$

As above, d_{ij} is the distance between the ith and jth points, and $I(x)$ is the indicator function. The weight function $w(l_i, l_j)$ provides the edge correction. It has the value of 1 when the circle centered at l_i and passing through the point l_j (i.e. with a radius of d_{ij}) is completely in the study area (i.e. if d_{ij} is larger than the distance from l_i to at least one boundary). If part of the circle falls outside the study area, then $w(l_i, l_j)$ is the proportion of the circumference of that circle that falls in the study area.The effects of edge corrections are more important for large t, because large circles are more likely to be outside the study area.

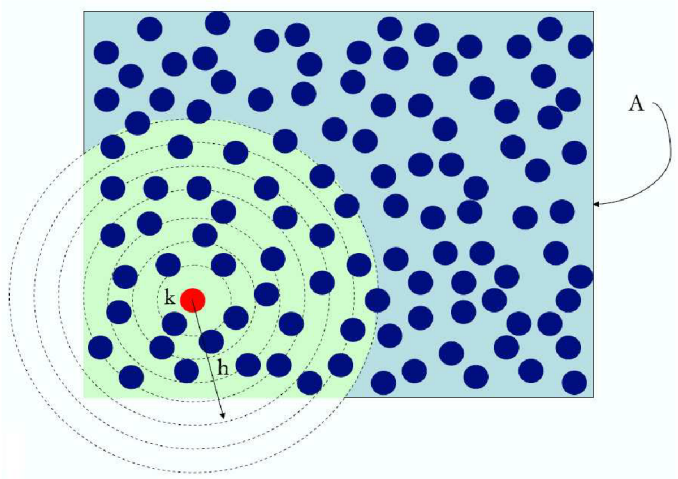


Figure 2.9: *A figure related to explanation to the Ripley's $K(t)$ function.*

The explicit formula for $w(l_i, l_j)$ can be deduced if A is rectangular, see [30]. Although $\hat{K}(t)$ can be determined for any $t > 0$, it is common practice to consider only t less than one-half the shortest dimension of the study area.

The simplest use of Ripley's $K(t)$ function is to test CSR. If CSR of a studied process holds, then $K(t) = \pi t^2$ for all t. In practice, it is easier to use

$$\hat{L}(t) = \sqrt{\frac{\hat{K}(t)}{\pi}},$$

because $\mathbf{D}(\hat{L}(t))$ is approximately under CSR. Under CSR is then then $L(t) = t$. Deviations from the expected value at each distance t are used to construct tests of CSR. One approach is to test $L(t) - t = 0$ at each distance t.

To test whether the data comes from a CSR process, a Monte Carlo test based on the Cramér-von Mises-type statistic

$$k = \int_0^{t_{max}} \left\{ \sqrt{\hat{K}(t)} - \sqrt{K_0(t)} \right\}^2 \mathrm{d}t,$$

where $\hat{K}(t)$ is the estimated K-function of the observed pattern, $K_0(t) = \pi t^2$ is the K-function under the hypothesis of CSR, and t_{max} is the maximum distance for which $\hat{K}(t)$ is computed.

For a given spatial point pattern, $\hat{D}(t) = \hat{K}(t) - \pi t^2$ can be used to evaluate its compatibility with the CSR assumption. The sampling distribution of $\hat{K}(t)$ under the CSR assumption is analytically intractable. However, when A is a rectangle, the variance of $\hat{K}(t)$ can be explicitly expressed, see [9](Lotwick & Silverman) as

$$var_{LS}(t) = \frac{|A|^2}{n(n-1)} \Big(2b(t) - a_1(t) + (n-2)a_2(t) \Big),$$

where

$$a_1(t) = \frac{(0,21Pt^3 + 1,3t^4)}{|A|^2}, \quad a_2(t) = \frac{(0,24Pt^5 + 2,62t^6)}{|A|^3},$$

$$b(t) = \frac{\pi t^2}{|A|}\left(1 - \frac{\pi t^2}{|A|}\right) + \frac{1,0716Pt^3 + 2,2375t^4}{|A|^2},$$

where P denotes the perimeter of A. All the above four equations are exact when t is smaller than or equal to a quarter of the length of the shorter side of A, see [6]. As suggested in [9], $\pm 2\sqrt{var_{LS}(t)}$ can be used as the upper/lower limits for $\hat{D}(t)$. If $\hat{D}(t)$ lies within these limits for all the valid values of t, then the spatial point pattern under investigation can be regarded as compatible to the CSR assumption; otherwise, a deviation from CSR is suggested. In [9] it is suggested to draw a D-curve ($\hat{D}(t)$ and $\pm 2\sqrt{var_{LS}(t)}$ against t) to visualize the CSR test result:

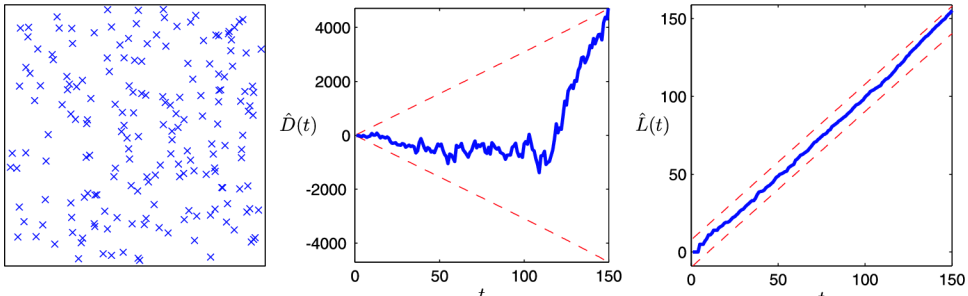


Figure 2.10: *Point pattern following CSR.*

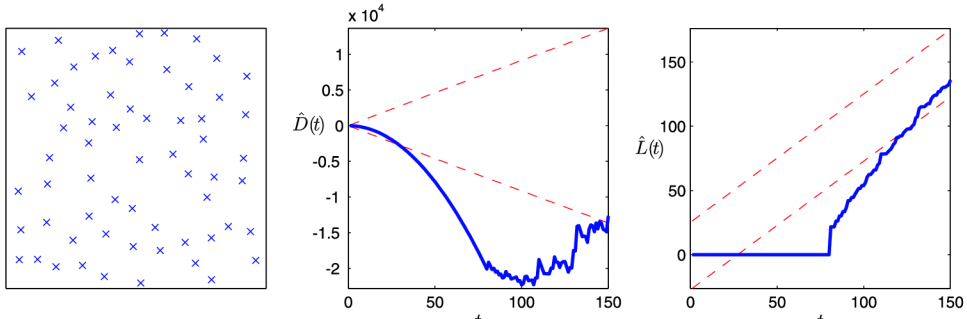


Figure 2.11: *Point pattern tending to regularity.*

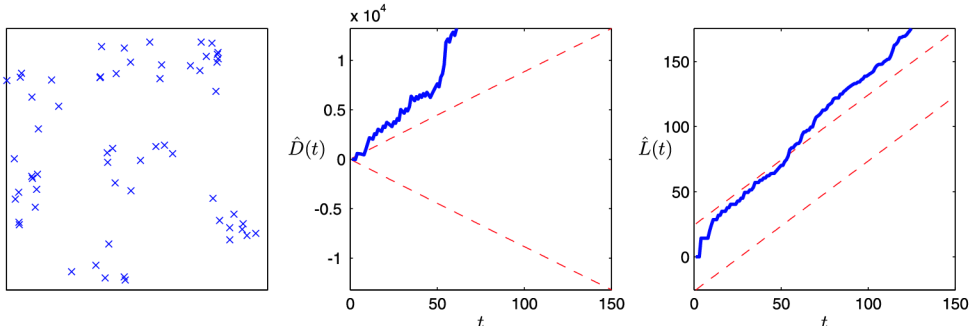


Figure 2.12: *Point pattern tending to clustering.*

Three typical spatial point patterns and their corresponding D-curves are shown in previous Figures. In Figure 2.10, the CSR assumption is supported. The D-curves in Figure 2.11 and 2.12 both suggest obvious deviation from the CSR assumption but in opposite directions. This can be explained by investigating the physical meaning of $\hat{K}(t)$. By definition, $\hat{K}(t)$ is essentially an average of point counts in circles of radius t. If the point pattern under investigation tends to cluster for certain values of t, the point counts in the circles will become much higher than the expectation under the CSR assumption

because it is very probable that a large number of points aggregate "into" the circles. However, if the point pattern has a tendency to regularity, the point counts in the circles will be essentially lower then expectation because t may not be big enough for the circles to "reach" enough number of points. In other words, if $\hat{D}(t)$ is smaller than the lower bound, the pattern tends to regularity; or if $\hat{D}(t)$ is bigger than the upper bound, the pattern tends to cluster; otherwise, the CSR assumption becomes applicable.

The most right graphs in the previous three pictures shows an acceptance region of a 5% test for CSR of n events in a square area $A = [0, a] \times [0, a]$, based on $\hat{L}(t)$, see [8]:

$$\left\{ \left(t - \frac{1,42\sqrt{|A|}}{n}, t + \frac{1,42\sqrt{|A|}}{n} \right) : 0 < t \leq \frac{a}{4} \right\}.$$

3 Microstructural Descriptors

3.1 Introduction

In this chapter we give a brief review of some statistic methods that are used for describing and distinguishing different structures of fibre composite materials.

3.2 Properties of random media

3.2.1 Homogenity and symmetry

The medium is strictly *spatially stationary* or strictly *statistically homogeneous* if the joint probability distributions describing the stochastic process are *translationally invariant*, i.e. invariant under a translation of the origin.

If descriptive functions depend generally on the absolute positions of inclusions, then we say that the medium is *statistically inhomogeneous*. Figure 3.1 depicts two examples of statistically inhomogeneous media, see [37].

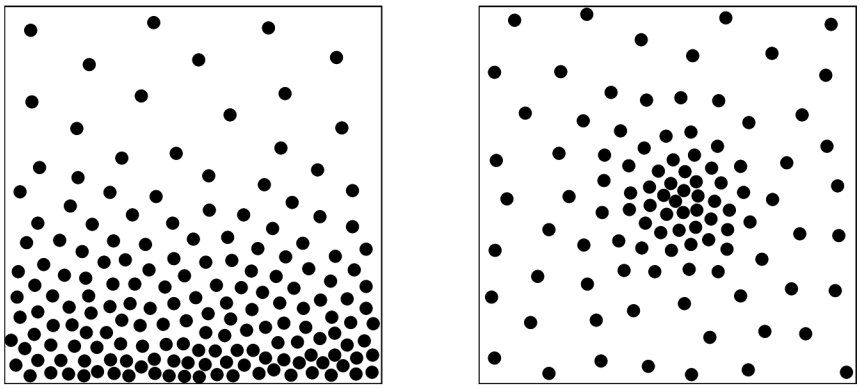


Figure 3.1: *Two examples of statistically inhomogeneous media. Density of the black phase decreases in the upward direction (left panel) and radially from the center (right panel).*

The medium is said to be strictly *statistically isotropic* if the joint probability distributions describing the stochastic process are *rotationally invariant,* i.e. invariant under rotation of the spatial coordinates, see Figure 3.2.

3.2.2 Ergodicity

Usually, the further assumption which is introduced when estimating random fields is the assumption of *ergodicity* of the field. A random field is said to be *ergodic*, when any information about it can be obtained from a single realization. By the term "realization" we understand the event for which a random variable obtains a definite and unique value, see [20]. Thus, complete probabilistic information can be obtained from a single realization

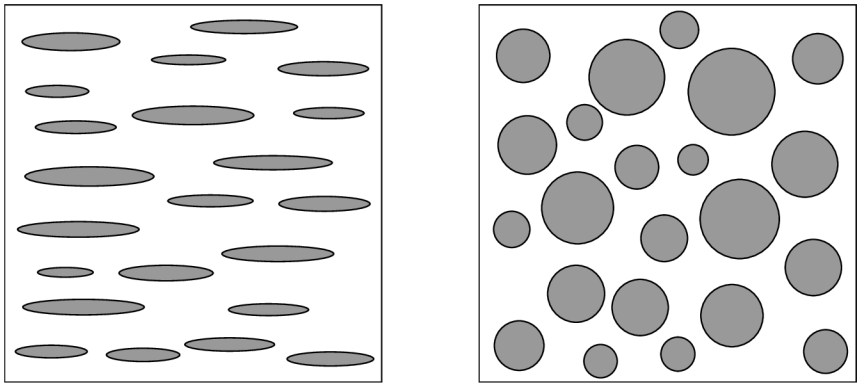


Figure 3.2: *Two examples of portions of statistically homogeneous media. The medium is anisotropic (left panel) and isotropic (right panel).*

of the infinite medium. This suggests an *ergodic hypothesis*, i.e., the result of averaging over all realizations of the ensemble is equivalent to averaging over the volume for one realization in the infinite-volume limit. Thus, complete probabilistic information can be obtained from a single realization of the infinite medium, see [37].

3.3 Statistic description of composites

This section provides some useful functions and formulas for the description of a composite materials:

- indicator function
- $n-$point probability functions
- second order intensity function
- lineal-path function
- nearest-neighbor functions
- pair distribution function

For the sake of simplicity, we will in the next assume only 2D-cases, i.e. cross-section of a material, which is a sufficient condition for us, because we consider "only" composites with unidirectional (paralel) fibres.

Let us consider a composite material made of $i = 1 \dots n$ (in our case $n = 2$) homogeneous and perfectly bounded phases. The volume fraction of the $i-$th phase we denote by ϕ_i.

3.3.1 The indicator function

The use of the term *random heterogeneous material* rests on the assumption that any sample of the medium is a realization of a specific random process. An ensemble is a

collection of all the possible realizations of a random medium generated by a specific stochastic process. Let us denote $(\Omega, \mathcal{F}, \mathbf{P})$ be some fixed probability space. Let each point $\omega \in \Omega$ corresponds to a realization of the random medium, see [37].

Each realization ω of the two-phase random composite random medium occupies the region of space $V \subset \mathbb{R}^2$ that is partitioned into two disjoint random phases: phase 1 of a region $V_1(\omega)$ and volume fraction ϕ_1, and phase 2 of a region $V_2(\omega)$ and volume fraction ϕ_2. The random sets $V_1(\omega)$ and $V_2(\omega)$ are the complements of each other, i.e. $V_1(\omega) \cap V_2(\omega) = \emptyset$ and $V_1(\omega) \cup V_2(\omega) = V$. Figure 3.3 shows a portion of a realization of a two-phase random medium. For a given realization ω, the *indicator function* $\mathcal{I}^{(r)}(\boldsymbol{x}, \omega)$ for phase r is given for $\boldsymbol{x} \in V$ by

$$\mathcal{I}^{(r)}(\boldsymbol{x}, \omega) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in V_r(\omega), \\ 0 & \text{otherwise.} \end{cases}$$

Next we will denote by index r the following:

$$r = \begin{cases} m \text{ instead of } 1 & \text{for a matrix,} \\ f \text{ instead of } 2 & \text{for a fiber.} \end{cases}$$

For such system the indicators functions $\mathcal{I}^{(f)}(\boldsymbol{x}, \omega)$ and $\mathcal{I}^{(m)}(\boldsymbol{x}, \omega)$ are related by

$$\mathcal{I}^{(f)}(\boldsymbol{x}, \omega) + \mathcal{I}^{(m)}(\boldsymbol{x}, \omega) = 1.$$

Unless otherwise stated, we will drop ω from the notation (as it is usual) and write $\mathcal{I}^{(r)}(\boldsymbol{x})$ instead of $\mathcal{I}^{(r)}(\boldsymbol{x}, \omega)$.

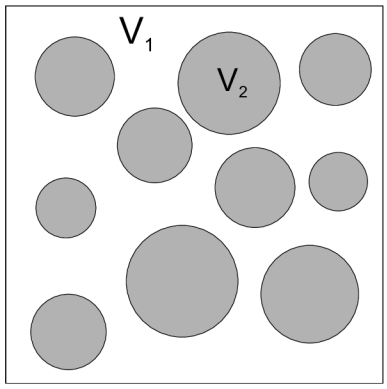


Figure 3.3: *Two-phase fibre composite material with phases V_1 and V_2.*

3.3.2 $n-$point probability functions

Now, we describe a set of general $n-$point probability functions, applicable to an arbitrary two-phase composite.

Definitions

The probabilistic description of $\mathcal{I}^{(r)}(\boldsymbol{x})$ is given by the probability that $\mathcal{I}^{(r)}(\boldsymbol{x})$ is equal to one, which we write as

$$\mathbf{P}\left\{\mathcal{I}^{(r)}(\boldsymbol{x}) = 1\right\}.$$

Given this probability, it follows that

$$\mathbf{P}\left\{\mathcal{I}^{(r)}(\boldsymbol{x})=0\right\}=1-\mathbf{P}\left\{\mathcal{I}^{(r)}(\boldsymbol{x})=1\right\}.$$

One-point probability function. Using the indicator function as it has been defined above, the probability of the location $\boldsymbol{x}$ belonging to phase r is defined by the ensemble average (denoted by angular brackets $\langle\cdot\rangle$) of the function $\mathcal{I}^{(r)}(\boldsymbol{x})$, see [37]:

$$S_1^{(r)}(\boldsymbol{x})\equiv\left\langle\mathcal{I}^{(r)}(\boldsymbol{x})\right\rangle=\mathbf{P}\{\mathcal{I}^{(r)}(\boldsymbol{x})=1\}.$$

The *one-point probability function* (also known as one-point correlation function) described in equation above is normally difficult to compute. However, if the material is assumed to be statistically homogeneous and ergodic, the following simplifications can be considered, see [20]:

$$S_1^{(r)}(\boldsymbol{x})=\lim_{V\to\infty}\int_V\mathcal{I}^{(r)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=\phi_r,$$

where symbol ϕ_r denotes the volume fraction of the phase r.

If we sample the domain V with a set of locations $\boldsymbol{x}_i$ with $i=1,\ldots,n$, then ϕ_r can be estimated easily:

$$\widehat{\phi_r}=\frac{1}{n}\sum_{i=1}^{n}S_1^{(r)}(\boldsymbol{x}_i),\quad r=f,\ m.$$

General n-point probability functions. Knowing a realization $V_r(\omega)$ is the same as knowing $\mathcal{I}^{(r)}(\boldsymbol{x},\omega)$ for all $\boldsymbol{x}\in V$. Therefore, we may regard the random set $V_r(\omega)$ as the collection of all random variables $\mathcal{I}^{(r)}(\boldsymbol{x})$ for $\boldsymbol{x}\in V$. Hence, the probability law of $V_r(\omega)$ is described by the finite-dimensional distributions of the random process $\left\{\mathcal{I}^{(r)}(\boldsymbol{x}):\boldsymbol{x}\in V\right\}$. Since the $\mathcal{I}^{(r)}(\boldsymbol{x})$ are either 0 or 1, this allows to specify the probabilities, see [37]:

$$\mathbf{P}\left\{\mathcal{I}^{(r)}(\boldsymbol{x}_1)=j_1,\mathcal{I}^{(r)}(\boldsymbol{x}_2)=j_2,\ldots,\mathcal{I}^{(r)}(\boldsymbol{x}_n)=j_n\right\},$$

where each of numbers $j_k,\ k=1,\ldots,n$ is either 0 or 1.

The expectation of the product $\mathcal{I}^{(r)}(\boldsymbol{x}_1)\mathcal{I}^{(r)}(\boldsymbol{x}_2)\ldots\mathcal{I}^{(r)}(\boldsymbol{x}_n)$ is a very important average. Similarly, see see [37], as in the case of one-point probability function we get:

$$\begin{aligned}S_n^{(r)}(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n)&\equiv\left\langle\mathcal{I}^{(r)}(\boldsymbol{x}_1)\mathcal{I}^{(r)}(\boldsymbol{x}_2)\ldots\mathcal{I}^{(r)}(\boldsymbol{x}_n)\right\rangle=\\&=\mathbf{P}\left\{\mathcal{I}^{(r)}(\boldsymbol{x}_1)=1,\mathcal{I}^{(r)}(\boldsymbol{x}_2)=1,\ldots,\mathcal{I}^{(r)}(\boldsymbol{x}_n)=1\right\},\end{aligned}$$

which features the probability that n points at positions $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n$ are found in phase r. According to see [37] we will refer to $S_n^{(r)}$ as the *n-point probability function* for phase r.

It is possible to express the probability $S_n^{(f)}$ of finding n points in phase formed by fibres (f) in terms of the set of phase formed by matrix (m) by means of probabilities $S_1^{(m)}, S_2^{(m)}, \ldots, S_n^{(m)}$. It is not difficult to show that:

$$\begin{aligned} S_n^{(f)}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) &= \left\langle \prod_{j=1}^{n} [1 - \mathcal{I}^{(m)}(\boldsymbol{x}_j)] \right\rangle \\ &= 1 - \sum_{j=1}^{n} S_1^{(m)}(x_j) + \sum_{j<k}^{n} S_2^{(m)}(\boldsymbol{x}_j, \boldsymbol{x}_k) \\ &- \sum_{j<k<l}^{n} S_3^{(m)}(\boldsymbol{x}_j, \boldsymbol{x}_k, \boldsymbol{x}_l) + \cdots + (-1)^n S_n^{(m)}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n). \end{aligned}$$

Remark 3.3.1. The probability that a point at $\boldsymbol{x}_1$ is in the phase f and a point at $\boldsymbol{x}_2$ is in the phase m is given by

$$S_2^{(fm)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left\langle \mathcal{I}^{(f)}(\boldsymbol{x}_1)[1 - \mathcal{I}^{(f)}(\boldsymbol{x}_2)] \right\rangle = S_1^{(f)}(\boldsymbol{x}_1) - S_2^{(f)}(\boldsymbol{x}_1, \boldsymbol{x}_2).$$

Geometrical interpretation of $S_n^{(r)}$. Let $F_n^{(r)}$ be a polyhedron with n vertices located at positions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Then for statistically *inhomogeneous* media, $S_n^{(r)}$ is the probability that all n vertices of $F_n^{(r)}$ with *fixed* positions $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ lie in V_r. For statistically *homogeneous* but *anisotropic* media, $S_n^{(r)}$ is the probability that all n vertices of $F_n^{(r)}$ lie in V_r when the polyhedron is randomly placed in the volume at fixed orientation, i.e. over all translations of the polyhedron. For statistically *isotropic* media, $S_n^{(r)}$ can be interpreted as the probability that all n vertices of $F_n^{(r)}$ lie in V_r when the polyhedron is randomly placed in the volume, i.e. over all translations and solid-body rotations of the polyhedron, see [37].

Remark 3.3.2. As we said before, the medium is statistically homogeneous, if the joint probability distributions describing the stochastic process are translationally invariant. Then we can write, see [37], for $n-$point probability functions for phase r:

$$\begin{aligned} S_n^{(r)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) &= S_n^{(r)}(\boldsymbol{x}_1 + \boldsymbol{y}, \boldsymbol{x}_2 + \boldsymbol{y}, \ldots, \boldsymbol{x}_n + \boldsymbol{y}) \\ &= S_n^{(r)}(\boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{1n}), \end{aligned}$$

for all $n \geq 1$, where $\boldsymbol{x}_{jk} = \boldsymbol{x}_k - \boldsymbol{x}_j$ and $\boldsymbol{y}$ is a constant vector. According to the previous notation, we can write the probability functions $S_2(\boldsymbol{r})$ or $S_3(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{t})$ for two- or three-point probability functions.

The one-point function S_1 is obtained by randomly throwing a single point onto the planar section many times and recording the fraction of times that it lands in one of the phases, say fibres in Figure 3.4. Thus, S_1 (if the number of attempts is sufficiently large) is the probability that a single point falls in the white phase. The two-point correlation function $S_2(\boldsymbol{r})$ is obtained by randomly throwing a line segment of length r into the sample many times and recording the fraction of times that its end points land in the

fibres, see Figure 3.4. By performing this experiment for all possible lengths $\boldsymbol{r}$, one can generate a graph of S_2 as a function of $\boldsymbol{r}$. Therefore, $S_2(\boldsymbol{r})$ is the probability that the two end points of a line segment of length $\boldsymbol{r}$ fall in the fibres. Clearly, variations in $S_2(\boldsymbol{r})$ contains more information than S_1, which is just a constant. Similarly, $S_3(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{t})$ is the probability that the three vertices of a triangle with sides of lengths $\boldsymbol{r}, \boldsymbol{s}$ and $\boldsymbol{t}$ fall in the fibres. The three-point probability S_3 gives more information than S_2. In general, S_n gives the probability that n points with specified positions lie in the fibres, see Figure 3.4.

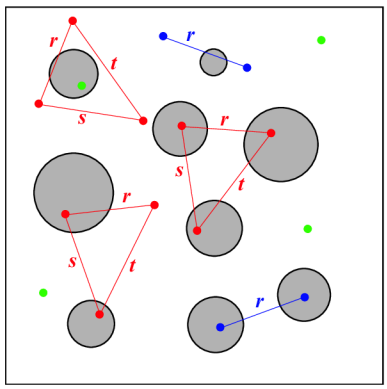


Figure 3.4: *A scheme showing attempts at sampling for the correlation functions* S_1, S_2 *and* S_3 *from a planar section.*

The probability of finding the phase r at the point $\boldsymbol{x}_i$ and the phase s at the point $\boldsymbol{x}_j$ (in other words, two-point probability function) can be expressed, see [20], [39], [38], [12], [37].

$$S_2^{(r,s)}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \overline{\mathcal{I}^r(\boldsymbol{x}_i)\mathcal{I}^s(\boldsymbol{x}_j)} \rangle.$$

Generally, we can define *n-point probability functions* as:

$$S_n^{(r_1,\ldots,r_n)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \langle \overline{\mathcal{I}^{r_1}(\boldsymbol{x}_1) \ldots \mathcal{I}^{r_n}(\boldsymbol{x}_j)} \rangle,$$

which gives the probability of finding n points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ randomly thrown into a medium located in the phases $r_1, \ldots, r_n$.

Hereafter, we limit our attention to functions of the order one and two, since higher-order functions are quite difficult to determine in practice, see [38], [37]. Therefore, description of a random medium will be provided by the *one-point probability function* and by the *two-probability function.*

One-point probability function

As we said before, the one-point probability function S_1 is obtained by randomly throwing a single point onto the planar section many times and recording the fraction of times that it lands in one of the phases.

Two-point probability function

Let us remark $S_2(\boldsymbol{r}) \equiv S_2(x_1, x_2)$, where $\boldsymbol{r} = x_1 - x_2$. As noted earlier, the two-point (sometimes called *autocorrelation function*) $S_2(\boldsymbol{r}) \equiv S_2^{(1)}(\boldsymbol{r})$ for statistically homogeneous media can be obtained by randomly tossing a line segments of length $r \equiv |\boldsymbol{r}|$ with a specified orientation and counting the fraction of times the end points fall in phase 1. The function $S_2(\boldsymbol{r})$ provides a measure of how the end points of a vector $\boldsymbol{r}$ are correlated. For isotropic media, $S_2(r)$ attains its maximum value of ϕ_1 at $r = 0$ and decays(usually exponencially fast) to its asymptotic value ϕ_1^2. For explanation of asymptotic properties and bounds of $S_n^{(i)}$ see [37], paragraph 2.2.4 for more details.

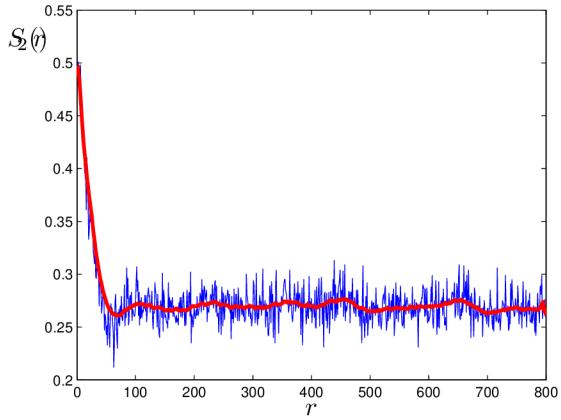


Figure 3.5: *Two-point probability function.*

3.3.3 Lineal-path function

Another interesting and useful statistical measure is what we call the *lineal-path function* $L^{(r)}(t)$, see [37]. For statistically isotropic media, it is defined as follows:

$$L^{(r)}(t) = \mathbf{P}(\text{a line segment of length } t \text{ lies wholly in a phase } r, \text{ when randomly thrown into the sample.})$$

The lineal-path function is a monotonically decreasing function of t, since the space available in phase r to a line segment of length t decreases with increasing t. At the extreme values of $L^{(r)}(t)$ we have

$$L^{(0)}(t) = \Phi_r, \qquad L^{(r)}(\infty) = 0.$$

The "tail" of $L^{(r)}(t)$ (i.e., large t behavior) provides information about the largest lineal paths in phase r. If we define $L^{(12)}(t)$ to be the probability that a line segment of length t intersects any parts of the two-phase interface when randomly thrown into the sample, then it is clear that

$$L^{(1)}(t) + L^{(2)}(t) + L^{(12)}(t) = 1.$$

In the next figure we can see one realization of the lineal path function for fibres(left) and for the matrix(right). By the blue curve represents this function obtained by Monte-Carlo method and the red curve is smoothed the blue one.

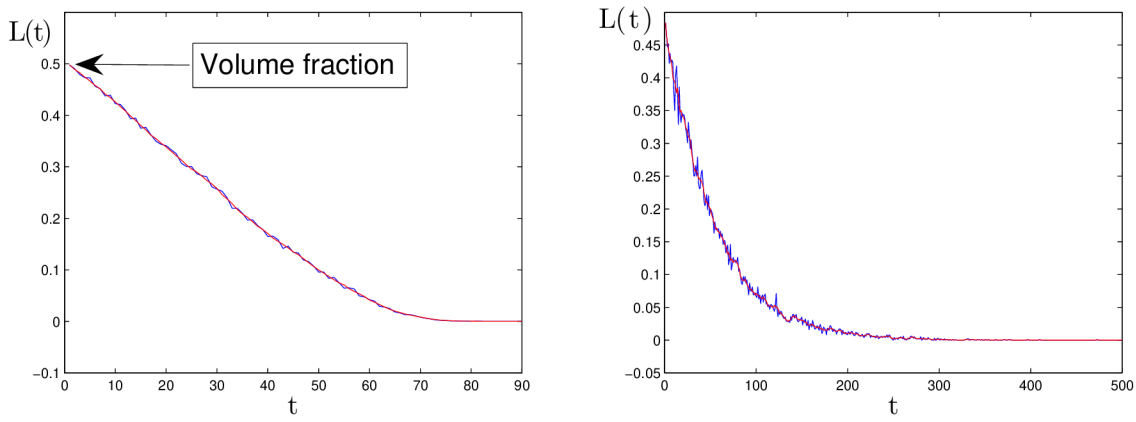


Figure 3.6: *Lineal path function for fibres(left) and matrix(right).*

Now, we will give brief definitions of functions which help identifying the type of distributions found on spatial patterns. They consists of Ripley's K function, which can be classified as a second-order measure, and distance measures which include the G (nearest neighbor), F (empty-space) and J functions.

3.3.4 Second order intensity function (Ripley's K-function)

The most important function of second order is Ripley's $K-$ function, as was said in Section 2. In that section is presented a detailed description of this function.

In the next figure we can see an example of the Ripley's K-function for real composite. The shape of the function was computed as a mean of 15 real samples. In the left figure there are plotted 15 $K-$functions corresponding to various samples of real composite and on the right figure there is an average $K-$function with blue dotted function, corresponding to CSR (Poisson process).

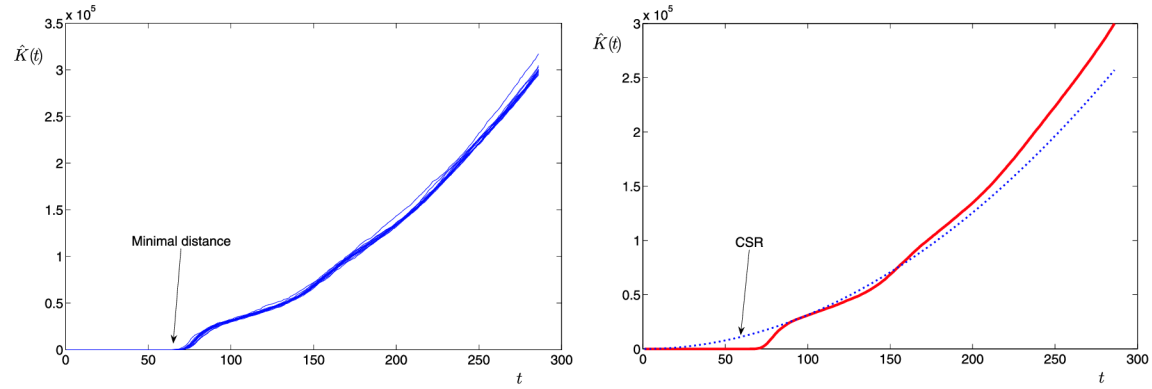


Figure 3.7: *Ripley's $K(t)$ function for the real composite.*

3.3.5 Nearest neighbor function

The G or *nearest neighbor distance distribution function* is a relatively simple description of the spatial distributions of the points based on the measurement of the distance from a typical point of the pattern to its nearest neighbor. So, this statistic focuses on short-range interactions between points, see [32]. The nearest neighbor distribution function $G(t)$ is defined for $t \in \mathbb{R}^+$ by:

$$G(t) = \mathbf{P}(\text{the circle of radius } t, \text{ centered on an arbitrary object, contains at least one other object}).$$

According to [32], an equivalent definition of $G(t)$ is defined by:

$$G(t) = \mathbf{P}(\text{the distance between an arbitrary object and its nearest neighbor, is less than or equal to } t).$$

An obvious way of estimating the G function from an observed pattern is simply to calculate for each point within a sampling space the distance to its nearest neighbor and we get an empirical G function, i.e.:

$$\hat{G}(t) = n^{-1} \times (\text{number of points whose nearest neighbor distance is } \leq t)$$

A practical computation proceed as follows: We denote by n a number of points in a certain region A and by y_i denote the distance from the $i-$th point to the nearest other point in A, i.e. the nearest neighbor point. The distances y_i are called *nearest neighbor distances.*

The simplest and most natural estimator of the G function is given by the proportion of members of an event set for which the distance to the nearest other member of the set $\{y_i\}_{i=1}^n$ is less than or equal t. It is provided by the following function:

$$\hat{G}(t) = n^{-1} \#(y_i < t),$$

where $\#(\cdot)$ is the *counting function* which records the number of points in the specified set. In the next figure we can see an average estimation of the G-function from the real 15 samples of composite:

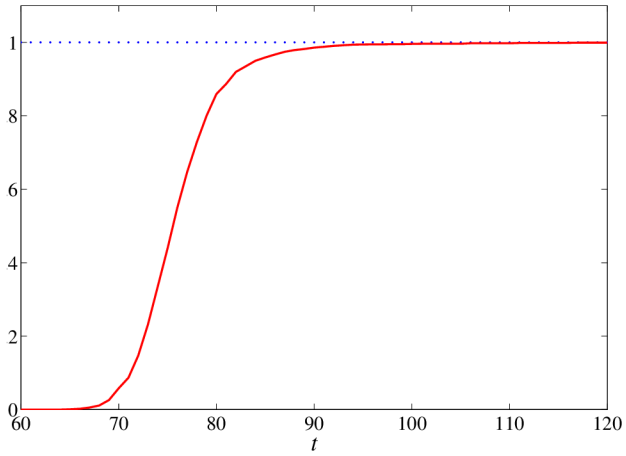


Figure 3.8: *Estimation of the $G(t)$ function for the real composite.*

3.3.6 Empty space function

Another descriptor of random pattern is the so called *empty space distribution function* (also called F-function), which is closely related to the nearest neighbor function. A definition of the F-function is given by:

$$F(t) = \mathbf{P}(\text{the distance between an arbitrary point and its nearest event is at most } t),$$

especially for two-dimensional spaces the definition can be rewritten like this:

$$F(t) = \mathbf{P}(\text{the circle of radius } t, \text{ centered on an arbitrary point contains at least one event}).$$

The empirical distribution function of the F function can be estimated by counting the distances d_i (from each of the m fixed points to the nearest point in the sample) that are less than t and dividing this total by the overall number of fixed points, m:

$$\hat{F}(t) = m^{-1}\#(d_i < t).$$

The choice of a number of m fixed points is not exactly defined. For example, according to [9] it is recommended to place m fixed points into a $k \times k$ grid, where $k \approx \sqrt{n}$.

In the next figure we can see the estimation of the F function for real sample:

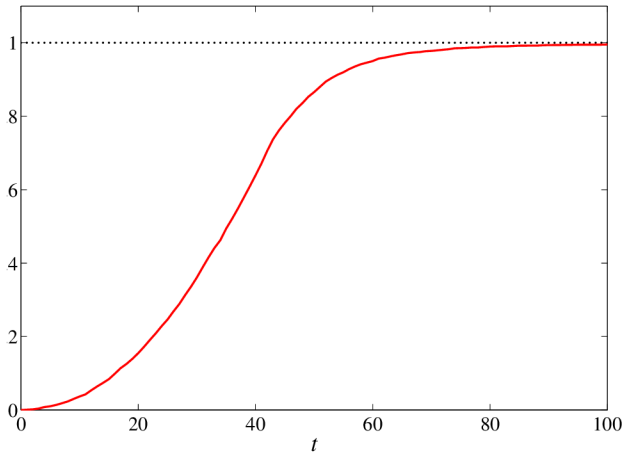


Figure 3.9: *Estimation of the $F(t)$ function for the real composite.*

The estimation of $F(t)$ is complicated by the bounded nature of the pattern being studied. As the distances t increase, the position of the nearest neighbor to a point will only be known with certainty for those points lying within the interior of the study region. Thus, edge effects play a significant role in the estimation of $F(t)$. For a Poisson process of intensity λ on a two dimensional space one obtains:

$$1 - F(t) = \exp(-\lambda\pi t^2), \ t \geq 0$$

and therefore, just like for the nearest neighbor distribution function, $F(t)$ is given by:

$$F(t) = 1 - \exp(-\lambda\pi t^2), \ t \geq 0.$$

Values of $F(t)$ greater than the Poisson value suggest that there is regularity or ordering in the point pattern; lower values suggest aggregation, see [32].

3.3.7 The J function

The J function was first introduced by Baddeley and van Lieshout who stated that the strength and range of interpoint interactions in a spatial point process can be qualified by the $J(t)$ function given by, see [5]:

$$J(t) = \frac{1 - G(t)}{1 - F(t)}, \ t > 0 \quad \text{such that } F(t) < 1,$$

where $G(t)$ is the nearest-neighbor distance distribution function and $F(t)$ is the empty-space function of the process. The J function is a nonparametric measure of the type of spatial interaction. The values of $J(t)$ function less or greater than one are indicative of clustering or regularity, respectively. If the point process is stationary and Poisson, then $F(t) \equiv G(t)$ and so $J(t) = 1$. Just as for the K function, J does not depend on the intensity parameter, a feature that affects both F and G. Also note, that $J(0) = 1$. In the next figure we can see an average of the J-function from 15 samples of the real composite.

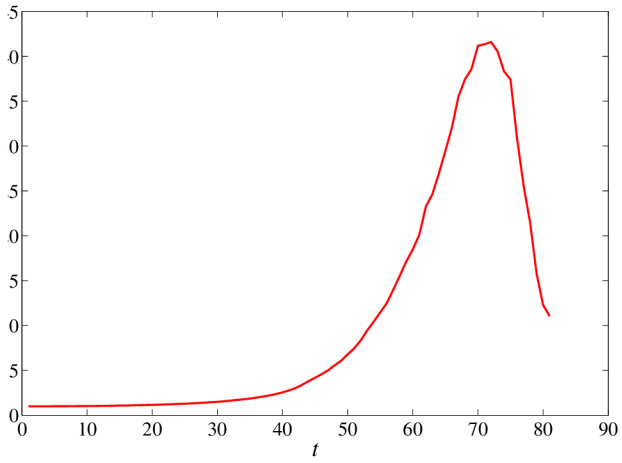


Figure 3.10: *The $J(t)$ function for the real composite.*

3.3.8 Pair distribution function

According to [20], *the pair distribution function* $g(t)$ describes the probability of finding an inclusion whose center lies in an infinitesimal circular region of radius dt about the point t, provided that the coordinate system is located at the center of a second inclusion. Next, according to [20] we can get the following relation between $g(t)$ and $K(t)$:

$$g(t) = \frac{1}{2\pi t} \frac{\mathrm{d}K(t)}{\mathrm{d}t}.$$

Although $g(t)$ and $K(t)$ are related, they provide quite different physical information. The Ripley's function $K(t)$ can distinguish between different patterns and detect regularities, whereas the *pair distribution function* $g(t)$ describes the occurrence intensity of inter-inclusion distances. In this function a local maxima indicates the most frequent distances between points and a local minima the least frequent ones in the pattern.

The following discretized estimation of pair distribution function for our computations was used, see [20]:

$$g(t) = \frac{1}{2\pi t \rho \mathrm{d}t} \frac{1}{N} \sum_{i=1}^{N} n_i(t),$$

where t is the radial distance from a fibre center, ρ the number of fibres per unit area, N the total number of fibre centers in the region considered, n_i the number of fibre centers which lay within an annulus of radius t and thickness $\mathrm{d}t$, with the same center as the fibre i. If the values of $g(t)$ are greater than one, the corresponding distances occur more frequently than in a complete random pattern, and conversely for smaller values.

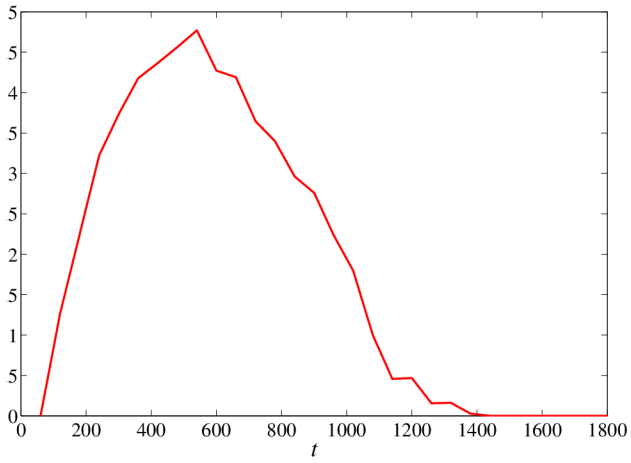


Figure 3.11: *The pair distribution function of the real composite.*

4 Applied Algorithms

4.1 Basic Terms

The mechanical behavior of composite materials in strongly dependent on the geometrical arrangement of distinct phases of the composite– *microstructure.* Unfortunately, microstructure of real composite materials is typically quite complicated. To illustrate this fact, we present a high-contrast micrograph of a part of the graphite fiber tow impregnated by the polymer matrix, see figure 4.1.

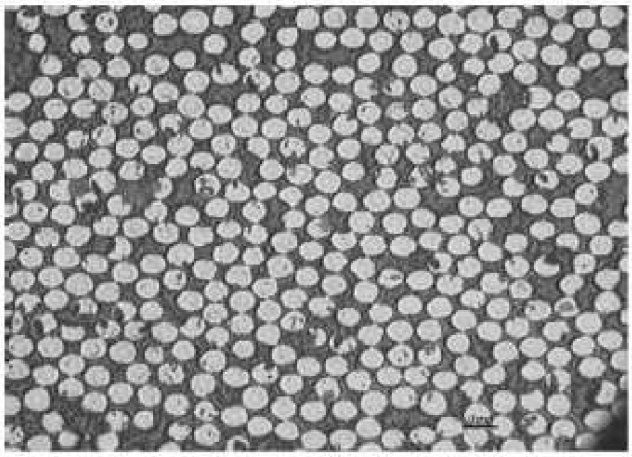
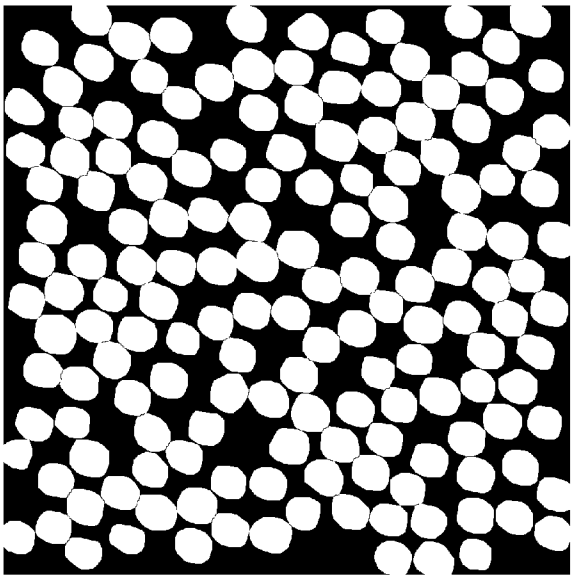
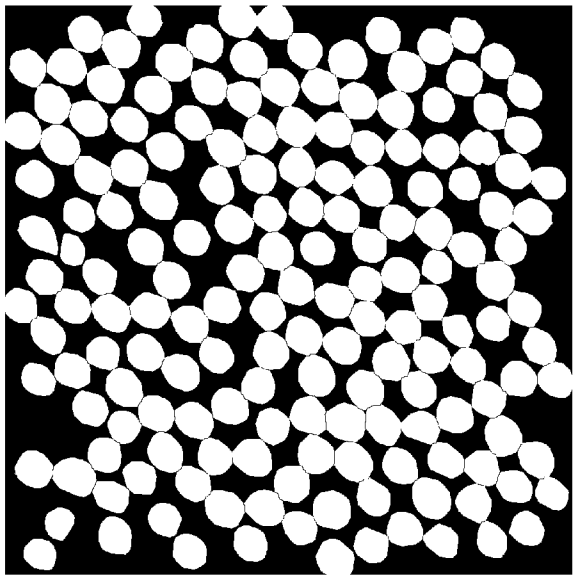


Figure 4.1: *A micrograph of a transverse plane section of a real graphite fiber tow.*

Before starting our description of the developped algorithms it is natural to describe the real composite, that was used as a "starting" model. All of them come from the photos of a real composite, that were gained from Ing. Jan Zeman, Ph.D from Czech Technical Univerzity in Prague. To see a sample of sent photos see the following figures. For more information about separating real structure from the photos see [12].





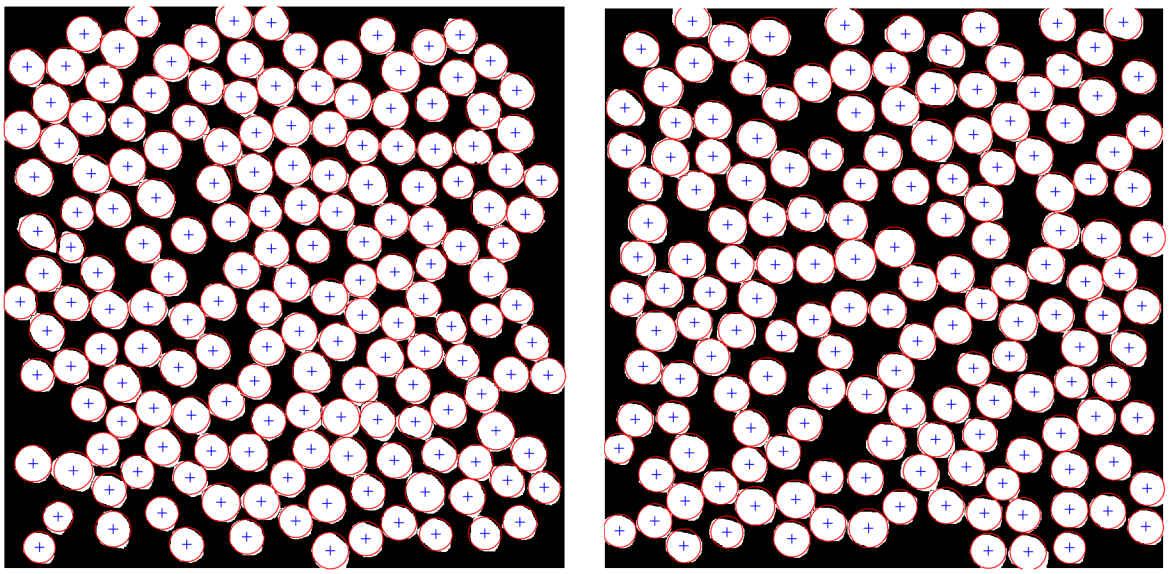


Figure 4.2: *Original(up) and corrected figures(down).*

In the upper pair of figures you can see two different real samples in the "raw" state and on the third and fourth figures the same samples with drawn circles representing fibres with their centers. We used an *Image Processing Toolbox for Matlab* to the next manipulation: First of all we determined surface areas of each white region representing one fibre. In the second step we interleaved an ellipse with the same area by the given region to get the center of a fibre(blue crosses) and then we placed a circle with the same area and center as the ellipse(red circles). After this steps we have at disposal model of a real material–non-constant diameters of fibres and non-periodic structure. From this data we stated the distribution of fibre's diameters that we used in the following algorithms.

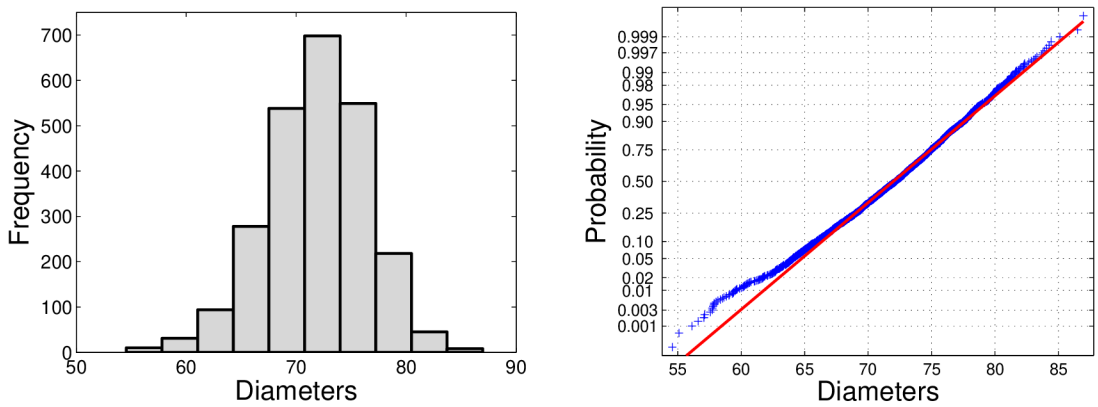


Figure 4.3: *Histogram(left) and normal probability plot of fibre's diameters.*

The resulting distribution agrees with normal distribution $N(71,87;\ 20,96)$.

It is good to remark, that major existing results dealing with generating random structures come from the fact, that the centers of the fibres are not nearer then diameter of a fibre, which is considered to be constant for all fibres in a sample. Algorithms generating such structures are based on so called *spatial point processes*, see [26] and

references therein. But, the previous ones differ from the algorithms presented later. Newly developed algorithms enable to work with non-constant diameters of the fibres with keeping the same *volume fraction,* i.e. the ratio of the area filled by fibres and total area of a sample. For our set of fifteen samples the mean value for total volume fractions amount 0,4869.

4.2 Algorithm AI

The main idea of this algorithm is based on stochastic process $S(t, \omega)$, see Appendix 13 for more information. This process, more specifically its separate trajectories, has a character of a *"wavy-random sinusoid curve"*. In other words, they have different amplitudes and periods. For the better imagination see figure 13.2 in Appendix.

Let D be a domain, representing our sample, into which we want to place fibres with random diameters corresponding to the established distribution of the real samples.

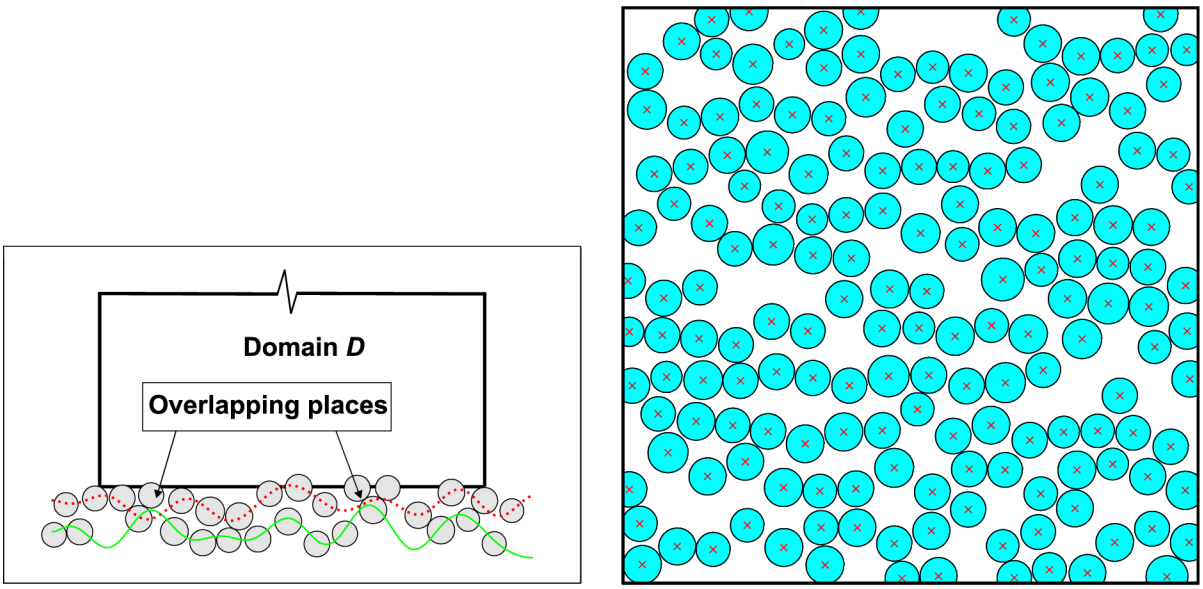


Figure 4.4: *To the description of the algorithm* ***AI****(left) and the final structure generated by the algorithm* ***AI****(right).*

A detailed procedure can be described in several steps. In the bottom of the picture there is a green curve which forms the centers of the fibres with random diameter. This green curve was received by means of one realization of the stochastic process $S(t, \omega)$ with $K = 3800$. During putting the fibres on the line we also have to check overlapping of the fibres. After the green line is filled we continue with a red one, which is generated as the green one, but is shifted up. Again, the fibres are placed to the line and checked with existing ones. In the case of overlapping(the arrows in the figure) they are shifted to the "safe" distance. In this way we continue until the whole domain D is not filled up. In the next figure there is finished a resulting structure of a two-phased fibre composite material according to the algorithm **AI**. By a different choice of a number K in a expansion of

the stochastic process $S(t, \omega)$ we change an amplitude and period of a curve. This fact causes, that we are able to generate structures with various volume fraction and number of fibres in a sample. Of course, it is possible to set a minimum distance between two fibres. By means of this algorithm fifteen samples were generated for the purpose to the next computations.

4.3 Algorithm AII

The principle of this algorithm we can describe as follows: Firstly, we generate one fibre with random diameter and situate it approximately in the middle of the domain sample. Then we choose a random direction and a distance, where we put a new fibre. This procedure is repeated until the resulting volume fraction does not reach the requested one. During every step we are checking whether a new fibre does not cross the existing ones. In case of overlaying of fibres, new position is generated. For the better illustration and result see the following pictures.

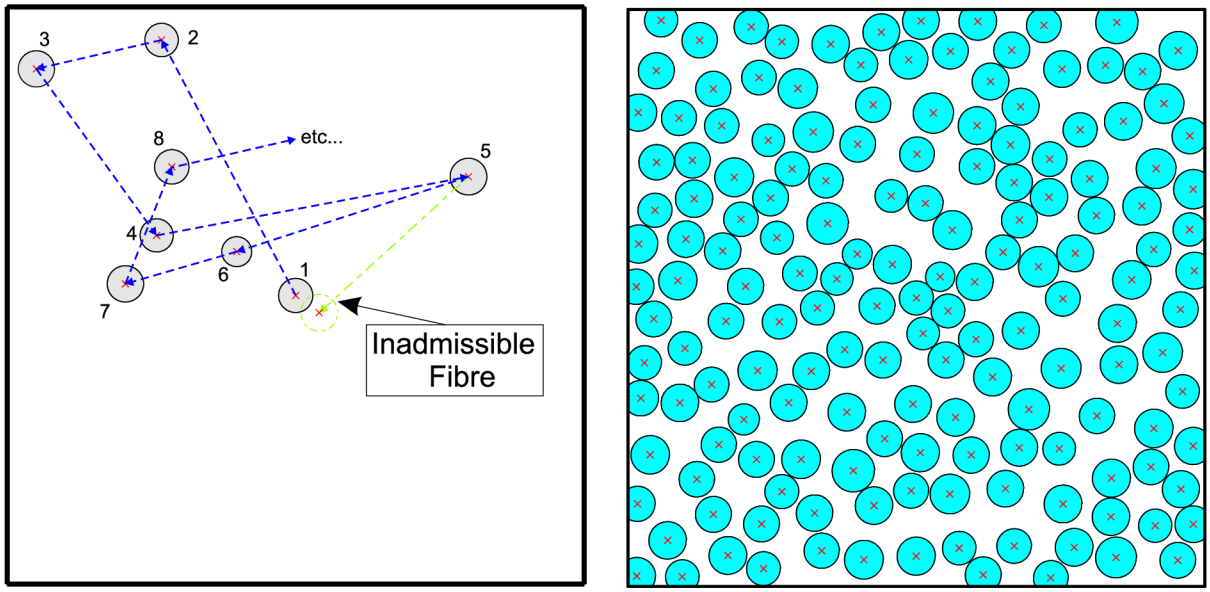


Figure 4.5: *To the description of the algorithm* ***AII****(left) and the final structure generated by the algorithm* ***AII****(right).*

4.4 Algorithm AIII

It is based on the Brownian motion of the suspended particles in a liquid medium. The simulation starts with generating a sample with complete periodic structure, i.e. constant diameters of fibres and the same distance between them. The diameter must be chosen in such a way, that the resulting volume fraction has the same value as in the real samples. After such structure is generated, the diameters of all fibres are changed according to the distribution of real samples. In a next step each Fibre is submitted to

the Brownian motion. In other words, we choose a random direction and distance of shifting a fibre. Simultaneously we check for the crossing with neighboring fibres and the minimum distance between them. If it occurs, we change the direction and the distance and the process is repeated. This is repeated until everything is all right. It is important to note, that generated amplitude of vibrations are in tenths of fiber's diameter, so they are relative small. This fact corresponds to the real concept of the Brownian motion, but we do not consider the collisions of particles and transmitting their quantity of motion during collision of one particle into another one.

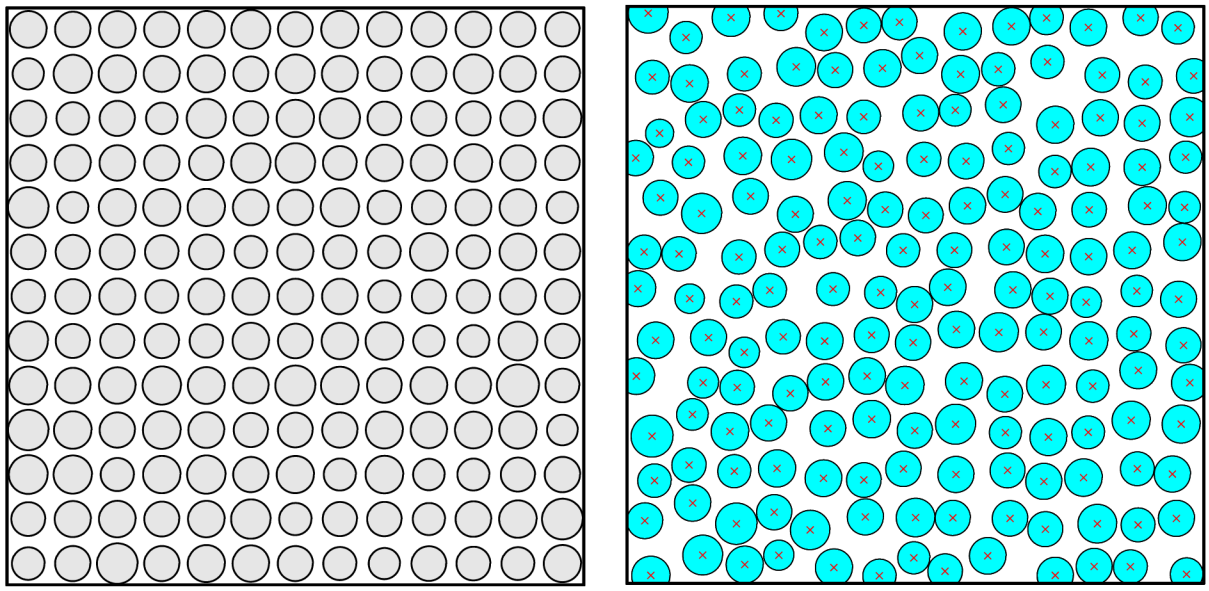


Figure 4.6: *To the description of the algorithm* ***AIII****(left) and the final structure generated by the algorithm* ***AIII****(right).*

4.5 Algorithm AIV

The principle of this algorithm is similar to the algorithm **AIII**. The difference is in processing overlaying of fibers: if the deflection of the fibre will cause overlaying with neighboring fiber, the shifting is canceled – the fiber stays in its old position. It causes, that the final structure is not so random as in the case of algorithm **AIII**, but the computing time is several times shorter. The only disadvantage of algorithms **AIII** and **AIV**is in a fact, that the amount of fibres is the same for all samples. We have to note that in each of the previous algorithms, the diameters of the fibres are driven by a known probability distribution.

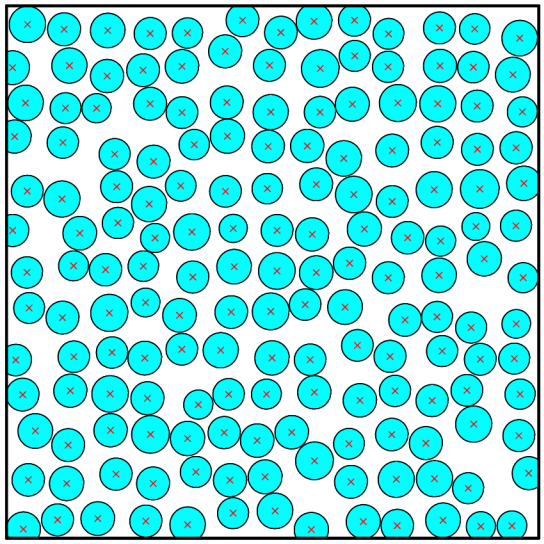


Figure 4.7: *The final structure generated by the algorithm* ***AIV****.*

Part II

Statistical Computations

I have a dream...

5 Disposal Data

According to the previous chapter, we had at disposal a file of fifteen square-shaped samples of a real two-fibre composite material. On the basis of the data obtained from these samples, we constructed four different algorithms (**AI-AIV**) generating analogous inner structure with almost the same volume fraction. Briefly speaking, we simulated fifteen different samples of the same size per each algorithm and therefore we have at disposal a set of seventy-five samples – i.e. five types per fifteen realizations.

Our aim is now to compare samples generated by algorithms **AI**–**AIV** to real ones. As a tool to this comparison we use e.g. descriptive statistics, methods of analysis of variance, variograms, etc. All these techniques will be presented in the next chapters.

6 Descriptive Statistics

6.1 Introduction

Descriptive statistics are used to describe the basic features of the data gathered from an experimental study in various ways. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. It is necessary to be familiar with primary methods of describing data in order to understand phenomena and make intelligent decisions. Various techniques that are commonly used are classified as:

- **Graphical displays** of the data in which graphs summarize the data or facilitate comparisons.
- **Tabular description** in which tables of numbers summarize the data.
- **Summary statistics** (single numbers) which summarize the data.

The summary statistics we can divide among these groups:

- **Location** - mean median, mode
- **Dispersion** - range, standard deviation
- **Moments** - variance, skewness, kurtosis

To start an analysis based on the descriptive statistics we need to receive detailed data from the samples. One of the possible ways is to imaginary divide requested sample by 10x10 grid into one hundred cells. After this procedure we compute a volume fraction in each of the elementary cell and obtain a file of one hundred elementary volume fractions indexed from 1 to 100. The previous ones will be the base of next computations.

6.2 Results

In the following picture we can see the idea presented in the previous section.

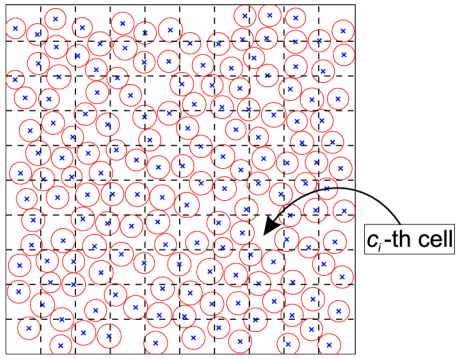


Figure 6.1: *A sample with an abstract grid.*

As was said in the previous chapter, we simulated fifteen samples from each algorithm **AI**–**AIV** and therefore we obtained fifteen sets per one hundred elements. Next step is based on computation mean value for each element from the fifteen samples. This operation leads to getting a data matrix 5x15 which we use for computation descriptive statistics.

Now, let's formulate it more formally. Let i be an index denoting the number of used algorithm. It runs from one to five according to this table:

Algorithm	$\boldsymbol{i}$
Alg. **AI**	1
Alg. **AII**	2
Alg. **AIII**	3
Alg. **AIV**	4
Real sample	5

Next, let index j denotes the number of a sample generated by the given algorithm. In our case it can assume a value from one to fifteen. Finally, index k represents a number of the elementary cells in our sample, see the figure above.

Denote by the symbol $\mathbf{X_i}$, $i = 1, \ldots, 5$ a data object with components $X_i^{j,k}$, $j = 1, \ldots, 15$, $k = 1, \ldots, 100$ and by the symbol $\overline{x}_i^k$ its mean value over all samples, i.e.

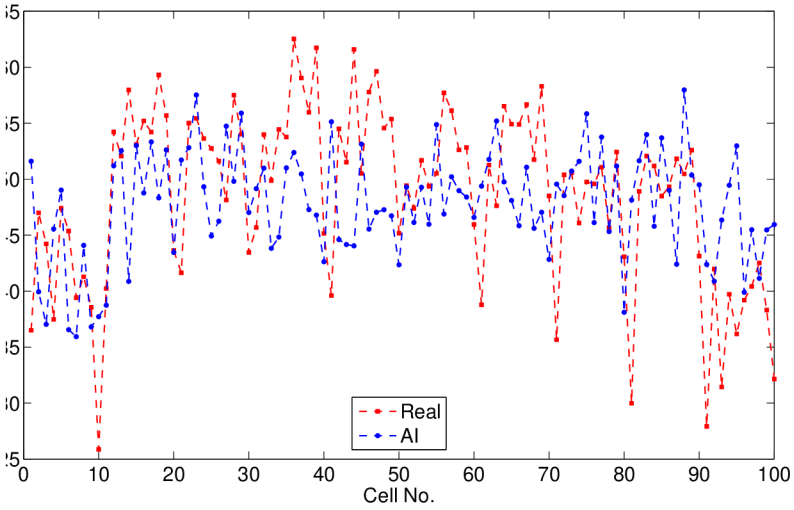
$$\overline{x}_i^k = \frac{1}{15} \sum_{j=1}^{15} X_i^{j,k}.$$

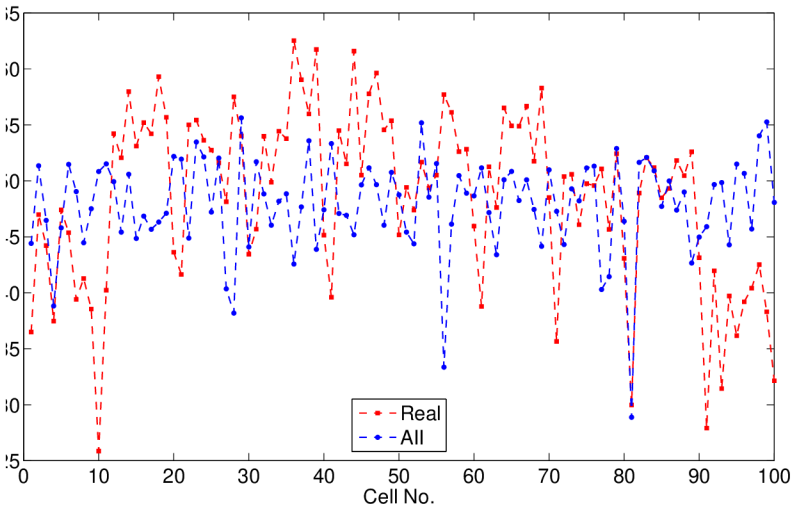
Then, by fixation of the index i in a matrix $\overline{x}_i^k$, we obtain a statistic file of an amount 100, where a classical methods of a descriptive statistic can be applied. For the right formulas for the computation of the characteristics, see e.g. [29] or [33]. Let us denote by a symbol f_i^k the volume fraction in a cell c_i^k (the notation of indexes is the same as above). Then we obtain statistic files for volume fractions in each cell for all realizations overall. The results are presented in the following table:

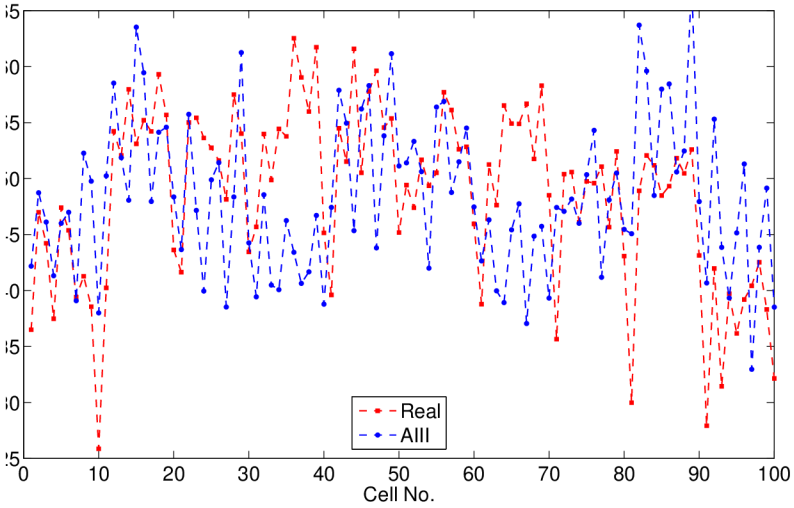
	Mean	**Median**	**Min.**	**Max.**	**Range**	**Var.**	**Std.Dev.**	**Kurt.**	**Skew.**
Real	48,69	50,49	25,86	62,53	36,67	60,16	7,76	3,17	-0,73
AI	47,73	48,23	35,93	57,97	22,04	24,15	4,91	2,77	-0,36
AII	47,87	48,23	28,88	55,63	26,75	18,46	4,3	6,6	-1,32
AIII	48,34	47,94	32,96	67,49	34,53	46,57	6,82	2,78	0,38
AIV	48,44	47,78	34,45	64,73	30,28	53,48	7,31	2,44	0,15

Table 6.1: *Computed values of descriptive statistics of all volume fractions for all samples.*

The following figures display the difference between volume fraction in each elementary cell in real samples (meaning their mean) and samples obtained by each algorithm. Each algorithm **AI**–**AIV** is from the reason of clearness presented in a separated figure:


Real
AI
0
10
20
30
40
50
60
70
80
90
100
Cell No.


Real
AII
0
10
20
30
40
50
60
70
80
90
100
Cell No.


Real
AIII
0
10
20
30
40
50
60
70
80
90
100
Cell No.

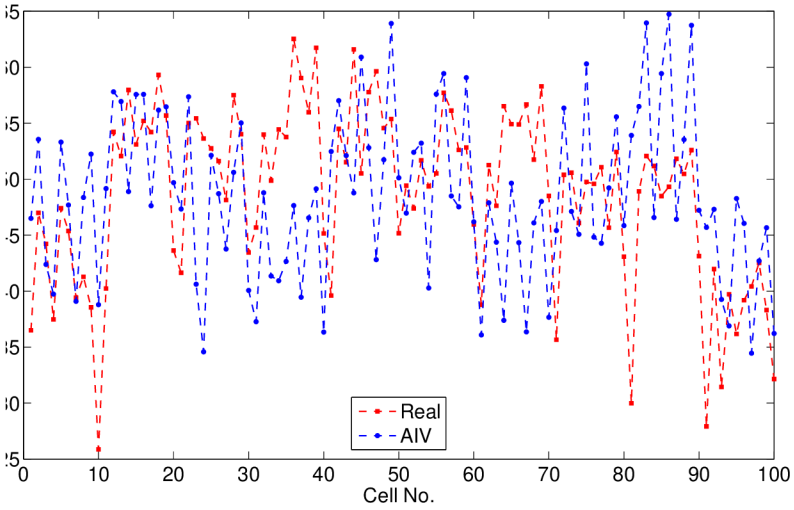


Figure 6.2: *Comparing elementary volume fractions of each algorithm to the real one.*

Next, we present descriptive statistics for the amount of fibres in the samples for each algorithm **AI**–**AIV** and real samples, see table:

	Mean	**Median**	**Min.**	**Max.**	**Range**	**Variation**	**Std. Dev.**	**Kurt.**	**Skew.**
Real	164,60	164	145	189	44	167,40	12,94	2,04	0,14
AI	164,93	163	155	177	22	45,21	6,72	1,89	0,21
AII	162,13	162	159	165	6	2,84	1,68	2,23	-0,22
AIII	169,00	169	169	169	0	0,00	0,00	undef.	undef.
AIV	169,00	169	169	169	0	0,00	0,00	undef.	undef.

Table 6.2: *Computed values of descriptive statistics for a total amount of fibres for several simulations computed by algorithms **AI**–**AIV**.*

We can see, that the amount of fibres for real samples and samples generated by algorithms **AI**–**AII** can vary, but for the algorithms **AIII**–**AIV** is still constant. It is caused by the fact, that we generate random structures from the starting position, where the amount of fibres is chose in such way to be the resulting volume fraction the same, see Figures 4.6 and 4.7. It is a difference in contrast to the algorithms **AI** or **AII**, where we do not know a priori the numbers of fibres that will be generated to a desired sample domain.

7 Anizotropy

7.1 Variograms

Consider two data sets; we will assume that common descriptive statistics for these two data sets are almost the same. According to this evidence the two data sets are almost identical. However, these two data sets are significantly different in ways that are not captured by the common descriptive statistics and histograms. Note that we can not say that data set A is "more variable" than data set B, since the standard deviations for the two data sets are the same. The variogram is a quantitative descriptive statistic that can be graphically represented in a manner which characterizes the spatial continuity of a data set.

In this section we present results concerning with variograms. A theoretical fundamental is introduced in section 1.2.4. In our variogram analysis we computed directional variograms in directions

$$0°;\ 22,5°;\ 45°;\ 67,5°;\ 90°;\ 112,5°;\ 135°;\ 157,5°$$

with angle toleration $11,25°$ and one omnidirectional variogram. Their graphical representation we can see in the next page. To the analysis we used **gstat 2.3.3** software. By help of implemented optimization methods we found out that the best fitting model is *cosine model* with nonzero nugget, see figure 11.8 in Appendix.

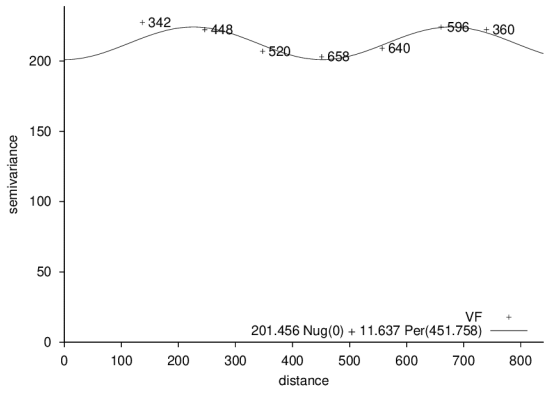


Figure 7.1: *Omni-directional variogram for one real sample.*

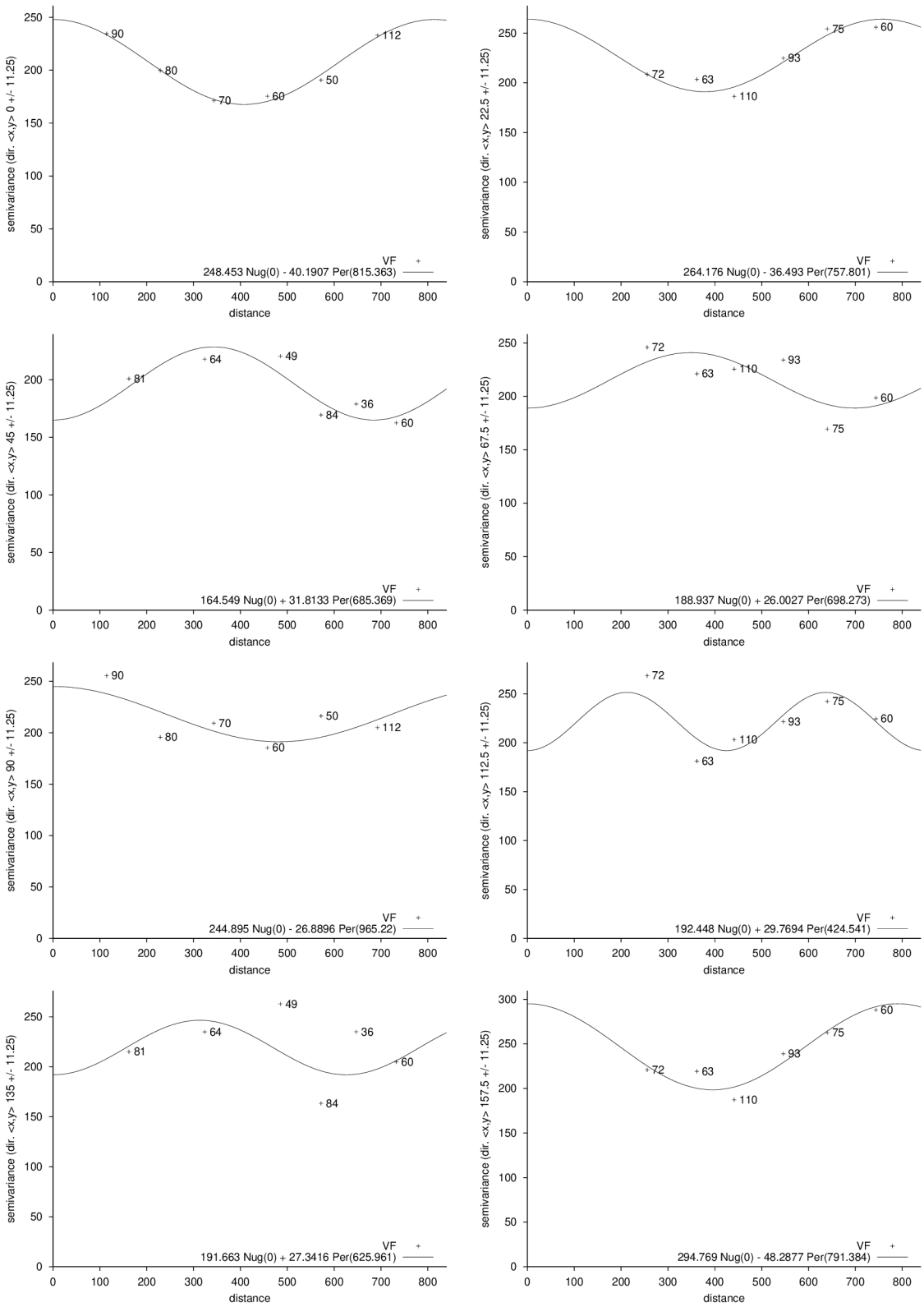


Figure 7.2: *Directional variograms for one real sample.*

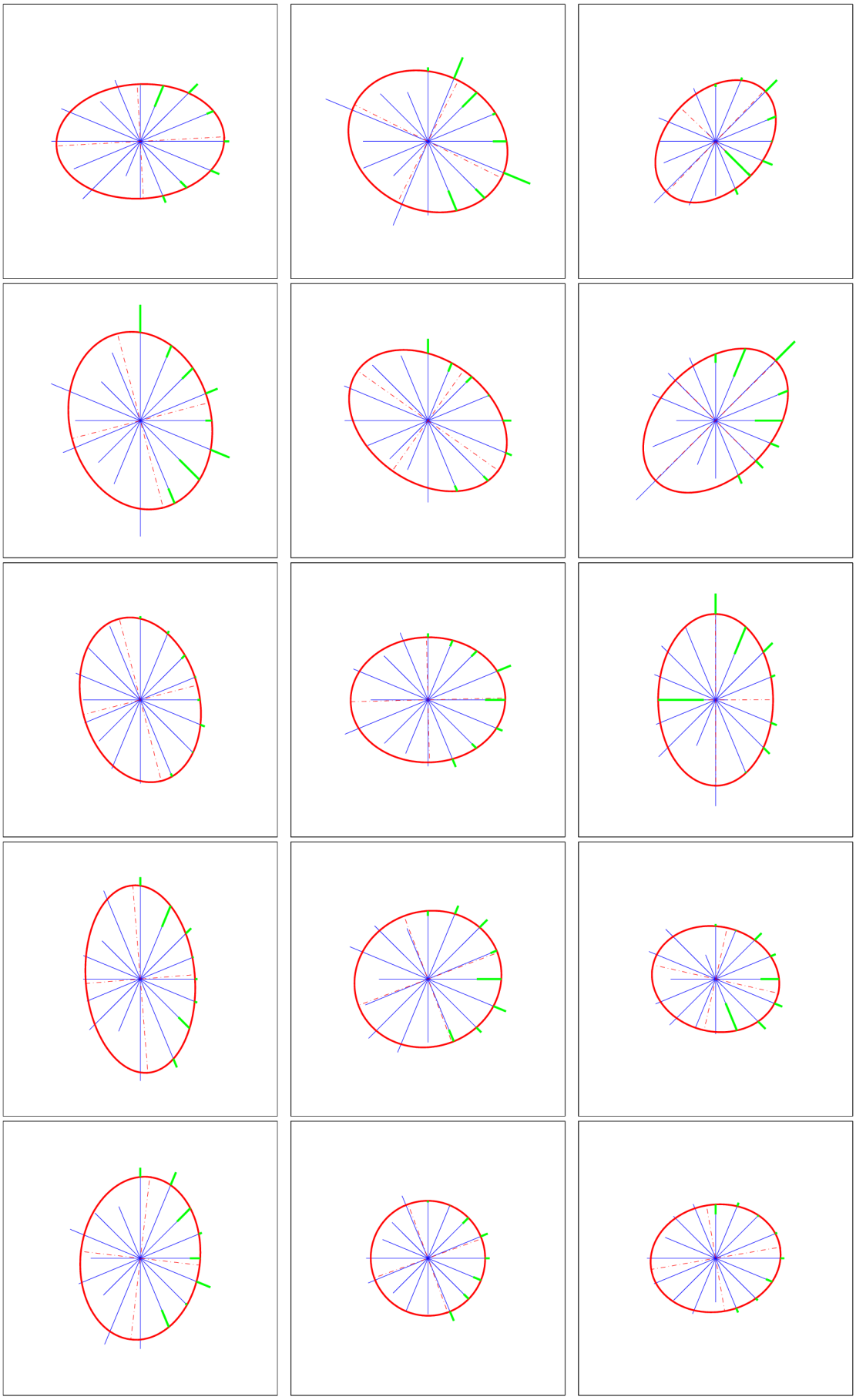


Figure 7.3: *Rose diagrams for samples generated by algorithm* ***AI****.*

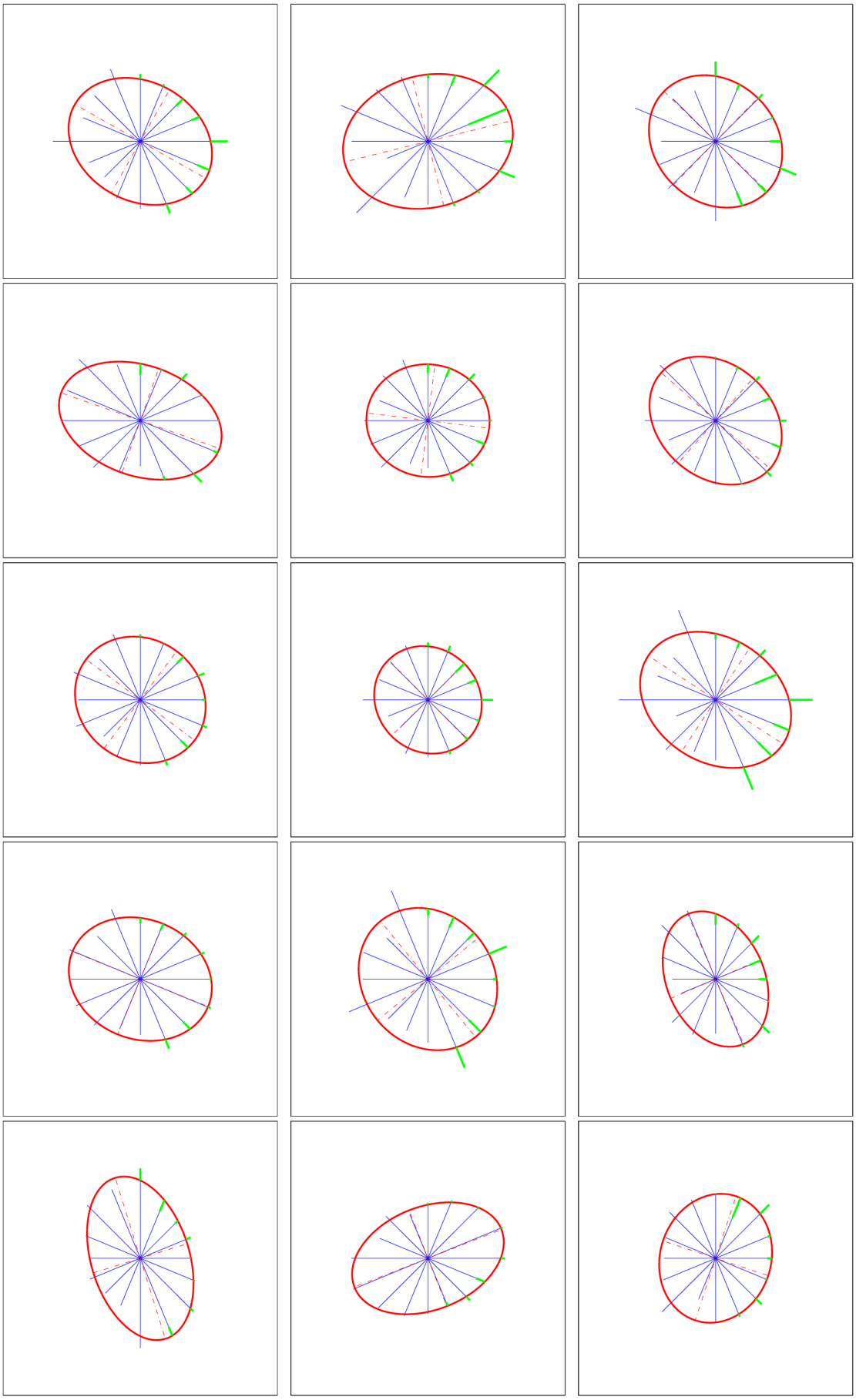


Figure 7.4: *Rose diagrams for samples generated by algorithm* ***AII***.

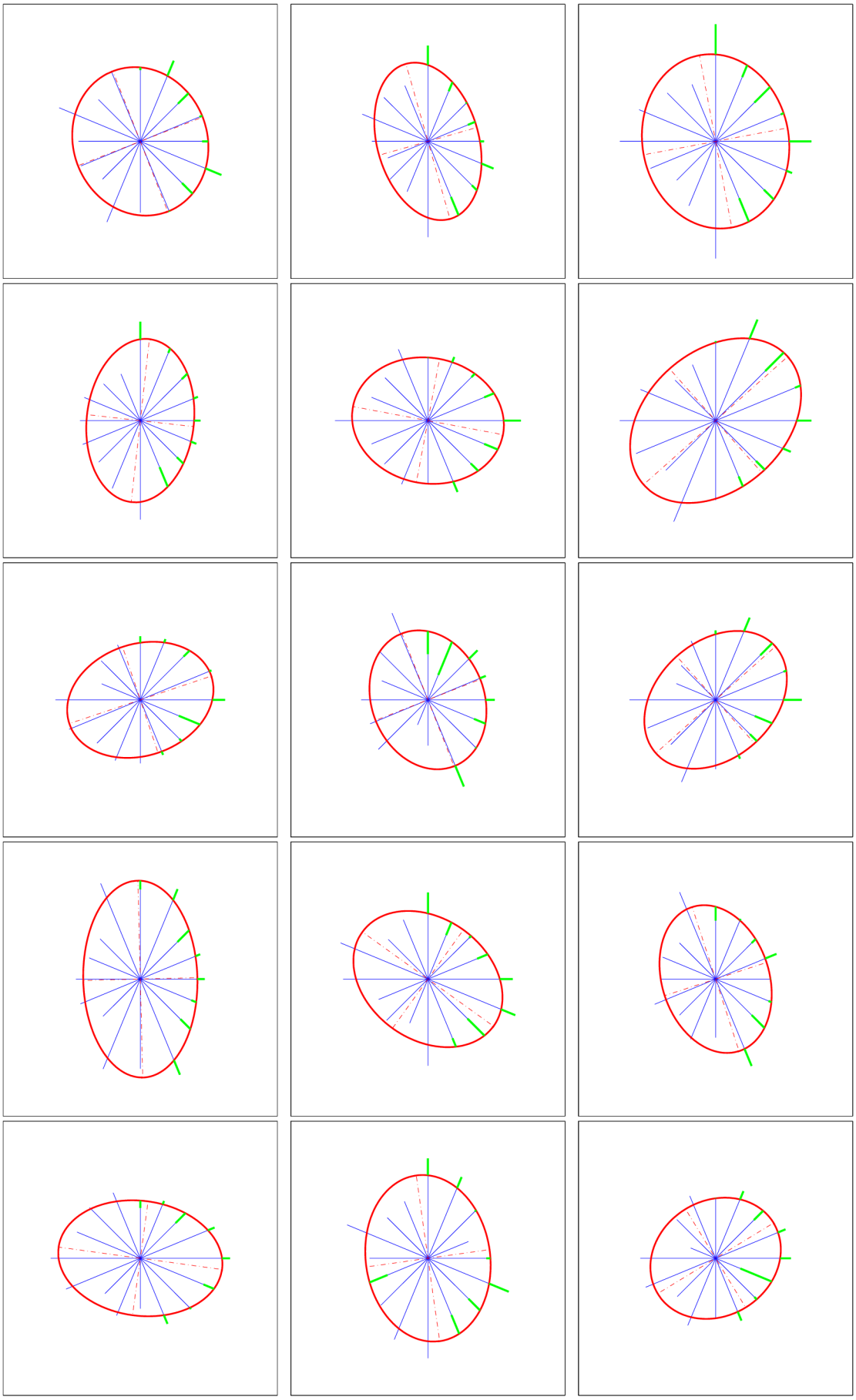


Figure 7.5: *Rose diagrams for samples generated by algorithm* ***AIII***.

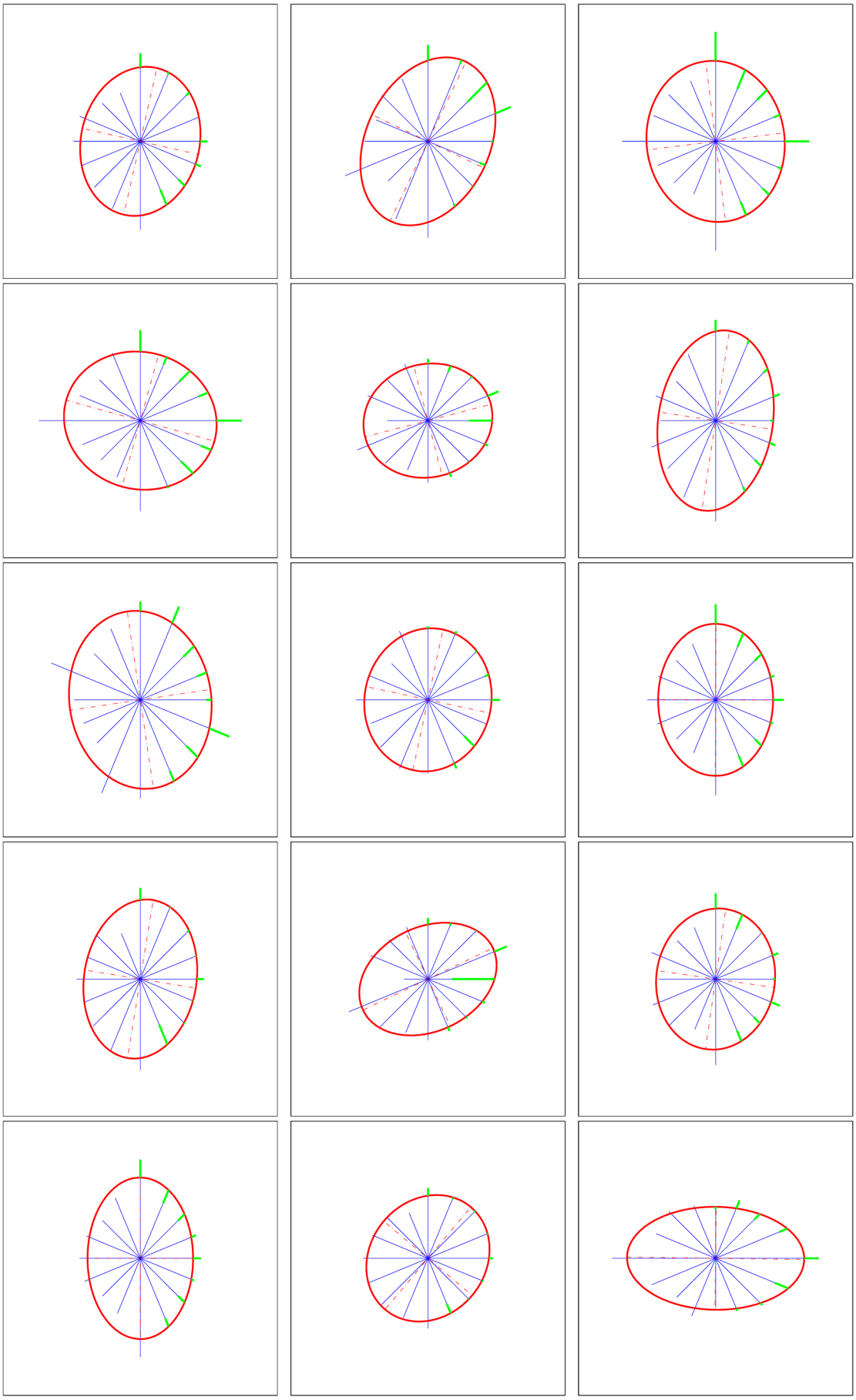


Figure 7.6: *Rose diagrams for samples generated by algorithm* ***AIV***.

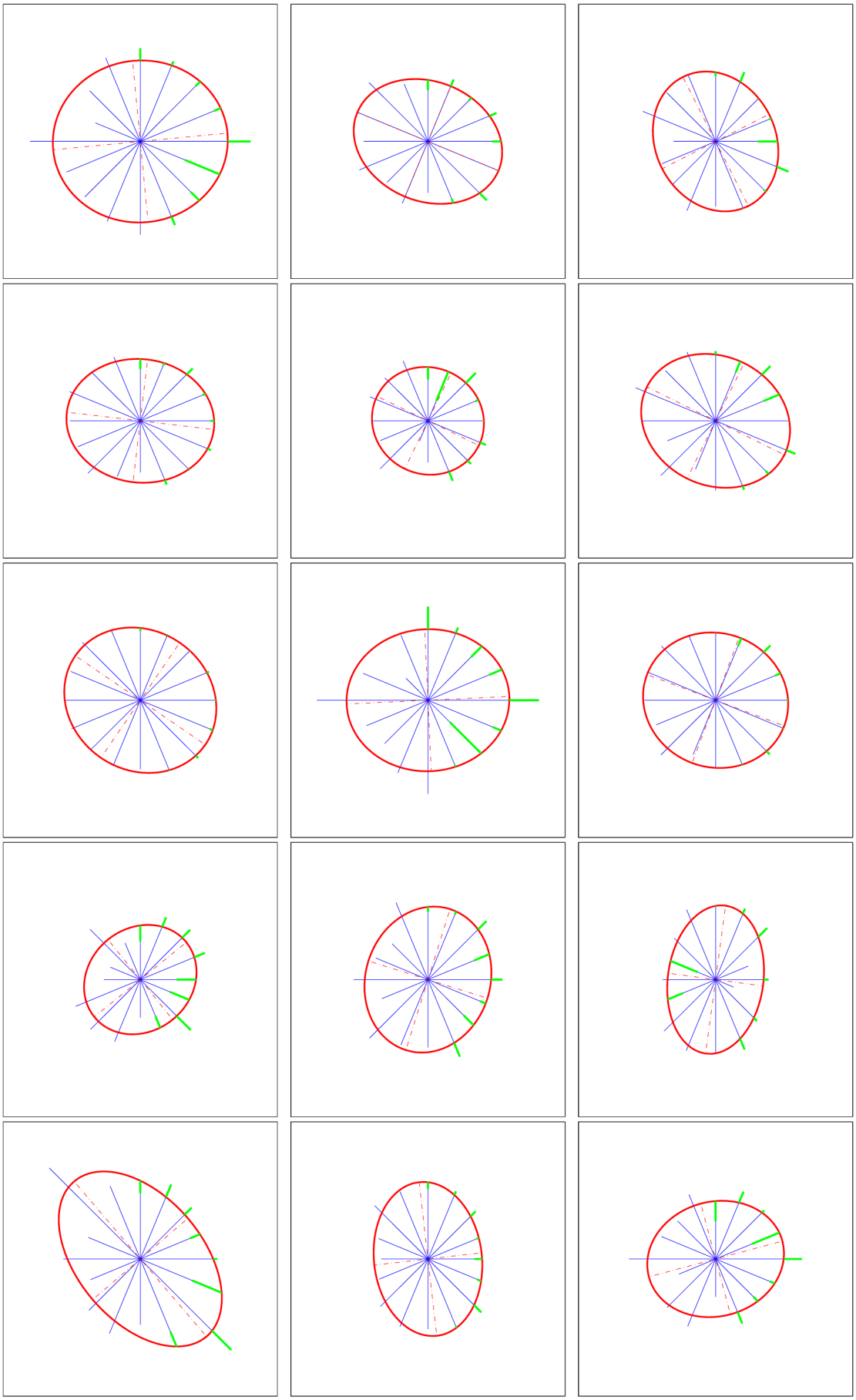


Figure 7.7: *Rose diagrams for real samples.*

7.2 Coefficients

As we said in the section 1.2.5, related to an anisotropy of a material, we distinguish geometric and zonal anisotropy, see figures in mentioned section. The most used parameter to describe geometric anisotropy is so called *anisotropic ratio* k, defined as

$$k = \frac{\text{range of min. variational axis}}{\text{range of max. variational axis}} = \frac{a_1}{a_2} \geq 1.$$

For an isotropic material, $k = 1$, i.e. an ellipse becomes to a circle with $a_1 = a_2$. In the following table we can see the values of anisotropic ratio for the real sample and for algorithms **AI-AIV**evaluated for each of fifteen realizations.

Sample No.	**Real**	**Alg. AI**	**Alg. AII**	**Alg AIII**	**Alg. AIV**
1	1,082	1,467	1,266	1,127	1,267
2	1,278	1,207	1,299	1,587	1,424
3	1,206	1,437	1,187	1,198	1,171
4	1,199	1,275	1,520	1,529	1,122
5	1,062	1,400	1,096	1,215	1,148
6	1,191	1,492	1,280	1,390	1,586
7	1,129	1,425	1,115	1,341	1,261
8	1,147	1,235	1,093	1,268	1,136
9	1,093	1,494	1,298	1,373	1,322
10	1,152	1,722	1,225	1,725	1,424
11	1,184	1,102	1,150	1,332	1,381
12	1,561	1,222	1,447	1,415	1,194
13	1,629	1,368	1,716	1,435	1,530
14	1,432	1,008	1,532	1,345	1,157
15	1,207	1,219	1,179	1,177	1,720

Table 7.1: *Computed values of anisotropic ratios for each algorithm.*

Next, we present summary descriptive statistics for the previous values.

	Mean	**Median**	**Min.**	**Max.**	**Range**	**Variation**	**Std.Dev.**	**Kurt.**	**Skew.**
Real	1,237	1,191	1,062	1,629	0,567	0,029	0,171	3,440	1,287
AI	1,338	1,368	1,008	1,722	0,714	0,033	0,181	2,795	0,149
AII	1,294	1,266	1,093	1,716	0,623	0,034	0,184	2,905	0,915
AIII	1,364	1,345	1,127	1,725	0,598	0,026	0,162	2,879	0,601
AIV	1,323	1,267	1,122	1,720	0,598	0,034	0,184	2,479	0,728

Table 7.2: *Computed descriptive characteristics for anizotropic ratios.*

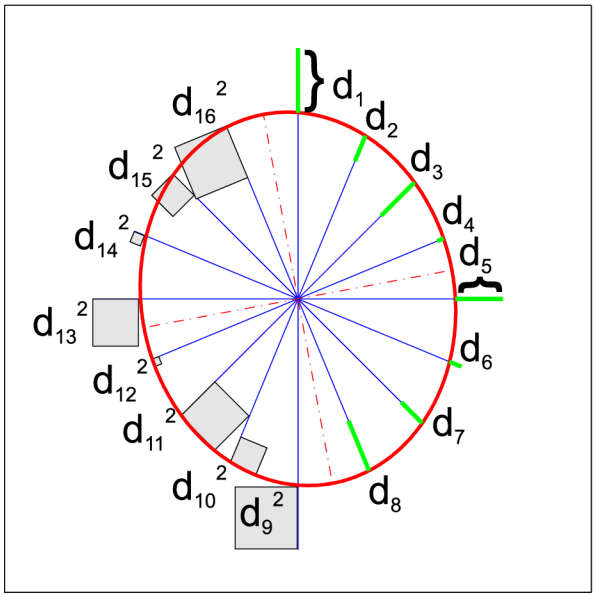


Figure 7.8: *Squares deviations.*

Now, we present a new characteristic-so called *proportional coefficient*, which we define as a ratio of area of the ellipse and a sum of squares of deviations variogram's ranges from an ellipse in estimated directions

$$p = \frac{P}{\sum_{i=1}^{16} d_i^2}.$$

Because of the symmetry of an ellipse, it holds $d_{i+8} = d_i$, so we can simplify the computation to the form

$$p = \frac{\pi ab}{2\sum_{i=1}^{8} d_i^2}.$$

This coefficient practically determines the accuracy of fitting an ellipse of the rose diagram to the separate abscissae obtained from directional variograms. The bigger the coefficient is, the better fitting we have. In an ideal case, the denominator equals to zero and the coefficient tends to infinity.

Sample No.	Real	Alg. AI	Alg. AII	Alg AIII	Alg. AIV
1	4,633	7,689	8,238	7,574	10,668
2	16,677	3,363	3,328	5,608	6,438
3	9,912	3,076	6,392	3,444	3,663
4	30,474	3,516	20,044	6,811	4,349
5	3,123	15,159	13,388	8,293	7,334
6	10,829	2,634	23,661	6,227	21,840
7	225,005	99,871	21,630	7,857	6,183
8	2,350	8,511	9,971	2,313	21,578
9	36,413	2,123	3,110	5,579	6,531
10	2,448	8,177	22,844	8,619	10,816
11	7,177	6,280	5,400	4,221	2,829
12	3,747	3,483	9,648	6,087	8,130
13	3,893	5,834	15,244	12,565	7,737
14	22,125	14,422	39,081	4,222	32,633
15	3,175	27,009	8,068	3,677	11,194

Table 7.3: *Computed values of proportional coefficients for each algorithm.*

	Mean	Median	Min.	Max.	Range	Variation	Std.Dev.	Kurt.	Skew.
Real	25,465	7,177	2,350	225,00	222,655	3162,221	56,234	12,148	3,267
AI	14,077	6,280	2,123	99,871	97,747	606,985	24,637	11,281	3,084
AII	14,003	9,971	3,110	39,081	35,971	97,513	9,875	3,653	1,067
AIII	6,206	6,087	2,313	12,565	10,252	6,672	2,583	3,540	0,751
AIV	10,795	7,737	2,829	32,633	29,804	68,432	8,272	4,340	1,506

Table 7.4: *Computed descriptive characteristics of proportional coefficients.*

8 Assumptions for the Analysis

Among the most important properties in statistical analysis are its *normality* and *homogeneity of variance*. In the following we try find out, whether sets of elementary volume fractions in the samples satisfy these conditions.

8.1 Normality

If the number of members in each group is fairly large, then deviations from normality do not matter much at all because of the central limit theorem. In our case, every sample has 100 data. In our analysis of normality we choose eight different tests to clarify this phenomena. The tests are: Anderson-Darling test, Chi-squared test, D'Agostino's K-squared test, Jarque-Bera test, Kolmogorov-Smirnov test, Lilliefors test, Ryan-Joiner test and Shapiro-Wilk test. Each of this tests is described in Appendix.

Tst.	Anderson-Darling					Chi-Squared					D'Agostino's K-squared					Jarque-Bera				
No.	AI	AII	AIII	AIV	Real	AI	AII	AIII	AIV	Real	AI	AII	AIII	AIV	Real	AI	AII	AIII	AIV	Real
1	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	X	✓	✓	✓	✓	X	✓	✓	✓	X	✓	✓	✓	✓	X	✓
3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	X
7	✓	X	✓	✓	X	✓	X	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
8	✓	✓	✓	X	X	✓	✓	✓	X	X	✓	✓	✓	X	✓	✓	✓	✓	X	✓
9	✓	✓	✓	X	✓	✓	✓	X	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
11	X	✓	✓	✓	X	X	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
12	✓	X	✓	✓	X	✓	X	✓	✓	X	X	✓	✓	✓	X	✓	✓	✓	✓	X
13	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	X	✓
14	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
15	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Tst.	Kolmogorov-Smirnov					Lilliefors					Ryan-Joiner					Shapiro-Wilk				
No.	AI	AII	AIII	AIV	Real	AI	AII	AIII	AIV	Real	AI	AII	AIII	AIV	Real	AI	AII	AIII	AIV	Real
1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	X	X	✓	✓	✓	X	✓
3	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓
5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓
7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	X	✓	X	✓	✓	X
8	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	X	✓	✓	✓	✓	X	✓
9	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	X	✓
10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
11	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	X	✓	✓	✓	✓	X	✓	✓	✓	✓
12	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	X
13	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	X	✓	✓	X	✓	✓	✓	✓	✓	✓
14	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
15	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 8.1: *Resulting values obtained by various tests for verification of normality.*

As we can see, almost every data fulfils the normality, so we can say, that the values of elementary volume fractions satisfy to the normality condition.

8.2 Homogeneity of Variances

This section will be denoted to the second important request – homogeneity of variances. This is very important for an ANOVA(Analysis of Variance), due to F-tests, which this method is based on. As in the previous, we will study the homogeneity on the set of elementary volume fractions for each sample. We will not compute the homogeneity for all five(real + 4 five) algorithms together, but always for real sample and one for some algorithm. To realize this, we need a pair of samples. We have four algorithms **AI**–**AIV** and real one. So, we will have four sets of computing. We can use parametric tests, because we know, that our data are normally distributed. The tests we use: Bartlett's test, Cochran test, Brown-Forsythe test, Levene test and O'Brien test. They are also described in Appendix. The results are in the table:

Test	**Bartlett's**				**Cochran**				**Brown-Forsythe**				**Levene**				**O'Brien**			
Alg.	**AI**	**AII**	**AIII**	**AIV**	**AI**	**AII**	**AIII**	**AIV**	**AI**	**AII**	**AIII**	**AIV**	**AI**	**AII**	**AIII**	**AIV**	**AI**	**AII**	**AIII**	**AIV**
1	✓	✓	X	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓
2	✓	X	X	X	✓	X	X	X	✓	X	X	X	✓	X	X	X	✓	X	X	X
3	X	✓	X	X	X	✓	X	X	X	✓	X	X	X	✓	X	X	X	✓	X	X
4	✓	X	X	X	✓	X	X	X	✓	X	✓	X	✓	X	X	X	✓	X	X	X
5	✓	X	X	✓	✓	X	X	X	✓	X	X	✓	✓	X	X	✓	✓	X	X	✓
6	✓	✓	X	X	✓	✓	X	X	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	X	✓
7	✓	X	X	X	✓	X	X	X	✓	X	X	X	✓	X	X	X	✓	X	X	X
8	✓	✓	X	✓	✓	✓	X	✓	✓	✓	X	✓	✓	✓	X	✓	✓	✓	X	✓
9	✓	✓	X	✓	✓	X	X	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	X	✓
10	✓	X	X	X	✓	X	X	X	✓	X	X	X	✓	X	X	X	✓	X	X	X
11	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
12	✓	X	X	X	X	X	X	X	✓	X	X	X	✓	X	X	X	✓	X	X	X
13	✓	✓	X	X	✓	✓	X	X	✓	✓	X	X	✓	✓	X	X	✓	✓	X	X
14	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
15	✓	✓	X	X	✓	X	X	X	✓	✓	X	X	✓	✓	X	X	✓	✓	X	X

Table 8.2: *Resulting values obtained by various tests for verification of homogeneity.*

Explanation of this table: in the header of the table, e.g. **AI** means that we compare samples from real material and a sample generated by algorithm **AI** by the test that is above. The checkmark(✓) means that its variances are homogeneous, meaning it is not rejected, a cross(X) means an opposite.

Here, according to this tests we can see, that these data do not have a character to be homogeneous in variance, generally. The "best" is **AI** and the "worst" one is **AIII**. So, for eventual analysis, we cannot use ANOVA, because the assumptions are not fulfilled. The only way is to use some nonparametric test, e.g. the *two-sample Kolmogorov-Smirnov test.*

8.2.1 Two-Sample Kolmogorov-Smirnov test

In statistics, the Kolmogorov–Smirnov test(K–S test) is a form of minimum distance estimation used as a nonparametric test of equality of one-dimensional probability distributions used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test).

The two-sample Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution. In each case, the distributions considered under the null hypothesis are continuous distributions.

The two-sample KS test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

The aim of this computation is to check out, whether two samples(real and simulated) come from the same distribution. We will by sequel apply the two-sample K-S test to the pair of samples combined from the real sample and simulated one from the algorithms **AI**–**AIV**. To the computation we use a Matlab function `kstest2`, see a guide book [35] about syntax, input and output arguments.

	Real - AI		**Real - AII**		**Real - AIII**		**Real - AIV**	
Sample No.	**Result**	**p-Value**	**Result**	**p-Value**	**Result**	**p-Value**	**Result**	**p-Value**
1	✓	0,140	X	0,013	✓	0,193	X	0,021
2	✓	0,193	X	0,005	X	0,001	X	0,003
3	✓	0,443	✓	0,677	✓	0,140	✓	0,261
4	✓	0,140	✓	0,193	✓	0,193	✓	0,344
5	X	0,021	X	0,013	X	0,031	X	0,031
6	X	0,003	✓	0,099	X	0,047	✓	0,140
7	✓	0,443	X	0,013	✓	0,099	✓	0,069
8	X	0,008	✓	0,069	✓	0,344	X	0,021
9	✓	0,443	✓	0,193	✓	0,140	✓	0,261
10	✓	0,140	X	0,021	X	0,008	X	0,002
11	X	0,000	✓	0,069	X	0,008	✓	0,069
12	✓	0,099	X	0,013	X	0,003	X	0,031
13	✓	0,261	✓	0,099	X	0,013	X	0,005
14	✓	0,894	✓	0,677	✓	0,344	✓	0,261
15	✓	0,556	✓	0,794	✓	0,677	✓	0,677

Table 8.3: *Resulting values of the two-sample Kolmogorov-Smirnov test.*

In the table we can see whether the H_0 hypothesis is rejected(X) or not(✓). Beside these markers it is also present the $p-$Value for each test. It is clear, that if the $p-$Value is greater then the significance level $\alpha = 0,05$, then we do not reject the null hypothesis(the samples come from the same distribution).

Now, for a consideration of the two-sample K-S test we will illustrate the mutual connection or similarity between algorithms **AI**–**AIV**.

	AI - AII		AI - AIII		AI - AIV	
Sample No.	**Result**	**p-Value**	**Result**	**p-Value**	**Result**	**p-Value**
1	✓	0,794	✓	0,677	✓	0,344
2	✓	0,261	✓	0,140	✓	0,261
3	✓	0,140	✓	0,443	✓	0,443
4	✓	0,140	✓	0,193	✓	0,193
5	✓	0,894	✓	0,794	✓	0,961
6	✓	0,140	✓	0,344	✓	0,344
7	✓	0,069	✓	0,677	✓	0,556
8	✓	0,794	✓	0,261	✓	0,961
9	✓	0,556	✓	0,344	✓	0,794
10	✓	0,344	✓	0,261	✓	0,140
11	✓	0,099	✓	0,099	✓	0,140
12	✓	0,794	✓	0,443	✓	0,894
13	✓	0,443	✓	0,261	✓	0,193
14	✓	0,140	✓	0,261	✓	0,140
15	✓	0,261	✓	0,140	✓	0,261

Table 8.4: *Two-sample Kolmogorov-Smirnov test for the algorithm **AI**.*

	AII - AIII		AII - AIV		AIII - AIV	
Sample No.	**Result**	**p-Value**	**Result**	**p-Value**	**Result**	**p-Value**
1	✓	0,443	✓	0,261	✓	0,894
2	✓	0,961	✓	0,794	✓	0,556
3	X	0,031	✓	0,443	✓	0,193
4	✓	0,443	✓	0,677	✓	0,961
5	✓	0,261	✓	0,677	✓	0,992
6	✓	0,794	✓	0,894	✓	0,961
7	✓	0,344	✓	0,261	✓	0,961
8	✓	0,261	✓	0,443	✓	0,443
9	✓	0,992	✓	0,556	✓	0,443
10	✓	0,894	✓	0,794	✓	0,556
11	✓	0,894	✓	0,894	✓	0,992
12	✓	0,099	✓	0,677	✓	0,443
13	✓	0,140	✓	0,344	✓	0,894
14	✓	0,961	✓	0,992	✓	0,677
15	✓	0,677	✓	0,961	✓	0,794

Table 8.5: *Two-sample Kolmogorov-Smirnov test for the algorithms **AII** and **AIII**.*

From the upper table it seems there is no significant difference between **AI** and the remaining ones. Almost the same we can say in the case of the algorithm **AII**, because the $p-$Values vary approximately from 0,1 to 0,95. A slight difference we can find in the last case, i.e. between **AIII** and **AIV**. Here, when looking at the $p-$Values we can say, that its variance is much more smaller – almost all values are greater then 0,6. It means, that there is quite no difference between **AIII** and **AIV**from the statistical point of view. This fact was slightly indicated in the passage about describing algorithms. So, statistically

we can change these ones. The only difference is in the fact, that the algorithm **AIII** is approximately four times faster then **AIV**.

8.3 Complete Spatial Randomness

Now, we examine tests for the CSR hypothesis of a point pattern in our samples. This hypothesis states that the observed pattern was generated by a homogeneous Poisson process. According to [9], CSR operates as a dividing hypothesis between aggregated and regular patterns and its rejection is a minimum requirement for further modeling.

8.3.1 The Quadrat Test of Randomness

It is the simplest and the most widely used method to investigate deviations from randomness and it is based on counting the numbers of points(centers) in each quadrat of a grid overlaid on the section of interest. The approach used to calculate the quadrat test involves analyzing the variation in the numbers of points in selected sub-areas of the region under investigation. This is called the quadrat method, see 2.3.1. The comparison will be as follows: For each sample we compute Pearson's test statistic Q and compare it with the critical value. In our cases we choose $n = 10$, i.e. the 10x10 grid (an assumption $n^2 > 6$ should be fulfilled). The results are in the following table. It holds, under CSR,

	Real	**AI**	**AII**	**AIII**	**AIV**
1	35,45	33,04	37,75	33,31	34,59
2	34,69	41,90	30,77	37,28	33,11
3	34,20	30,65	33,30	31,90	37,74
4	39,97	40,75	36,19	24,06	25,27
5	26,70	34,23	26,93	25,73	35,67
6	26,96	47,56	45,29	26,16	30,84
7	35,53	46,29	37,99	25,58	31,28
8	32,38	29,12	35,08	31,76	28,17
9	36,89	36,10	30,85	29,38	33,98
10	38,95	48,47	33,97	29,55	22,21
11	23,54	34,78	32,84	29,20	37,67
12	34,35	43,46	30,03	31,85	34,54
13	36,75	37,67	42,84	25,38	34,20
14	30,39	38,52	31,54	25,84	41,82
15	39,09	55,13	31,20	26,42	27,20

Table 8.6: *Pearson's statistics Q for the quadrat test of randomness.*

the Pearson's test statistic has χ^2-distribution with $f = n^2 - 1 = 10^2 - 1 = 99$ degrees of freedom.

If the value for Q is less than the $100\alpha/2$ percentile of the chi-squared distribution with $n^2 - 1$ degrees of freedom, the test rejects the stationary Poisson point process hypothesis in favour of regularity at level α. If it is greater than the $100(1 - \alpha/2)$ percentile, then the same hypothesis is rejected at level α, this time in favour of clustering(meaning that the variability in the process is greater than that for the Poisson process).

According to [32], a constant problem in designing a study using quadrats is to establish what would be a suitable size for the quadrat. Various suggestions have been made as to the optimal size, however, most authors agree that the size of the quadrats depends on the specific problem in hand, like the type and range of the events' interactions with each other.

In our case, $n = 10$, so $\chi^2_{99}(0.025) = 73,36$ and $\chi^2_{99}(0,975) = 128,42$. Since in our case, all values of Pearson's test statistic Q are smaller than 73,36, it indicates significant departure from the CSR.

8.3.2 Tests Based on Ripley's K Function

Now, we present to check non-CSR not directly the Ripley's $K-$function, but the $\hat{D}-$ and $\hat{L}-$functions, which are defined by means of the Ripley's $K-$function. They are defined as, see 2.4.1

$$\hat{D} = \hat{K}(t) - \pi t^2 \qquad \text{and} \qquad \hat{L}(t) = \sqrt{\frac{\hat{K}(t)}{\pi}}.$$

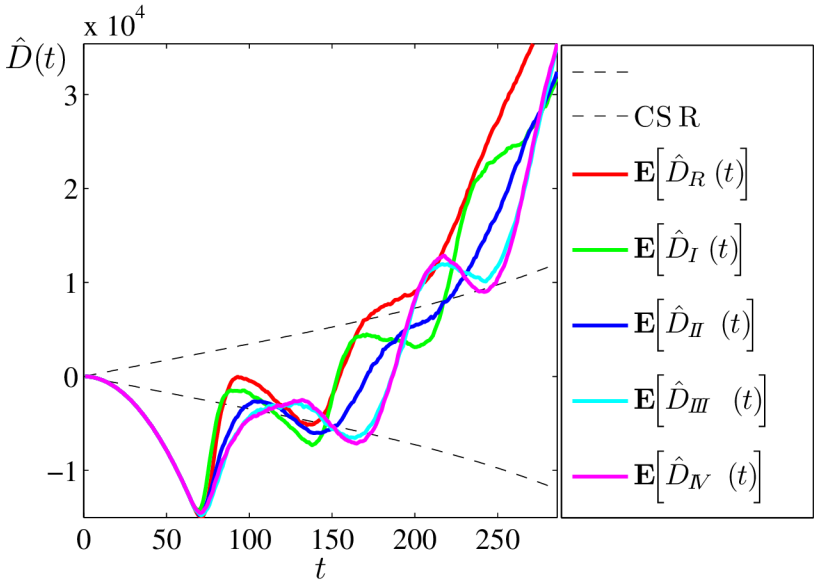


Figure 8.1: *Comparison of $\hat{D}-$functions.*

From the preceding figures it is clear a big departure from the CSR, especially around the beginning. It is caused by the fact, that no two fibres can overlap, so the distance of their centers is greater than the sum of the mutual radii of the fibres. The same situation occurs in the distance approximately 180. The behavior of the $L-$ and $D-$ curves between

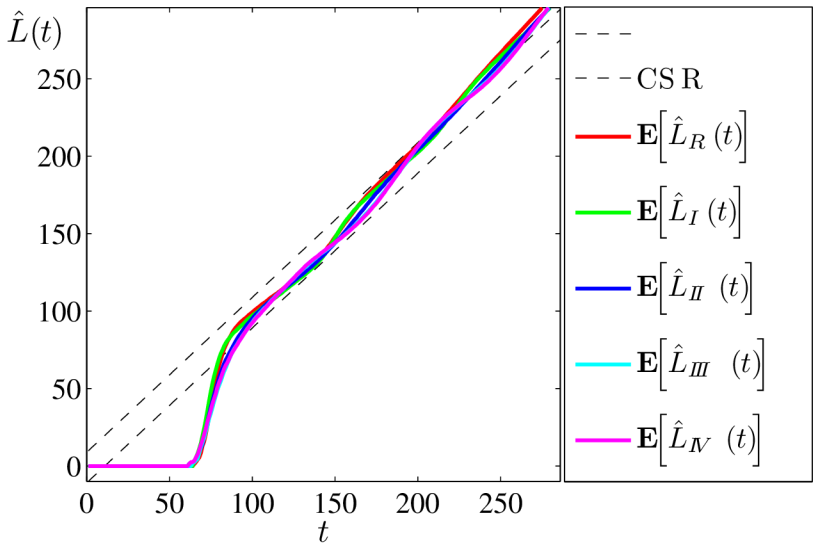


Figure 8.2: *Comparison of $\hat{L}-$functions.*

75–180 indicates the CSR, but, as we can see in figures 2.10–2.12, therefore suggesting some evidence of deviation from randomness towards a regularity.

8.3.3 Clark-Evans Test

The Clark-Evans test is based on the index of the degree of the non-randomness for a spatial configuration. It consists of comparing the observed mean nearest neighbor distance to that expected for a random configuration of the same density.

As was pointed out in the subsection 2.3.3, the results of this Clark-Evans test depend on the particular sample of the nn-distances chosen. If we proceed a Clark-Evens test several times, always with different set of samples, we obtain different Z-values. The results of this simulated sampling scheme yield a distribution of Z-values that is approximately normal. While this normality property is again a consequence of the Central Limit Theorem, it should not be confused with the normal distribution in 2.3.6 upon which the Clark-Evans test is based (that requires n to be sufficiently large). However, this normality property does suggest that a 50% sample($m = n/2$) in this case yields a reasonable amount of independence among nn-distances, as it was intended to do.

On the next figure we can see a realization of Z-values for the real media:

Now, we present the Z-means of all samples obtained by simulating algorithms **AI**–**AIV** and of the real ones.

And here we can see the extremes from the previous table for a better clarify. If we choose a significant level of 0,05, the critical value $z_{\alpha/2} = z_{0,025} = 1,96$ and thus we reject the hypothesis of CSR. Since $z_\alpha = z_{0,05} = 1,65$, we conclude significant uniformity of the patterns.

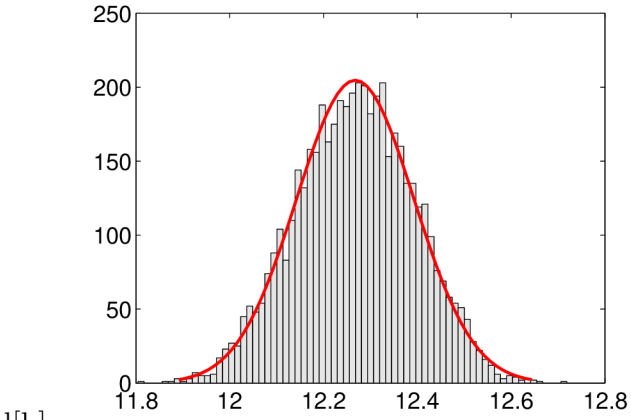


Figure 8.3: *Histogram of the Z-means for the real material.*

	Real	AI	AII	AIII	AIV
1	12,457	12,367	12,472	13,612	13,029
2	12,827	11,251	12,383	13,975	13,716
3	11,342	12,938	12,864	13,651	13,540
4	10,586	12,476	12,849	13,606	13,334
5	13,101	12,218	12,796	13,251	13,366
6	12,625	11,193	12,075	13,740	13,211
7	10,316	12,422	12,298	13,287	13,523
8	14,062	10,979	12,484	13,675	13,201
9	13,207	11,749	12,279	13,890	13,512
10	10,401	11,728	12,299	13,538	13,053
11	13,456	11,149	12,926	13,739	13,481
12	12,052	12,114	12,493	13,272	13,504
13	14,067	12,544	12,287	13,810	13,133
14	11,398	11,609	12,293	13,463	13,177
15	12,270	11,160	12,490	13,230	12,800

Table 8.7: *The values of the mean values obtained by Monte-Carlo simulation of the Clark-Evans test.*

	Minimum	Maximum
Real	10,316	14,067
AI	10,979	12,938
AII	12,075	12,926
AIII	13,230	13,975
AIV	12,800	13,716

Table 8.8: *Extremes of the mean values obtained by Monte-Carlo simulation of the Clark-Evans test.*

8.3.4 Skellam statistic

As a second example of distance methods we present Skellam statistic. A theory, concerning to this, are in the Appendix. The results obtained by this method are below: And

	Real	**AI**	**AII**	**AIII**	**AIV**
1	445,96	453,67	452,01	481,20	479,57
2	468,99	400,88	441,36	504,86	488,83
3	390,72	482,46	461,85	486,18	494,10
4	362,57	453,55	467,44	485,92	484,93
5	481,53	446,36	457,60	484,58	475,39
6	459,40	406,96	426,79	498,39	475,45
7	364,61	463,54	443,61	478,93	487,03
8	541,02	390,66	444,18	485,95	477,93
9	489,81	425,82	442,38	494,32	481,45
10	366,78	434,08	452,05	491,20	478,54
11	496,99	406,91	460,92	492,22	477,35
12	427,88	445,93	441,40	469,35	484,12
13	523,05	467,34	438,24	497,28	477,37
14	404,41	420,55	456,61	483,42	483,16
15	435,36	402,12	452,35	478,79	459,57
Range	**178,45**	**91,79**	**40,65**	**35,52**	**34,53**

Table 8.9: *The values of the Skellam statistic for all samples.*

the extremes are: Similarly, as in the case of the Clark-Evans test, we have $\alpha = 0,05$ and

	Minimum	**Maximum**	**Range**
Real	362,57	541,02	178,45
AI	390,66	482,46	91,79
AII	426,79	467,44	40,65
AIII	469,35	504,86	35,52
AIV	459,57	494,10	34,53

Table 8.10: *Extremes of the Skellam statistic for all algorithms.*

approximately $n = 165$, see Table 6.2 and then $\chi^2_{2n}(0,025) = \chi^2_{330}(0,025) = 281,6$ and $\chi^2_{330}(0,975) = 382,2$. From the values in the previous table and the value of $\chi^2_{330}(0,975)$, we can deduce rejecting CSR, because the minimum values are greater than the critical value. The only exception is the real sample. When we look at the ranges, we can easy compare, which algorithms are more random than the others. In our case the most random are the real ones and on the other hand, the smallest variability has algorithm **AIV**.

9 Computational Circumstances

The programs for all computations and simulations were written in Matlab R14. The used PC's hardware parameters were CPU 1100MHz and 256MB of RAM.

Part III

Appendix

Thank you for the music...

10 Distance Methods

Among the simplest of these is based on the observation that if one simply looks at distances between points and their nearest neighbors in A, then this provides a natural test statistic that requires no artificial partitioning scheme. More precisely, for any given points, $s = (s_1, s_2)$ and $v = (v_1, v_2)$ in A we denote the *Euclidean distance* between s and v by

$$d(s, v) = \sqrt{(s_1 - v_1)^2 + (s_2 - v_2)^2}$$

and denote each *point pattern* of size n in A by $S_n = (s_i : i = 1, ..., n)$, then for any point $s_i \in S_n$, the *nearest neighbor distance (nn-distance)* from s_i to all other points in S_n is given by

$$d_i = d_i(S_n) = \min\{d(s_i, s_j) : s_j \in S_n,\, j \neq i\}.$$

10.1 Skellam's Statistic

To make ideas of nearest-neighbor distances precise, we have to determine the probability

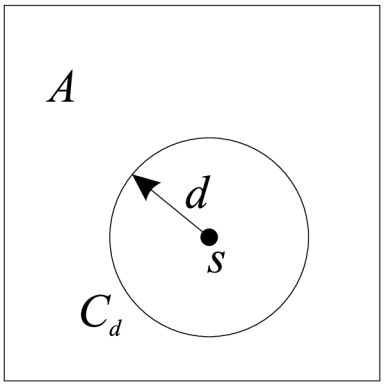


Figure 10.1: *Cell of radius d*

distribution of nn-distance under CSR and compare the observed nn-distance with this distribution. To begin, suppose that the implicit reference region A is large, so that for any given point density λ, we may assume that cell-counts are Poisson distributed under CSR. Now suppose that s is a randomly selected point in a pattern realization of this CSR process, and let the random variable, say D, denote nn-distance from s to the rest of the pattern. To determine the distribution of D, we next consider a circular region C_d of radius d around s, as shown in Figure 10.1. Then, according to the picture, the probability that D is at least equal to d is precisely the probability that here are no other points in C_d. Hence, if we now let $C_d(s) - \{s\}$, then this probability is given by

$$\mathbf{P}(D > d) = \mathrm{e}^{-\lambda\pi d^2} \tag{10.1.1}$$

and that's why we finally obtain

$$F_D(d) = 1 - \mathrm{e}^{-\lambda\pi d^2}. \tag{10.1.2}$$

As we can see, this is an instance of the Rayleigh distribution. Next, for a random sample of n nearest-neighbor distances $\{W_1, \ldots, W_m\}$ from this distribution, the scaled sum (Skellam's statistics)

$$S_w = 2\lambda\pi \sum_{i=1}^{m} W_i^2 \tag{10.1.3}$$

is *chi-square distributed* with $2n$ degrees of freedom. So, finally, this statistic provides a test of the CSR hypothesis based on nearest neighbors.

11 Theoretical Models of Variograms

11.1 Valid Models

The experimental variogram obtained from measured data is in practice impossible to use for the next analysis. So, we have to approximate point-estimated variogram by a theoretic model of the variogram. But the values of the variogram we cannot approximate by an arbitrary function. In other words, a theoretical variogram is not an arbitrary function, but it has to fulfil some conditions (it is similar to a density function of a random variable). The most important condition is, that it must not be negative. It is quite difficult to prove, that the model of a variogram must be conditionally negative definite, see e.g. [8] for detailed information and proves. To check this condition is very difficult, so we always try to use predefined models of variograms, as will be described in the following section.

11.2 Review of the Most Used Models

Models of variograms we can divide according behavior near origin and "infinity" into several groups.

1. *Models with sill* – spherical, quadratic, gaussian, exponential, linear
2. *Models without sill* – power, logarithm
3. *Models with oscillating sill* – sine, cosine
4. *Pure random model*

The first three types of models we can remark a nugget. The models with a sill are weakly stationary, whereas unbounded models are intrinsically stationary.

11.2.1 Models with Sill

Spherical Model

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C_1 \left(\frac{3h}{2a} - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right), & 0 < h \leq a \\ C_0 + C1, & h > a \end{cases}$$

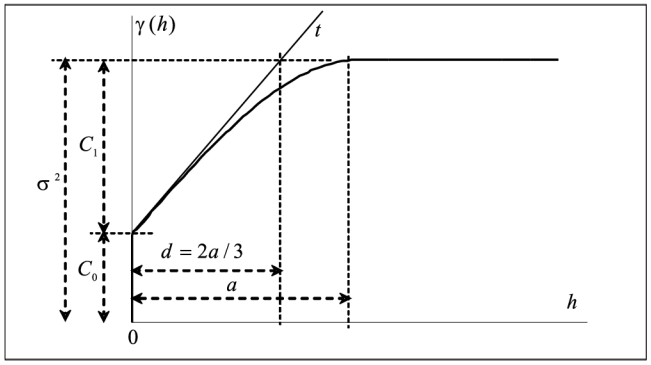


Figure 11.1: *Spherical model*

Quadratic Model

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C_1 \left(2\frac{h}{a} - \left(\frac{h}{a} \right)^2 \right), & 0 < h \leq a \\ C_0 + C1, & h > a \end{cases}$$

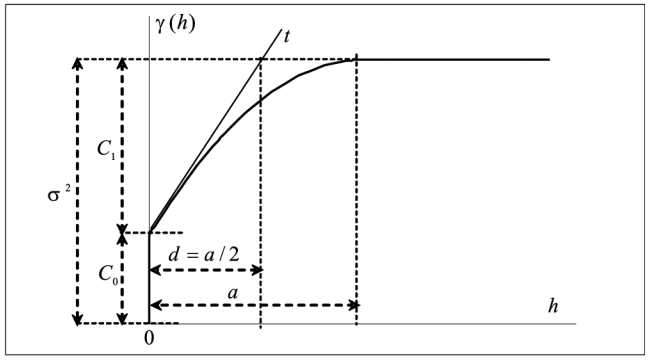


Figure 11.2: *Quadratic model*

Exponential Model

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C1\left(1 - \exp\left(-\frac{h}{d}\right)\right), & h > 0 \end{cases} \qquad a \approx 3d$$

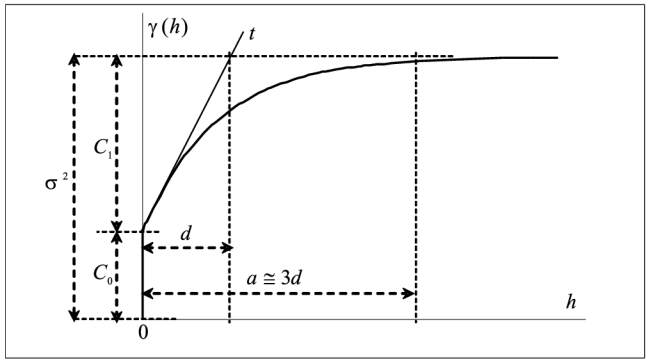


Figure 11.3: *Exponential model*

Gaussian Model

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C_1\left(1 - \exp\left(-\frac{h}{d}\right)^2\right), & h > a \end{cases} \qquad a \approx \sqrt{3}d$$

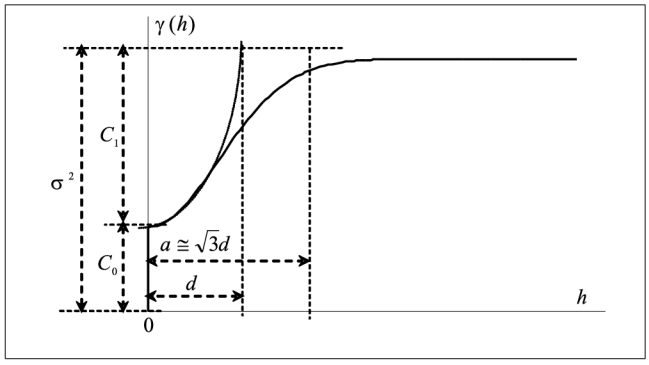


Figure 11.4: *Gaussian model*

Linear Model

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C_1 \frac{h}{a}, & 0 < h \leq a \\ C_0 + C1, & h > a \end{cases}$$

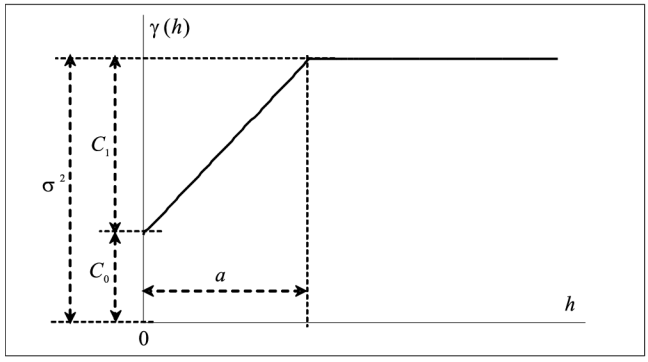


Figure 11.5: *Linear model*

11.2.2 Models Without Sill

Power Model

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C_1 h^{\alpha}, & h > 0 \end{cases} \qquad \alpha \in (0, 2)$$

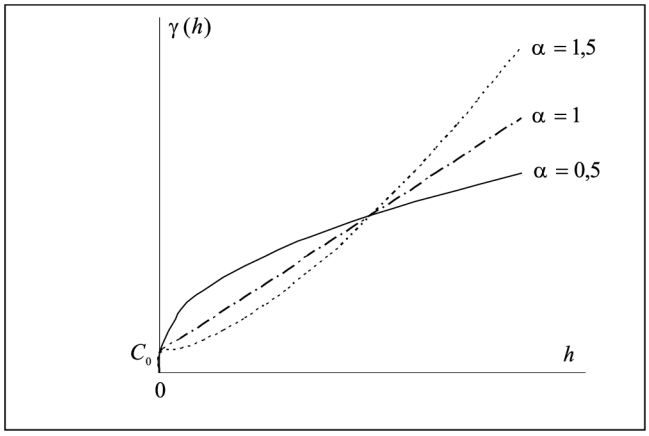


Figure 11.6: *Power model*

Logarithmic Model

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C_1 \ln h, & h > 0 \end{cases}$$

11.2.3 Oscillating Models

Sine Model

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C_1 \left(1 - \frac{\sin gh}{gh}\right), & h > 0 \end{cases} \qquad g = \frac{\pi}{\omega}$$

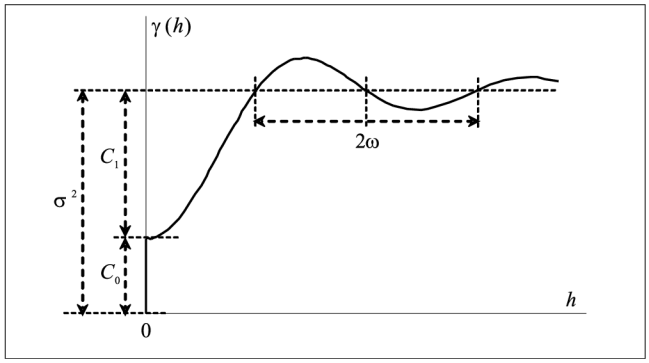


Figure 11.7: *Sine model*

Cosine Model

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C_1(1 - \cos gh), & h > 0 \end{cases} \qquad g = \frac{\pi}{\omega}$$

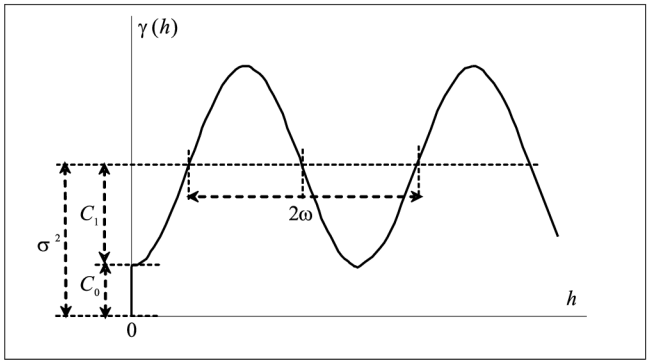


Figure 11.8: *Cosine model*

11.2.4 Pure Random Model

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0, & h > 0 \end{cases}$$

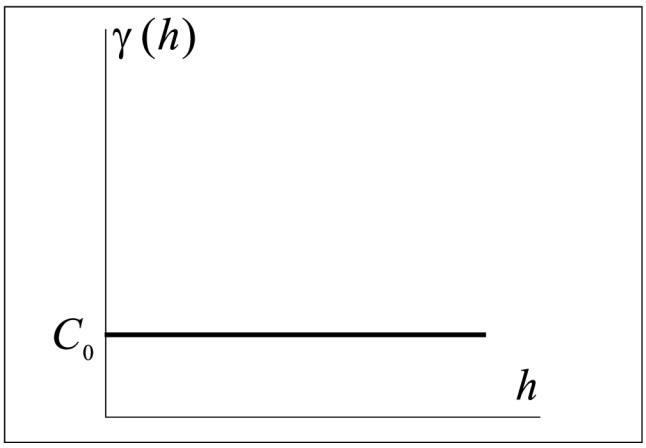


Figure 11.9: *Pure random model*

12 Spatial Autocovariance

12.1 Global Moran's and Geary's Indexes

As was written in Chapter 1, both indexes describe global spatial autocorrelation of the process. There were also introduced the computational formulas for them.

Of course, that the values both indexes depend on the $w(i,j)$, which is specified by the spatial weighting scheme chosen. In literature are presented several approaches of choosing these weights. The most popular is to choose

$$w(i,j) = \frac{A}{||\boldsymbol{x}_i - \boldsymbol{x}_j||^m},$$

where $||\boldsymbol{x}_i - \boldsymbol{x}_j||$ is the distance of the points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$; m is a parameter chosen by the user and A is a constant. Usually we put $m = A = 1$.

The variances of I and c will differ according to the data model employed. According to [7], under an assumption of normality we obtain

$$\mathbf{E}(I) = \frac{1}{n-1}, \qquad \mathbf{D}(I) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{S_0^2(2n-1)} - \left(\frac{1}{n-1}\right)^2,$$

where auxiliary variables $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w(i,j)$, $S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(w(i,j) + w(j,i))^2$ and $S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w(i,j) + \sum_{j=1}^{n} w(j,i)\right)^2$.

Standardized random variable

$$z = \frac{I - \mathbf{E}(I)}{\sqrt{\mathbf{D}(I)}} \sim \mathrm{N}(0,1)$$

and it is possible to test significance of Moran's index I.

Similarly, for the Geary's index c we have

$$\mathbf{E}c = 1, \qquad \mathbf{D}(c) = \frac{(n-1)(2S_1 + S_2) - 4S_0^2}{2S_0^2(n+1)},$$

where S_0, S_1 and S_2 have the same meaning as in the case of Moran's index. The next technique of testing of significance is the same as above.

13 Stochastic Processes: A Spectral Approach

13.1 White Noise Process

Gaussian white noise process is a good approximation of many real-world situations and generates mathematically tractable models, especially in physics and electrotechnics. But on the other side, it has many applications in many others area. We will use it in one of our algorithm for generating a random structure.

The White Noise Process: A *white noise process* is a random process of random variables that are uncorrelated, have mean zero, and a finite variance.

Formally, W_t is a white noise process if $\mathbf{E}(W_t) = 0$, $\mathbf{D}(W_t) = \lambda$, and $\mathbf{E}(W_t W_j) = 0$ for all $t \neq j$.

A common, slightly stronger condition is that they are independent from one another; this is an *independent white noise process.*

Often one assumes a normal distribution for the variables, in which case the distribution was completely specified by the mean and variance; these are "normally distributed" or "Gaussian" white noise processes. From the previous it follows, that white noise process is not continuous process and that is the reason why we are not able to draw it. For more information see e.g. [17], [23], [27] or [13].

13.2 Karhunen-Loève Expansion

As we said before, we would like to apply some properties of stochastic processes to develop an algorithm for generating random structure. This tool is called a spectral decomposition of a stochastic process. A theoretically appealing approach is to expand it in a Fourier-type series as

$$w(\boldsymbol{x}, \omega) = \sum_{n=0}^{\infty} \sqrt{\lambda_n} \xi_n(\omega) \Phi_n(\boldsymbol{x}),$$

where $\{\xi_n(\omega)\}$ is a set of random variables to be determined, λ_n is some constant and $\{\Phi_n(\boldsymbol{x})\}$ is an orthonormal set of deterministic functions. This is exactly what the Karhunen-Loève expansion achieves. The Karhunen-Loève expansion of a stochastic process is based on the following analytical properties of its covariance function.

Let $w(\boldsymbol{x}, \omega)$ be a random process, function of the position vector $\boldsymbol{x}$ defined over domain D, with ω belonging to the space of random events Ω. Next, let $\bar{w}(\boldsymbol{x}) = \mathbf{E}[w(\boldsymbol{x}, \omega)]$ denotes the expected value of $w(\boldsymbol{x}, \omega)$ over all possible realizations of the process and $C(x_1, x_2)$ denotes its covariance function. By definition of the covariance function, it is bounded, symmetric and positive definite. This fact simplifies the ensuing analysis

considerably in that it guarantees a number of properties for the eigenfunctions and the eigenvalues that are solution to the previous equation:

- The set $\Phi_i(x)$ of eigenfunctions is orthogonal and complete.
- For each eigenvalue λ_k, there correspond at most a finite number of linearly independent eigenfunctions.
- There are at most a countably infinite set of eigenvalues.
- The eigenvalues are all positive numbers.
- The kernel $C(x_1, x_2)$ admits of the following uniformly convergent expansion.

$$C(x_1, x_2) = \sum_{n=0}^{\infty} \lambda_n \Phi_n(x_1)\Phi_n(x_2),$$

where λ_n and $\Phi_n(\boldsymbol{x})$ denote the eigenvalues and eigenvectors of the appropriate covariance kernel, which we obtain by solving the following Fredholm equation of a second type

$$\int_D C(x_1, x_2)\Phi(x_2)\,\mathrm{d}x_2 = \lambda\Phi(x_1).$$

Due to the symmetry and the positive definiteness of the covariance kernel, see [13], its eigenfunctions are orthogonal and form a complete set. They can be normalized according to the following criterion

$$\int_D \Phi_n(\boldsymbol{x})\Phi_m(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \delta_{nm},$$

where δ_{nm} is the Kronecker delta. Then, $w(\boldsymbol{x}, \omega)$ can be written as

$$w(\boldsymbol{x}, \omega) = \bar{w}(\boldsymbol{x}) + \alpha(\boldsymbol{x}, \omega),$$

where $\alpha(\boldsymbol{x}, \omega)$ is a process with zero mean and covariance function $C(x_1, x_2)$. The process $\alpha(\boldsymbol{x}, \omega)$ can be expanded in terms of the eigenfunctions $\Phi_n(\boldsymbol{x})$ as

$$\alpha(\boldsymbol{x}, \omega) = \sum_{n=0}^{\infty} \xi_n(\omega)\sqrt{\lambda_n}\Phi_n(\boldsymbol{x}). \tag{13.2.1}$$

Thus, the random process $w(\boldsymbol{x}, \omega)$ can be written as

$$w(\boldsymbol{x}, \omega) = \bar{w}(\boldsymbol{x}) + \sum_{n=0}^{\infty} \xi_n(\omega)\sqrt{\lambda_n}\Phi_n(\boldsymbol{x}),$$

where $\mathbf{E}\left[\xi_n(\omega)\right] = 0$, $\mathbf{E}\left[\xi_n(\omega)\xi_m(\omega)\right] = \delta_{nm}$ and λ_n, $\Phi_n(\boldsymbol{x})$ are solution to the integral equation. Truncating the series in previous equation at the K^{th} term, gives

$$w(\boldsymbol{x}, \omega) = \bar{w}(\boldsymbol{x}, \omega) + \sum_{n=0}^{K} \xi_n(\omega)\sqrt{\lambda_n}\Phi_n(\boldsymbol{x}), \quad (\boldsymbol{x}, \omega) \in D \times \Omega.$$

An explicit expression for $\xi_n(\omega)$ can be obtained by multiplying equation 13.2.1 by $\Phi_n(\boldsymbol{x})$ and integrating over the domain D. That is

$$\xi_n(\omega) = \frac{1}{\sqrt{\lambda_n}} \int_D \alpha(\boldsymbol{x}, \omega) \Phi_n(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

It can be proved, that $\mathbf{E}\left[(w - w_k)^2(\boldsymbol{x}, \omega)\right] \to 0$ for $K \to \infty$. The most important value of spectral decomposition lies in the fact, that spatial random deviations we can express as a sum of deterministic functions in spatial coordinates multiplied by random variables, which are independent on these coordinates.

13.3 Brownian Motion

Brownian motion(also *Wiener process*) plays a very important role in probability theory, the theory of stochastic processes, physics, finance, etc. Brownian motion is named after the biologist Robert Brown whose research dates to the 1820s. Wiener(1923) was the first to put Brownian motion on a firm mathematical basis.

Brownian Motion: A stochastic process $B = (B_t,\ t \in \langle 0, \infty))$ is called *(standard) Brownian motion* or a *Wiener process* if the following conditions are satisfied:

1. $B_0 = 0$ and it has continuous sample paths
2. For $0 < t_0 < t_1 < \ldots < t_n$ the increments $B_{t_1} - B_{t_0}, \ldots, B_{t_n} - B_{t_{n-1}}$ independent
3. For every $t > 0$ and $h > 0$, $B_{t+h} - B_t$ has a normal distribution with zero mean and variance h, i.e. $B_{t+h} - B_t \sim \mathrm{N}(0,\, h)$.

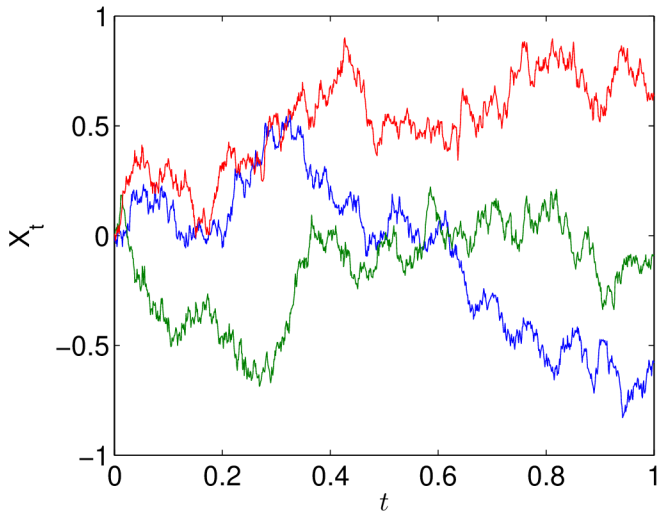


Figure 13.1: *Three realizations of the Brownian motion.*

Next, for a Brownian motion B_t it holds:

1. $\mathbf{E}\left[B_t^2\right] = t$
2. $\mathbf{E}\left[B_t B_s\right] = \min\left\{t, s\right\}, \quad t \leq 0,\ s \leq 0$
3. $\mathbf{E}\left[(B_t - B_s)^2\right] = |t - s|$.

Now, we return to the spectral decomposition of a stochastic process. According to [19] or [13], for the Brownian motion defined on the set $D = \langle 0, T\rangle$ it can be proved, that for eigenvalues and eigenfunctions stand the following relations

$$\Phi_n(t) = \sqrt{2}\sin\left(\frac{x}{\sqrt{\lambda_n}}\right), \qquad \lambda_n = \frac{4T^2}{\pi^2(2n+1)^2}, \quad n = 0, 1, 2, \ldots, \quad t \in \langle 0, T\rangle.$$

Brownian motion we can then then write

$$B_t = \sum_{n=0}^{\infty} \frac{2\sqrt{2}T}{\pi(2n+1)} \sin\left(\frac{(2n+1)\pi t}{2T}\right) \xi_n(\omega).$$

From the previous it follows that replacing the previous sum by the only finite members we get also trajectory similar to the Brownian motion, but it will be "smoother" than then real one.

Now we use Karhunen-Loève expansion of Brownian motion with finite number of members ($K < \infty$) for a introduction of a new stochastic process $S(t, \omega)$ defined on an interval $\langle 0, T\rangle$. This process we later use for simulation of a random structure:

$$S(t, \omega) = \sum_{n=0}^{K} \sin\left(\frac{(2n+1)\pi t}{2T}\right) \xi_n(\omega),$$

where $\{\xi_n\}_{n=1}^{\infty}$ is a sequence of mutually independent random variables. The process $S(t, \omega)$ has a character of a *"white noise process"* and for our simulation is quite sufficient. In the following picture we can see the realizations of the process with different K:

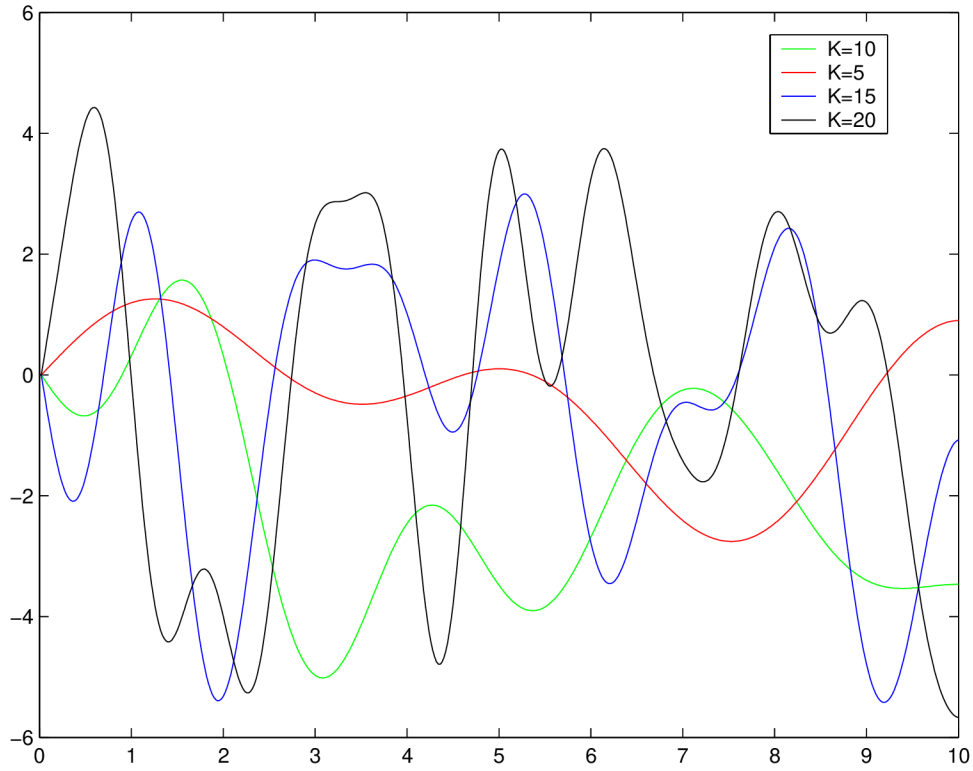


Figure 13.2: *Trajectories of a stochastic process* $S(t,\omega)$ *for different* K.

13.4 Brownian Bridge

A Brownian bridge is a continuous-time stochastic process whose probability distribution is the conditional probability distribution of a Brownian motion $B(t)$ given the condition that $W(0) = W(1) = 0$.

The expected value of the bridge is zero, with variance $t(1-t)$, implying that the most uncertainty is in the middle of the bridge, with zero uncertainty at the nodes. The covariance of $W(s)$ and $W(t)$ is $s(1-t)$ if $s < t$. The increments in a Brownian bridge are not independent. If $W(t)$ is a standard Wiener process (i.e., for $t \geq 0$, $B(t)$ is normally distributed with expected value 0 and variance t, and the increments are stationary and independent), then $W(t) - tW(1)$ is a Brownian bridge. Conversely, if B is a Brownian bridge and Z is an independent standard Gaussian random variable, then the process $W(t) = B(t) + tZ$ is a Brownian motion for $t \in [0,1]$. More generally, a Brownian motion $W(t)$ for $t \in [0,T]$ can be decomposed into

$$x(t) = B\left(\frac{t}{T}\right) + \frac{t}{\sqrt{T}}Z.$$

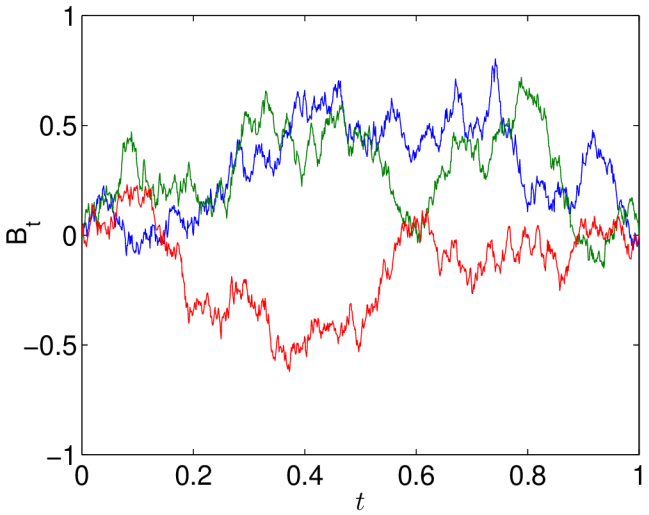


Figure 13.3: *Three realizations of the Brownian bridge.*

A Brownian bridge is the result of Donsker's theorem in the area of empirical processes. It is also used in the Kolmogorov-Smirnov test in the area of statistical inference. A standard Brownian motion satisfies $W(0) = 0$ and is therefore "tied down" to the origin, but other points are not restricted. In a Brownian bridge process on the other hand, not only is $B(0) = 0$ but we also require that $B(1) = 0$, that is the process is "tied down" at $t = 1$ as well. Just as a literal bridge is supported by pylons at both ends, a Brownian Bridge is required to satisfy conditions at both ends of the interval $[0, 1]$. (In a slight generalization, one sometimes requires $B(t_1) = \text{a}$ and $B(t_2) = b$ where t_1, t_2, a and b are known constants.)

Suppose we have generated a number of points $W(0), W(1), W(2), W(3)$, etc. of a Brownian motion path by computer simulation. It is now desired to fill in additional points in the interval $[0, 1]$, that is to interpolate between the already generated points $W(0)$ and $W(1)$. The solution is to use a Brownian bridge that is required to go through the values $W(0)$ and $W(1)$.

For the general case when $W(t_1) = a$ and $W(t_2) = b$, the distribution of W at time $t \in (t_1, t_2)$ is normal, with mean

$$a + \frac{t - t_1}{t_2 - t_1}(b - a) \qquad \text{and variance} \quad \frac{(t - t_1)(t_2 - t)}{t_2 - t_1}.$$

14 Selected Distributions

14.1 Weibull Distribution

In probability theory and statistics, the Weibull distribution is a continuous probability distribution. It is often used to describe the size distribution of particles.

14.1.1 Motivating the Weibull model

Assume a CSR pattern(Poisson process) of intensity λ(λ= mean number of events per unit area). Let X be a number of events in an area of size $A = \pi r^2$. Then $X \sim \text{Po}(\lambda A)$, where

$$\mathbf{P}(X = x) = \frac{(\lambda A)^x \mathrm{e}^{-\lambda A}}{x!}, \quad x = 0, 1, 2, \ldots.$$

Let the random variable R denotes the distance from a randomly selected point(cross-mark) to the nearest event(dot). Hence,

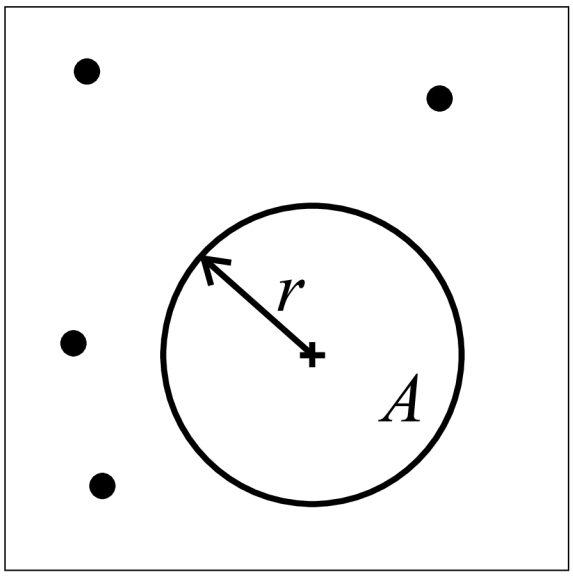


Figure 14.1: *Circle of radius r in area A centered on a randomly selected point.*

$$\begin{aligned}
\mathbf{P}(R > r) &= \mathbf{P}(\text{no events occur inside the circle of radius r}) = \\
&= \mathbf{P}(\text{no events occur in an area } A = \pi r^2) = \\
&= \mathbf{P}(X = 0), \text{ where } X \sim \text{Po}(\lambda A) = \\
&= \exp(-\lambda A) = \\
&= \exp(-\lambda \pi r^2).
\end{aligned}$$

Therefore the cumulative distribution function of R is $F(r) = \mathbf{P}(R \leq r) = 1 - \exp -\lambda \pi r^2$. Hence the probability density function of R is

$$f(r) = \frac{\mathrm{d}F(r)}{\mathrm{d}r} = 2\lambda \pi r \mathrm{e}^{-\lambda \pi r^2}, \quad r \geq 0.$$

14.1.2 Properties of Weibull Distribution

A continuous random variable R which has the probability distribution function

$$f(r) = 2\lambda\pi r \mathrm{e}^{-\lambda\pi r^2}, \quad r \geq 0,\ \lambda > 0,$$

follows a Weibull distribution. We derive the expectation as follows. By definition we have

$$\mathbf{E}[R] = \int_0^\infty r f(r)\,\mathrm{d}r = \int_0^\infty 2\lambda\pi r^2 \mathrm{e}^{-\lambda\pi r^2}\,\mathrm{d}r.$$

Let $y = \lambda\pi r^2$, hence $r = \sqrt{\frac{y}{\lambda\pi}}$ and $\mathrm{d}r = \frac{\mathrm{d}y}{2\sqrt{\lambda\pi y}}$. Recall, the gamma function $\Gamma(k)$ has the form

$$\Gamma(k) = \int_0^\infty z^{k-1}\mathrm{e}^{-z}\,\mathrm{d}z.$$

It may be shown, that $\Gamma(k) = (k-1)\Gamma(k-1)$ and $\Gamma(1/2) = \sqrt{\pi}$. Using these facts we find that

$$\mathbf{E}[R] = \int_0^\infty 2y\frac{\mathrm{e}^{-y}}{2\sqrt{\lambda\pi y}}\,\mathrm{d}y = \int_0^\infty \frac{1}{\sqrt{\lambda\pi}}\sqrt{y}\mathrm{e}^{-y}\,\mathrm{d}y = \frac{1}{\sqrt{\lambda\pi}}\Gamma\left(\frac{3}{2}\right) =$$

$$= \frac{1}{\sqrt{\lambda\pi}}\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{1}{\sqrt{\lambda\pi}}\frac{1}{2}\sqrt{\pi} = \frac{1}{2\sqrt{\lambda}}.$$

Next, we use a similar approach to find $\mathbf{E}[R^2]$

$$\mathbf{E}\left[R^2\right] = \int_0^\infty 2\lambda\pi r^3 \mathrm{e}^{-\lambda\pi r^2}\,\mathrm{d}r = \int_0^\infty \frac{1}{\sqrt{\lambda\pi}}\sqrt{y}\mathrm{e}^{-y}\frac{\sqrt{y}}{\sqrt{\lambda\pi}}\,\mathrm{d}y =$$

$$= \frac{1}{\lambda\pi}\int_0^\infty y\mathrm{e}^{-y}\,\mathrm{d}y = \frac{1}{\lambda\pi}\Gamma(2) = \frac{1}{\lambda\pi}.$$

Then we obtain

$$\mathbf{D}[R] = \mathbf{E}\left[R^2\right] - \mathbf{E}^2[R] = \frac{1}{\lambda\pi} - \frac{1}{4\lambda} = \frac{4-\pi}{4\lambda\pi}.$$

The obtained mean value and variance were utilized in the Clark-Evans test of CSR.

14.2 Fischer-Snedecor's Distribution

The following section provide an overview of the F distribution.

Background of the F distribution.

The F distribution has a natural relationship with the chi-square distribution. If χ_1 and χ_2 are both chi-squared with m and n degrees of freedom respectively, then the statistic F below is F distributed:

$$F(m,n) = \frac{\frac{\chi_1}{m}}{\frac{\chi_2}{n}}.$$

Definition of the F distribution.

Fischer-Snedecor distribution with m and n degrees of freedom has probability density function

$$f_{m,n}(x) = \begin{cases} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{x^{\frac{m-2}{2}}}{\left(1+\left(\frac{m}{n}\right)x\right)^{\frac{m+n}{2}}}, & y > 0; \\ 0, & y \leq 0, \end{cases}$$

where $\Gamma(\cdot)$ is the Gamma function, defined by

$$\Gamma(x) \equiv \lim_{n\to\infty} \prod_{v=0}^{n-1} \frac{n!\, n^{x-1}}{x+v} = \lim_{n\to\infty} \frac{n!\, n^{x-1}}{x(x+1)(x+2)\dots(x+n-1)} \equiv \int_0^\infty \mathrm{e}^{-t} t^{x-1}\, \mathrm{d}t.$$

The integral definition is valid only for $x > 0$ (2^{nd} Euler integral).
The most common application of the F distribution is in standard tests of hypotheses in analysis of variance and regression.

The next figure shows that the F distribution exists on the positive real numbers and is skewed to the right.

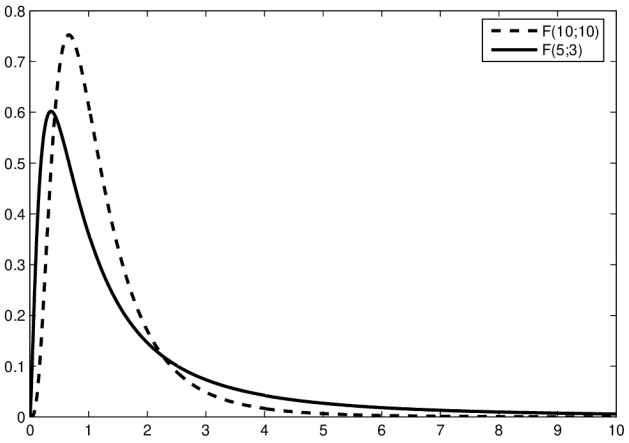


Figure 14.2: *Examples of Fischer-Snedecor probability functions.*

The mean, variance, skewness and kurtosis are, see [1]:

$$\mu = \frac{n}{/}(n-2) \quad \text{for } n \geq 2$$

$$\sigma^2 = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad \text{for } n > 4$$

$$a_3 = \frac{(2(n+2m-2)}{n-6}\sqrt{\frac{2(n-4)}{m(m+n-2)}} \quad \text{for } n > 6$$

$$a_4 = \frac{12(-16+20n-8n^2+n^3+44m-32mn+5n^2m-22m^2+5mn^2)}{m(n-6)(n-8)(n+m-2)} \quad \text{for } n > 8.$$

Note.

- It can be proved, that for $m+n$ joint independent variables $X_1, \ldots, X_m,\ Y_1, \ldots, Y_n$ with the same distribution $\mathrm{N}(0;1)$, the variable

$$Y = \frac{n(X_1^2 + \ldots + X_m^2)}{m(Y_1^2 + \ldots + Y_n^2)}$$

  has Fischer-Snedecor's distribution with probability function $f_{m,n}$.

- If $X \sim \mathrm{F}(m,n)$, then $Y = \lim_{n\to\infty} mX$ has the chi-square distribution χ^2_m.

- If $X \sim \mathrm{F}(m,n)$, then $\dfrac{1}{X} \sim \mathrm{F}(n,m)$.

- If $X \sim \mathrm{t}(\nu)$ has Student's distribution, then $X^2 \sim \mathrm{F}(m=1, n=\nu)$.

- For the critical values of the F distribution it holds

$$\mathrm{F}_\alpha(m,n) = \frac{1}{\mathrm{F}_{1-\alpha}(m,n)}.$$

- A generalization of the (central) F-distribution is the *noncentral F-distribution.* It is the distribution of the test statistic in analysis of variance problems when the null hypothesis is false. One uses the noncentral F-distribution to find the power function of such a test.

14.3 F-Tests

An F-Test is any statistical test in which the test statistic has an F-distribution if the null hypothesis is true.

14.3.1 Two-Sample F-Test

In order to compare two methods, it is often important to know whether the variabilities for both methods are the same. In order to compare two variances v_1 and v_2, one has to calculate the ratio of the two variances. This ratio is called the F-statistic (in honor of R.A. Fisher) and follows an F distribution:

$$F = \frac{v_1}{v_2}.$$

The null hypothesis H_0 assumes that the variances are equal and the ratio F is therefore one. The alternative hypothesis H_1 assumes that v_1 and v_2 are different, and that the ratio deviates from unity. The F-test is based on two assumptions:

- the samples are normally distributed,

- the samples are independent of each other.

When these assumptions are fulfilled and H_0 is true, the statistic F follows a F-distribution. The following is a decision table for the application of the two-sample F-test.

	One-tailed test		**Two-tailed test**
Hypothesis	$H_0 : \sigma_1^2 \geq \sigma_2^2$ $H_1 : \sigma_1^2 < \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$
Test statistics	$F=\frac{s_2^2}{s_1^2}$	$F=\frac{s_1^2}{s_2^2}$	$F=\frac{\text{larger sample variance}}{\text{smaller sample variance}}$
Deg. of freedom	$df_1=n_1-1$		$df_2=n_2-1$
Rejection	reject H_0 if $F>F_\alpha$		reject H_0 if $F>F_{\alpha/2}$

Table 14.1: *A decision table for the two-sample F-test.*

Remarks:

- When the normality assumption is not fulfilled, one should use a non-parametric method. In general the F-test is more sensitive to deviations from normality than the t-test.
- The F-test can be used to check the equal variance assumption needed for the two sample t-test, but the non-rejection of H_0 does not imply that the assumption (of equal variance) is valid, since the probability of the type II error is unknown.
- Note that when there are only two groups for the F-test, $F = t^2$, where t is the Student's t statistic.

Types of Errors

In general, there are two different types of error that can occur when making a decision:

- the error of the **first kind** ("Type I errors") are those errors which occur when we reject the null hypothesis although the null hypothesis is true.
- the error of the **second kind** ("Type II errors") arise when we accept the null hypothesis although the alternative hypothesis is true.

		Reality	
		H_0=true	H_0=false
Our Decision	H_0=true	OK	Error Type II
	H_0=false	Error Type I	OK

Table 14.2: *To the explanation of the Type I and Type II error.*

In summary:

- Rejecting a null-hypothesis when it should have been accepted creates a Type I error.
- Accepting a null-hypothesis when it should have been rejected creates a Type II error.
- In either case, a wrong decision or error in judgment has occurred.
- Decision rules (or tests of hypotheses), in order to be good, must be designed to minimize errors of decision.
- Minimizing errors of decision is not a simple, because for any given sample size, any effort to reduce one type of error is generally associated with an increase in the other type of error.
- In practice, one type of error may be more serious than the other.
- In such cases, a compromise should be reached in favor of limiting the more serious type of error.
- The only way to minimize both types of error is to increase the sample size; and such a move may or may not be feasible.

14.3.2 N-sample F-Test

In the case of multiple-comparison ANOVA problems, the F-test is used to test if the variance measuring the differences between groups in a certain pre-defined grouping of observations is large compared to the variance measuring the differences within the groups: a large value would tend to suggest that grouping is good or valid in some sense or that there are real differences between the groups. The formula for an F-test is:

$$F = \frac{\text{(explained variance)}}{\text{(unexplained variance)}}$$

or

$$F = \frac{\text{(between-group variability)}}{\text{(within-group variability)}},$$

where the quantities on the top and bottom of this ratio are each unbiased estimates of the within-group variance on the assumption that the between group variance is zero. An F test in ANOVA can only tell you if there is a relationship between two variables – it can't tell you what that relationship is. Mathematically, this means it can only tell you if one of the means of the groups is different from another one. It can't tell you which mean is different. More information about F-Test in ANOVA see e.g. [3], [2], [21], [16] or [15].

15 Ellipse Fitting

In the analysis of an izotropy we need to fit an ellipse-so called *rose diagram* to the set of points obtained by directional variograms. We will focus on least-square fitting, see [11] and references therein for details. Least-squares techniques center on finding the set of parameters that minimize some distance between the data points and the ellipse.

The equation describing a general conic by an implicit second order polynomial can be written as

$$F(\mathbf{a}, \mathbf{x}) = \mathbf{a} \cdot \mathbf{x} = ax^2 + bxy + cy^2 + dx + ey + f = 0, \tag{15.0.1}$$

where $\mathbf{a} = [a, b, c, d, e, f]^T$ and $\mathbf{x} = [x^2, xy, y^2, x, y, 1]^T$. $F(\mathbf{a}, \mathbf{x})$ is called *algebraic distance* of a point (x, y) to the conic $F(\mathbf{a}, \mathbf{x}) = 0$. The fitting of a general conic may be approached by minimizing the sum of squared algebraic distances

$$\mathcal{D}(\mathbf{a}) = \sum_{i=1}^{N} F(\mathbf{x}_i)^2 \tag{15.0.2}$$

of the curve to the N data points $\mathbf{x}_i$. In order to fit ellipses specifically while retaining the efficiency of solution of the linear least-squares problem 15.0.2, we would like to constrain the parameter vector $\mathbf{a}$ so that the conic that it represents is forced to be an ellipse. The appropriate constraint is well known, namely, that the *discriminant of quadratic members* be negative(see e.g. [29]), i.e.

$$b^2 - 4ac < 0.$$

However, this constrained problem is difficult to solve in general as the Karush-Kuhn-Tucker conditions(necessary for a solution in nonlinear programming to be optimal, see e.g. [36]), do not guarantee a solution.

Although the imposition of this inequality constraint is difficult in general, in this case we have a freedom to arbitrarily scale the parameters so we may simply incorporate the scaling into the constraint and impose the *equality* constraint $4ac - b^2 = 1$. This is a quadratic constraint which may be expressed in the matrix form $\mathbf{a}^T C a = 1$ as

$$\mathbf{a}^T \begin{pmatrix} 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{a} = 1 \tag{15.0.3}$$

Now, the constrained ellipse fitting problem reduces to

$$\min_{\mathbf{a}} \|\mathbf{D}\mathbf{a}\|^2 \quad \text{subject to the constraint } \mathbf{a}^T\mathbf{C}\mathbf{a} = 1, \tag{15.0.4}$$

where the *design matrix* $\mathbf{D}$ is defined as $\mathbf{D} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$. Introducing the Lagrange multiplier λ and differentiating, we arrive at the system of simultaneous equations

$$\begin{aligned} 2\mathbf{D}^T\mathbf{D}\mathbf{a} - 2\lambda\mathbf{C}\mathbf{a} &= 0 \\ \mathbf{a}^T\mathbf{C}\mathbf{a} &= 1. \end{aligned} \tag{15.0.5}$$

This may be rewritten as a system

$$\mathbf{Sa} = \lambda \mathbf{Ca} \tag{15.0.6}$$

$$\mathbf{a}^T \mathbf{Ca} = 1, \tag{15.0.7}$$

where $\mathbf{S}$ is the *scatter matrix* $\mathbf{D}^T\mathbf{D}$. This system is readily solved by considering the generalized eigenvectors of 15.0.6. If $(\lambda_i, \mathbf{u}_i)$ solves 15.0.6, then so does $(\lambda_i, \mu\mathbf{u}_i)$ for any μ and from 15.0.7 we can find the value of μ_i as $\mu_i^2 \mathbf{u}_i^T \mathbf{C} \mathbf{u}_i = 1$, giving

$$\mu_i = \sqrt{\frac{1}{\mathbf{u}_i^T \mathbf{C} \mathbf{u}_i}} = \sqrt{\frac{1}{\mathbf{u}_i^T \mathbf{S} \mathbf{u}_i}}. \tag{15.0.8}$$

Finally, setting $\hat{\mathbf{a}}_i = \mu_i \mathbf{u}_i$ solves 15.0.5.

We note that the solution of the eigensystem 15.0.6 gives six eigenvalue-eigenvector pairs $(\lambda_i, \mathbf{u}_i)$. Each of these pairs gives rise to a local minimum if the term under the square root of 15.0.8 is positive. In general, $\mathbf{S}$ is positive definite, so the denominator $\mathbf{u}_i^T \mathbf{S} \mathbf{u}_i$ is positive for all $\mathbf{u}_i$. Therefore, the square root exists if $\lambda_i > 0$, so any solutions to 15.0.5 must have positive generalized eigenvalues. It can be proved, that the minimization $\|\mathbf{Da}\|^2$ subject to $4ac - b^2 = 1$ yields exactly one solution, which corresponds, by virtue of the constraint, to an ellipse, see [11].

16 Normality Tests

In statistics, normality tests are used to determine whether a random variable is normally distributed, or not.

One application of normality tests is to the residuals from a linear regression model. If they are not normally distributed, the residuals should not be used in Z tests or in any other tests derived from the normal distribution, such as t tests, F tests and chi-square tests. If the residuals are not normally distributed, then the dependent variable or at least one explanatory variable may have the wrong functional form, or important variables may be missing, etc. Correcting one or more of these systematic errors may produce residuals that are normally distributed.

Normality tests include D'Agostino's K-squared test, the Jarque–Bera test, the Anderson–Darling test, the Cramér–von-Mises criterion, the Lilliefors test for normality (itself an adaptation of the Kolmogorov–Smirnov test), the Shapiro–Wilk test, the Pearson's chi-square test and the Shapiro–Francia test for normality.

Instead of using formal normality tests, another option is to compare a histogram of the residuals to a normal probability curve. The actual distribution of the residuals (the histogram) should be bell-shaped and resemble the normal distribution. This might be difficult to see if the sample is small. In this case one might proceed by regressing the measured residuals against a normal distribution with the same mean and variance as the sample. If the regression produces an approximately straight line, then the residuals can safely be assumed to be normally distributed. Among other graphical tools are the quantile-quantile plot and the normal probability plot.

16.1 Jarque–Bera Test

In statistics, the Jarque-Bera test is a goodness-of-fit measure of departure from normality, based on the sample kurtosis and skewness. This test is based on the fact that skewness and kurtosis of normal distribution equal to zero. Therefore, the absolute value of these parameters could be a measure of deviation of the distribution from normal. The test statistic JB is defined as

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right),$$

where n is the number of observations (or degrees of freedom in general); S is the sample skewness, K is the sample kurtosis, defined as

$$S = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\left(\sigma^2\right)^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right)^{3/2}},$$

$$K = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\left(\sigma^2\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right)^2}.$$

where μ_3 and μ_4 are the third and fourth central moments, respectively, $\overline{x}$ is the sample mean, and σ^2 is the second central moment, the variance. Therefore, this can be considered as a sort of portmanteau test, since the four lowest moments about the origin are used jointly for its calculation.

The statistic JB has an asymptotic chi-square distribution with two degrees of freedom and can be used to test the null hypothesis that the data are from a normal distribution. The null hypothesis is a joint hypothesis of the skewness being zero and the excess kurtosis being 0, since samples from a normal distribution have an expected skewness of 0 and an expected excess kurtosis of 0 (which is the same as a kurtosis of 3). As the definition of JB shows, any deviation from this increases the JB statistic. The Jarque-Bera test is an asymptotic test, and should not be used with small samples.

16.2 Ryan-Joiner Test

This test basically compares the unknown distribution with a normal distribution to see if they differ in shape. A correlation coefficient r is used as the test statistic and the closer r is to 1.0 the greater confidence we have that the unknown distribution is indeed normal. The exact values of r for a given confidence interval depend upon the number of points considered.

The Ryan-Joiner test, which is similar to Shapiro-Wilk test, is based on regression and correlation. The test tends to work well in identifying a distribution as not normal when the distribution under consideration is skewed. It is less discriminating when the underlying distribution is a t-distribution and non-normality is due to kurtosis. We can use the Ryan-Joiner statistic RJ to test the hypothesis, H_0: the data $\{x_1, ..., x_n\}$ are a random sample of size n from a normal distribution, H_1: the data are a random sample from some other distribution. The test statistic RJ is the correlation between the data and the normal scores. If the data are a sample from a normal distribution then the normal probability plot will be close to a straight line. The correlation RJ will be close to one and if the data are sampled from a non-normal distribution then the plot will exhibit some degree of curvature, resulting in a smaller correlation RJ. Small values of RJ are therefore regarded as strong evidence against H_0. The Ryan-Joiner test is given by the formula for the correlation coefficient, namely

$$RJ = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \sum_{i=1}^{n}(b_i - \overline{b})^2}}.$$

Since $\overline{b} = 0$, RJ can be simplified to

$$RJ = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})b_i}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \sum_{i=1}^{n} b_i^2}},$$

where Y_i are the ordered observations in a sample of size n and b_i is the p^{th} percentage point of the standard normal distribution, that is, $b_i = \Phi^{-1}(p_i) = \sqrt{2}\text{erf}^{-1}(2p-1)$, where

$\Phi^{-1}(\cdot)$ is the inverse cumulative distributive function, or quantile function, which can be expressed in terms of the *inverse error function.* This quantile function is sometimes called the probit function. The values p_i can be obtained by $p_i = \frac{i-\frac{3}{8}}{n+\frac{1}{4}}$. The statistic RJ can be used to provide an indication of how non-normal the revisions are. This will be particularly true with larger samples. The test has the desirable feature of linking together a graphical display of the data with a simple, objective test statistic. Some may object to the use of the term correlation coefficient since the b_i are not random variables. However, given any set of points in the plane, one can use the correlation coefficient associated with those points as a descriptive measure of how close they are to a straight line. In this sense, RJ can be thought of as a correlation coefficient. Since RJ does not arise from sampling a bivariate distribution, it is not the same as the usual correlation coefficient. Approximate critical values $CV(n)$ of RJ were obtained from Monte Carlo simulations. The results were then smoothed, and for $\alpha = 0,05$ it holds

$$CV(n) = 1,0063 - \frac{0,1288}{\sqrt{n}} - \frac{0,6118}{n} + \frac{1,3505}{n^2}.$$

More detailed description of this test you can find in [31].

16.3 D'Agostino's K-squared Test

In statistics, D'Agostino's K^2 test is a goodness-of-fit measure of departure from normality, based on transformations of the sample kurtosis and skewness. The test statistic K^2 is obtained as follows: In the following derivation, n is the number of observations (or degrees of freedom in general); a_3 is the sample skewness, a_4 is the sample kurtosis, defined as

$$a_3 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{(\sigma^2)^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}},$$

$$a_4 = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{(\sigma^2)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2},$$

where μ_3 and μ_4 are the third and fourth central moments, respectively, $\overline{x}$ is the sample mean, and σ^2 is the second central moment, the variance.

Transformed Skewness

First, calculate $Z(a_3)$, a transformation of the skewness a_3, that is approximately normally distributed under the null hypothesis that the data are normally distributed. However, in practice it can be used only for large sample. Denote by

$$U(a_3) = \frac{a_3}{\sqrt{\mathbf{D}(a_3)}} = a_3 \cdot \sqrt{\frac{(n+1)(n+3)}{6(n-2)}},$$

$$b = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)},$$

$$W^2 = \sqrt{2\,(b-1)} - 1, \qquad \delta = \frac{1}{\sqrt{\ln(W)}}, \qquad \alpha = \sqrt{\frac{2}{W^2-1}},$$

$$Z\,(a_3) = \delta \ln\left(\frac{U(a_3)}{\alpha} + \sqrt{\left(\frac{U(a_3)}{\alpha}\right)^2 + 1}\right)$$

Transformed Kurtosis

Next, calculate $Z(a_4)$, a transformation of the kurtosis a_4 that is approximately normally distributed under the null hypothesis that the data are normally distributed.

$$\mathbf{E}\,[a_4] = \frac{3(n-1)}{n+1}, \qquad \mathbf{D}[a_4] = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}, \qquad U(a_4) = \frac{a_4 - \mathbf{E}\,[a_4]}{\mathbf{D}[a_4]}.$$

Next, compute the skewness of the kurtosis:

$$B = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}},$$

$$A = 6 + \frac{8}{B}\left[\frac{2}{B} + \sqrt{1+\frac{4}{B^2}}\right],$$

$$Z(a_4) = \left(\left(1-\frac{2}{9A}\right) - \sqrt[3]{\frac{1-\frac{2}{A}}{1+U(a_4)\sqrt{\frac{2}{A-4}}}}\right)\sqrt{\frac{9A}{2}}.$$

Omnibus K^2 statistic

Now, we can combine $Z(a_3)$ and $Z(a_4)$ to define D'Agostino's Ombibus K^2 test for normality:

$$\mathrm{K}^2 = Z(a_3)^2 + Z(a_4)^2.$$

K^2 is approximately distributed as χ^2 with 2 degrees of freedom. The null hypothesis we reject, if $\mathrm{K}^2 \geq \chi^2_2(\alpha)$. It is sufficient for $n \geq 20$. More tests utilizing the previous statistics are presented in [2].

16.4 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov statistic is defined as

$$D = \sup_x |F_n(x) - F(x)|.$$

The Kolmogorov-Smirnov statistic belongs to the supremum class of EDF statistics. This class of statistics is based on the largest vertical difference between $F(x)$ and $F_n(x)$. The

Kolmogorov-Smirnov statistic is computed as the maximum of D^+ and D^-, where D^+ is the largest vertical distance between the EDF and the distribution function when the EDF is greater than the distribution function, and D^- is the largest vertical distance when the EDF is less than the distribution function.

$$\begin{aligned} D^+ &= \max_i \left(\tfrac{i}{n} - U_{(i)}\right) \\ D^- &= \max_i \left(U_{(i)} - \tfrac{i-1}{n}\right) \\ D &= \max\left(D^+, D^-\right) \end{aligned}$$

The empirical cumulative distribution function F_n for n iid observations x_i is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 1 & \text{if } x_i \le x, \\ 0 & \text{otherwise.} \end{cases}$$

The Kolmogorov–Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x |F_n(x) - F(x)|,$$

where $F(x)$ is the hypothesized distribution or another empirical distribution. By the Glivenko–Cantelli theorem, if the sample comes from distribution $F(x)$, then D_n converges to 0 almost surely, i.e.

$$\mathbf{P}\left(\lim_{n\to\infty} D_n = 0\right) = 1.$$

The Kolmogorov distribution is the distribution of the random variable

$$K = \sup_{t\in[0,1]} |B(t)|,$$

where $B(t)$ is the *Brownian bridge*. The cumulative distribution function of K is given by

$$\mathbf{P}(K \le x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty} e^{-(2i-1)^2\pi^2/(8x^2)}.$$

Under null hypothesis that the sample comes from the hypothesized distribution $F(x)$,

$$\sqrt{n}D_n \xrightarrow{n\to\infty} \sup_t |B(F(t))|$$

in distribution, where $B(t)$ is the Brownian bridge. If F is continuous then under the null hypothesis $\sqrt{n}D_n$ converges to the Kolmogorov distribution, which does not depend on F. This result may also be known as the Kolmogorov theorem. The goodness-of-fit test or the Kolmogorov–Smirnov test is constructed by using the critical values of the Kolmogorov distribution. The null hypothesis is rejected at level α if

$$\sqrt{n}D_n > K_\alpha,$$

where K_α is found from

$$\mathbf{P}(K \le K_\alpha) = 1 - \alpha.$$

The asymptotic power of this test is 1. If the form or parameters of $F(x)$ are determined from the X_i, the inequality may not hold. In this case, Monte Carlo or other methods are required to determine the rejection level α.

16.5 Anderson-Darling Test

The Anderson-Darling statistic and the Cramér-von Mises statistic belong to the quadratic class of EDF statistics. This class of statistics is based on the squared difference $(F_n(x) - F(x))^2$. Quadratic statistics have the following general form:

$$Q = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \psi(x) dF(x).$$

The function $\psi(x)$ weights the squared difference $(F_n(x) - F(x))^2$. The Anderson-Darling statistic (A^2) is defined as

$$A^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \left[F(x)\left(1 - F(x)\right)\right]^{-1} dF(x).$$

Here the weight function is $\psi(x) = \left[F(x)\left(1 - F(x)\right)\right]^{-1}$. The Anderson-Darling statistic is computed as

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i - 1) \log F(x_i) + (2n + 1 - 2i) \log(1 - F(x_{n-i+1}))\right]$$

H_0: The data follow the specified distribution.
H_A: The data do not follow the specified distribution.

The hypothesis regarding the distributional form is rejected at the chosen significance level (alpha) if the test statistic, A^2, is greater than the critical value computed by auxiliary formulas, see [4] for details.

If testing for normal distribution of the variable X:

1. The data X_i, for $i = 1, \ldots n$, of the variable X that should be tested is sorted from low to high.

2. The mean $\overline{X}$ and standard deviation s are calculated from the sample of X.

3. The values X_i are standardized as

$$Y_i = \frac{X_i - \overline{X}}{s}$$

4. With the standard normal CDF Φ, A^2 is calculated using

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1)(\log \Phi(Y_i) + \log(1 - \Phi(Y_{n+1-i})))$$

5. A^{*2}, an approximate adjustment for sample size, is calculated using

$$A^{*2} = A^2 \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$$

6. If A^{*2} exceeds 0.752 then the hypothesis of normality is rejected for a 5% level test.

Note:

1. If $s = 0$ or any $\Phi(Y_i) = (0 \text{ or } 1)$ then A^2 cannot be calculated and is undefined.
2. Above, it was assumed that the variable X_i was being tested for normal distribution. Any other theoretical distribution can be assumed by using its CDF. Each theoretical distribution has its own critical values, and some examples are: lognormal, exponential, Weibull, extreme value type I and logistic distribution.
3. Null hypothesis follows the true distribution (in this case, N(0,1)).

16.6 Chi-Squared Test

The Chi-Squared test is used to determine if a sample comes from a population with a specific distribution. This test is applied to binned data, so the value of the test statistic depends on how the data is binned. Although there is no optimal choice for the number of bins k, there are several formulas which can be used to calculate this number based on the sample size N. For example, it can be used the following empirical formula:

$$k \sim +\log_2 N.$$

The data can be grouped into intervals of equal probability or equal width. The first approach is generally more acceptable since it handles peaked data much better. Each bin should contain at least 5 or more data points, so certain adjacent bins sometimes need to be joined together for this condition to be satisfied. The Chi-Squared statistic is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(Q_i - E_i)^2}{E_i},$$

where Q_i is the observed frequency for bin i, and E_i is the expected frequency for bin i calculated by

$$E_i = F(x_2) - F(x_1),$$

where $F(\cdot)$ is the CDF of the probability distribution being tested, and x_1, x_2 are the limits for bin i. The hypotheses are:

H_0: The data follow the specified distribution.

H_A: The data do not follow the specified distribution.

The hypothesis regarding the distributional form is rejected at the chosen significance level (α) if the test statistic is greater than the critical value defined as

$$\chi^2_{1-\alpha}(k-1)$$

meaning the Chi-Squared inverse CDF with $k - 1$ degrees of freedom and a significance level of α.

16.7 Shapiro-Wilk Test

In statistics, the Shapiro–Wilk test tests the null hypothesis that a sample $x_1, \ldots, x_n$ came from a normally distributed population. The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

where $x_{(i)}$ (with parentheses enclosing the subscript index i) is the i-th order statistic, i.e., the $i-$th smallest number in the sample; $\overline{x} = (x_1 + \cdots + x_n)/n$ is the sample mean and the constants a_i are given by

$$(a_1, \ldots, a_n) = \frac{m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}},$$

where

$$m = (m_1, \ldots, m_n)^\top$$

and $m_1, \ldots, m_n$ are the expected values of the order statistics of independent and identically-distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.

The user may reject the null hypothesis if W is too small. Accuracy is claimed for samples size from 3 to 5000. Sample size less than three will not produce a Shapiro-Wilk statistic.

16.8 Lilliefors test

In statistics, the Lilliefors test, is an adaptation of the Kolmogorov–Smirnov test. It is used to test the null hypothesis that data come from a normally distributed population, when the null hypothesis does not specify which normal distribution, i.e. does not specify the expected value and variance.

The test proceeds as follows: First estimate the population mean and population variance based on the data. Then find the maximum discrepancy between the empirical distribution function and the cumulative distribution function (CDF) of the normal distribution with the estimated mean and estimated variance. Just as in the Kolmogorov-Smirnov test, this will be the test statistic. Finally, we confront the question of whether the maximum discrepancy is large enough to be statistically significant, thus requiring rejection of the null hypothesis. This is where this test becomes more complicated than the Kolmogorov-Smirnov test. Since the hypothesized CDF has been moved closer to the data by estimation based on those data, the maximum discrepancy has been made smaller than it would have been if the null hypothesis had singled out just one normal distribution. Thus we need the "null distribution" of the test statistic, i.e. its probability distribution assuming the null hypothesis is true. This is the Lilliefors distribution. To date, tables for this distribution have been computed only by Monte Carlo methods.

17 Homogeneity Tests

The homogeneity of variance assumption is one of the critical assumptions underlying most parametric statistical procedures such as the analysis of variance and it is important to be able to test this assumption.

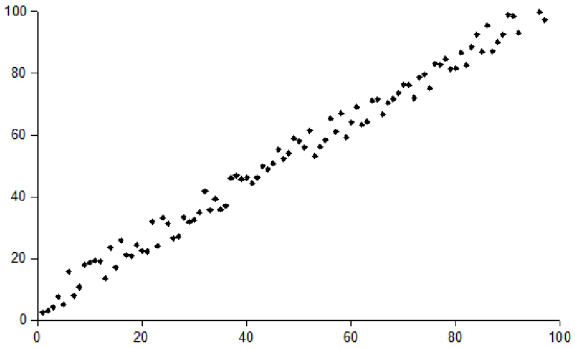


Figure 17.1: *Plot with random data showing homoscedasticity.*

In statistics, a sequence or a vector of random variables is *homoscedastic* if all random variables in the sequence or vector have the same finite variance. This is also known as homogeneity of variance. The complementary notion is called *heteroscedasticity*. In a scatterplot of data, homoscedasticity looks like an oval (most x values are concentrated around the mean of y, with fewer and fewer x values as y becomes more extreme in either direction). If a scatterplot looks like any geometric shape other than an oval, the rules of homoscedasticity may have been violated. Sometimes, so called *outliers* may be present in the sample and then it is suitable to remove such points and not include them into the analysis. More information about this problems you can find in any more extensive statistic book, see e.g. [3], [2], [21] or [14].

17.1 Bartlett's Test

Bartlett's test is used to test if k samples have equal variances. Equal variances across samples is called homoscedasticity or homogeneity of variances. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Bartlett test can be used to verify that assumption.

Bartlett's test is sensitive to departures from normality. That is, if your samples come from non-normal distributions, then Bartlett's test may simply be testing for non-normality. The Levene test and Brown-Forsythe test are alternatives to the Bartlett test that are less sensitive to departures from normality.

Bartlett's test is used to test the null hypothesis, H_0 that all k population variances are equal against the alternative that at least two are different.

If there are k samples with size n_i and sample variance S_i^2 then Bartlett's test statistic is

$$X^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i-1)\ln(S_i^2)}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(\frac{1}{n_i-1}) - \frac{1}{N-k}\right)},$$

where $N = \sum_{i=1}^{k} n_i$ and $S_p^2 = \frac{1}{N-k}\sum_i (n_i-1)S_i^2$ is the pooled estimate for the variance.

The test statistic has approximately a χ^2_{k-1} distribution. Thus the null hypothesis is rejected if $X^2 > \chi^2_{k-1,\alpha}$ (where $\chi^2_{k-1}(\alpha)$ is the upper tail critical value for the χ^2_{k-1} distribution).

17.2 Brown-Forsythe Test

In statistics, when a usual one-way ANOVA is performed, it is assumed that the group variances are statistically equal. If this assumption is not valid, then the resulting F-test is invalid. The Brown-Forsythe test is a statistical test for the equality of group variances based on performing an ANOVA on a transformation of the response variable.

Suppose we have k samples of response data, where y_{ij} represents the value of i-th observation ($i = 1, 2, \ldots, n_)$ on the $j-$th factor level ($j = 1, 2, \ldots, k$). The hypotheses of Brown-Forsythe test can be expressed as:

$$H_0 : \sigma_1 = \sigma_2 = \ldots = \sigma_k$$

$$H_1 : \sigma_p \neq \sigma_q, \text{ for at least one pair } (p, q),\ 1 \leq p,\ q \leq k.$$

Transformation

The transformed response variable is constructed to measure the spread in each group. Define z_{ij} as the following

$$z_{ij} = |y_{ij} - \tilde{y}_j|,$$

where $\tilde{y}_j$ is the median of group j. In order to correct for the artificial zeros that come about with odd numbers of observations in a group, any z_{ij} that equals zero is replaced by the next smallest z_{ij} in group j. The Brown-Forsythe test statistic is the model F statistic from a one way ANOVA on z_{ij}:

$$F = \frac{(N-k)}{(p-1)} \frac{\sum_{j=1}^{k} n_j (\overline{z_j} - \overline{z})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (z_{ij} - \overline{z_j})^2},$$

where k is the number of groups, n_j is the number of observations in group j, and N is the total number of observations. If the variances are indeed heterogeneous, techniques that allow for this may be used instead of the usual ANOVA.

Under the null hypothesis of homogeneous variances, Brown-Forsythe statistic will have approximately an F distribution with $k-1$ and $N-k$ degrees of freedom. The test rejects the hypothesis that the variances are equal if

$$F > F_\alpha(k-1, N-k).$$

17.3 Levene's Test

Levene's test is an inferential statistic used to assess the equality of variance in different samples. Some common statistical procedures assume that variances of the populations from which different samples are drawn are equal. Levene's test assesses this assumption. It tests the null hypothesis that the population variances are equal. If the resulting $p-$value of Levene's test is less than some critical value α (typically .05), the obtained differences in sample variances are unlikely to have occurred based on random sampling. Thus, the null hypothesis of equal variances is rejected and it is concluded that there is a difference between the variances in the population.

Procedures which typically assume homogeneity of variance include analysis of variance and t-tests. Advantage of Levene's test is no requirement of normality assumption. Levene's test is often used before a comparison of means. When Levene's test is significant, modified procedures are used that do not assume equality of variance.

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2},$$

where

- W is the result of the test;
- k is the number of different groups to which the samples belong,
- N is the total number of samples,
- N_i is the number of samples in the $i-$th group,
- Y_{ij} is the value of the $j-$th sample from the $i-$th group,
- $Z_{ij} = |Y_{ij} - \overline{Y}_{i\cdot}|$ with $\overline{Y}_{i\cdot}$ the median of group i,
- $Z_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$ is the mean of all Z_{ij},
- $Z_{i\cdot} = \frac{1}{N_i} \sum_{j=1}^{N_i} Z_{ij}$ is the mean of the Z_{ij} for group i.

The significance of W is tested against $F_\alpha(k-1, N-k)$, where F is the F-test, $k-1$ and $N-k$ are the degrees of freedom and α is the chosen level of significance. The test rejects the hypothesis that the variances are equal if

$$F > F_\alpha(k-1, N-k).$$

Levene's test may also test a meaningful question in its own right if a researcher is interested in knowing whether population group variances are different.

Comparison with the Brown-Forsythe test

The Brown-Forsythe test uses the median instead of the mean. Although the optimal choice depends on the underlying distribution, the definition based on the median is recommended as the choice that provides good robustness against many types of non-normal data while retaining good statistical power. If one has knowledge of the underlying distribution of the data, this may indicate using one of the other choices. Brown and Forsythe performed Monte Carlo studies that indicated that using the trimmed mean performed best when the underlying data followed a Cauchy distribution (a heavy-tailed distribution) and the median performed best when the underlying data followed a Chi-square distribution with four degrees of freedom (a heavily skewed distribution). Using the mean provided the best power for symmetric, moderate-tailed, distributions.

Another modifications of Levene's test are presented in [18].

17.4 O'Brien Test

In the Obrien's test the data are transforming to

$$y_{ij} = \frac{(n_j - 1,5)n_j(x_{ij} - \overline{x_j}^2) - 0,5\sum_{j=1}^{n_j}(x_{ij} - \overline{x_i})}{(n_j - 1)(n_j - 2)}$$

and uses the F distribution performing an one-way ANOVA using y as the dependent variable.

17.5 Hartley's Test

In statistics, Hartley's test, also known as the F_{max} test, is used in the analysis of variance to verify that different groups have a similar variance, an assumption needed for other statistical tests. The requirement is that the samples have to be of an equal size.

The test involves computing the ratio of the largest group variance, $\max(s_j^2)$ to the smallest group variance, $\min(s_j^2)$.

$$F_{max} = \frac{\max(s_j^2)}{\min(s_j^2)}.$$

The resulting ratio, F_{max}, is then compared to a critical value from a table of the sampling distribution of F_{max}. If the computed ratio is less than the critical value($\nu = n_1 - 1$), the groups are assumed to have similar or equal variances.

Hartley's test assumes that data for each group are normally distributed. This test, although convenient, is quite sensitive to violations of the normality assumption. Alternatives to Hartley's test that are robust to violations of normality are O'Brien's procedure, and the Brown-Forsythe test.

17.6 Cochran's Test

Similarly, as in the case of Hartley's test, it also requires the frequencies in each group to be the same, i.e. $n_1 = \ldots = n_k$. Cochran's test statistic is

$$G_{max} = \frac{\max(s_j^2)}{\sum_{i=1}^{k} s_i^2}.$$

Large values of G_{max} leads to the rejection of the null hypothesis. The critical values are tabulated.

18 Monte-Carlo Method

Monte Carlo methods are a class of computational algorithms that rely on repeated random sampling to compute their results. Monte Carlo methods are often used when simulating physical and mathematical systems. Because of their reliance on repeated computation and random or pseudo-random numbers, Monte Carlo methods are most suited to calculation by a computer. Monte Carlo methods tend to be used when it is infeasible or impossible to compute an exact result with a deterministic algorithm.

Monte Carlo simulation methods are especially useful in studying systems with a large number of coupled degrees of freedom, such as fluids, disordered materials, strongly coupled solids, and cellular structures. More broadly, Monte Carlo methods are useful for modeling phenomena with significant uncertainty in inputs, such as the calculation of risk in business. These methods are also widely used in mathematics: a classic use is for the evaluation of definite integrals, particularly multidimensional integrals with complicated boundary conditions.

There is no single Monte Carlo method; instead, the term describes a large and widely-used class of approaches. However, these approaches tend to follow a particular pattern:

1. Define a domain of possible inputs.
2. Generate inputs randomly from the domain.
3. Perform a deterministic computation using the inputs.
4. Aggregate the results of the individual computations into the final result.

For example, the value of π can be approximated using a Monte Carlo method:

1. Draw a square on the ground, then inscribe a circle within it.
2. Uniformly scatter some objects of uniform size throughout the square. For example, grains of rice or sand.

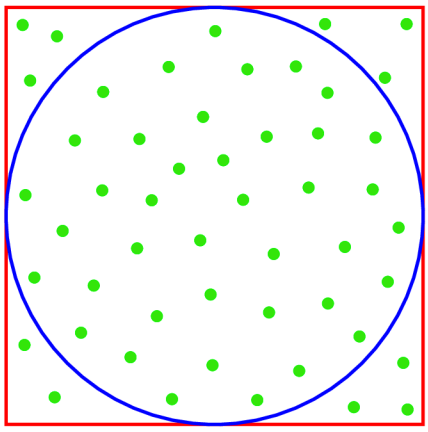


Figure 18.1: *To the computation of π by Monte-Carlo method.*

3. Count the number of objects in the circle, multiply by four, and divide by the total number of objects in the square.

4. The proportion of objects within the circle versus objects within the square will approximate $\pi/4$, which is the ratio of the circle's area to the square's area, thus giving an approximation to π.

Notice how the π approximation follows the general pattern of Monte Carlo algorithms. First, we define a domain of inputs: in this case, it's the square which circumscribes our circle. Next, we generate inputs randomly (scatter individual grains within the square), then perform a computation on each input (test whether it falls within the circle). At the end, we aggregate the results into our final result, the approximation of π. Note, also, two other common properties of Monte Carlo methods: the computation's reliance on good random numbers, and its slow convergence to a better approximation as more data points are sampled. If grains are purposefully dropped into only, for example, the center of the circle, they will not be uniformly distributed, and so our approximation will be poor. An approximation will also be poor if only a few grains are randomly dropped into the whole square. Thus, the approximation of π will become more accurate both as the grains are dropped more uniformly and as more are dropped.

In our thesis we used Monte-Carlo method for obtaining empirical $F-, G-, J-$ and $K-$functions.

For this section was used the material from the web site `www.wikipedia.com` and references therein. This section has only informative character and does not directly relate to this thesis.

19 Conclusion

The main aim of this thesis was to explore the use of statistical methods for the analysis of spatial distributions of particles in composite materials. The second aim was to provide better analyzes of patterns of finite sized events in finite regions. Moreover, the possible extensions and directions of further research are indicated.

The most common spatial patterns arising in composite materials can be classified into one of three main types. *Random*, when the particles' distribution do not obey to any specific requirement and, as a result, the particles or fibres will be randomly dispersed inside the matrix of the material. *Clustered*, when the particles tend to be grouped together forming several distinct aggregations. *Regular*, if the particles are distributed in a systematic way, especially when there is some sort of inhibition keeping them at a certain minimum distance from each other (the so-called *threshold distance).*

The self-contained file of methods for describing a random material is presented here. Spatial statistical techniques can be used to analyze patterns by detecting deviations from randomness. Complete spatial randomness (CSR) is considered to be the null hypothesis of the statistical tests and the interest lies in finding alternative types of patterns towards either clustering or regularity, from a random pattern. To find out this fact, Clark–Evans test and Skellam statistic were determined. In both cases, the CSR was rejected. This implies from the reality, that no two fibres cannot be nearer than the sum of their radii. In other words, penetration of fibres can not occur in the real situation.

In order to undertake this research, we started by acquainting ourselves with the features and characteristics of composite materials - the one of interest in this research.

Evidence shows that the way the particles or fibres, that are dispersed inside a matrix of material affect the materials' quality and performance. Therefore, one of the objectives of materials scientists is to find the particles' spatial distribution. The only feasible approach is to observe two dimensional cross-sections of the material. This is done by analyzing the pattern formed by the intercepted fibres and from the results obtained, infer the type of the distribution.

In the simulations, the three dimensional space was considered to be the parallelepiped with squared base which is intercepted at planes parallel to its edges and, as a consequence, the cross-sections were represented as squares. The bitmaps of such real samples were obtained from Klokner Institute of the Czech Technical Univerzity in Prague by Ing. Jan Zeman, Ph.D. The analyzes performed on the two dimensional patterns employed standard and recently developed statistical methods. They consist e.g. of the Kolmogorov-Smirnov test, quadrat and Clark–Evans tests as well as Monte Carlo tests using statistics based on Ripley's K function, nearest neighbor G function, empty space F and J functions. A detailed description of these standard tools, originally implemented to analyze and model spatial point patterns, was provided in the early chapters of this thesis.

From a review study we found, that of all the above functions, Ripley's K is the function whose statistic provides the most effective of all tests (especially the test based

on the square root transformation of K, i.e. L - function). This function was followed, in order of effectiveness, by the G and F (being G most effective against regularity and F against clustering) and the J functions. Note that the effectiveness is measured by the power of these functions at detecting deviations from randomness.

A problem encountered when estimating any of the above functions is, that the finite sample region is commonly assumed as being infinite. As a consequence, events that lay near the boundary of the region might have their nearest neighbor events outside the border and these will not be included in the analysis. It creates the so-called *edge effect problem* and causing bias in the functions' estimates.

Several methods have been introduced to correct the edge effects, see [9], [25] or [8]. Depending on the function to be estimated, authors seemed to have concentrated on different approaches to this problem. Evidently, it depends on the function to be estimated and also on the shape of the study region. Briefly, it can be viewed in the literature that the Doguwa-Upton or the Ohser- Stoyan estimators gave the least biased Ripley's K function's estimator and could also be applied to any shaped finite region. For the same reasons as for the choice of Ripley's K function estimator is recommend, the Floresroux-Stein estimator is suitable for estimating the G, F and J functions. In our analysis, the edge correctors were not used for the sake of simplicity and also from the reason, that it was not the main aim of this thesis to investigate them.

In literature, see e.g. [9] we can find many point processes that are very similar to ours, that represents the centers of the fibres, but they assume the equal sized circles, which is not true in the reality. For patterns formed from different sized events (in our case different sized circles), the best approach is to simulate several random patterns consisting of events whose sizes follow the same distribution as the real ones. The required functions are then calculated from each of those simulated patterns and their average is obtained. The functions thus achieved, will give the best approximation to the values of theoretical functions corresponding to random patterns, whose events have the same size distribution.

The main part of the thesis was devoted to the detailed description of developed algorithms **AI**–**AIV**. The new algorithms were created in such way to be similar to the real patterns as much as possible. The main mathematical properties of them was the theme of the second part.

The second part was devoted to the statistical computations and comparisons by mean of descriptive statistics. Moreover, the assumptions of homogeneity and normality were implemented. All the results are presented in tables, from which the difference of various algorithms is clear. The last part, named *Appendix* contains auxiliary techniques used in the second part.

20 Perspectives

Now, we briefly outline the possibilities for the next research. A very interesting and challenging topic is the deeper analysis of the algorithms generating random structures as the real one. It is conditioned by having at disposal real samples, resp. bitmaps of them. It would be suitable to include to the algorithms such parameters, which will be able in some ranges to influence the final structure to the reason of the best fitting of the real samples. Next challenging improvement can be made by using edge correctors for the computations of F, G, J and K functions and their better comparison in order to the refitting of algorithms generating structures similar to the real ones as much as possible. Such obtained structures can be then used as a base for computations of the e.g. equations of mathematical physic and their subsequent use to the general domain.

I'm nothing special,
in fact I'm a bit of a bore.

If I tell a joke,
you've probably heard it before.

But I have a talent,
a wonderful thing,

'cause everyone listens,
when I start to sing,

I'm so grateful and proud.
All I want is to sing it out loud...

Nejsem ničím zvláštní,
vlastně jsem spíš nudná.

Když řeknu vtip,
zřejmě jste ho už slyšeli.

Ale mám nadání –
úžasnou věc,

protože všichni zpozorní,
jakmile začnu zpívat.

Jsem tak vděčná a hrdá
a chci to vyzpívat nahlas...

...ABBA–never dying and unequalled texts in their songs. Their songs wrote the whole life. That is a pity, for their ending in 1981. The three parts of this thesis are commented by ABBA's songs. The chosen citations are according to the author´s best consideration leading to the best characterization of the parts.

The AƁBA pop-group in the seventieth.

When all is said and done...

Bibliography

[1] Abramowitz, M., Stegun, I.A.: *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, New York, 1964.

[2] Anděl, J.: *Statistické metody.* Matfyzpress, Praha, 2005.

[3] Anděl, J.: *Základy matematické statistiky.* Matfyzpress, Praha, 2005.

[4] Anderson, T.W., Darling, D.A.: Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Annals of Mathematical Statistics*, **23**:193–212, 1952.

[5] Baddeley, A.J., Kerscher, M., Schladitz, K., Scott, B.T.: Estimating the J-function without edge corrections.

[6] Chen, J.S., Moon, Y.S.: Fingerprint Matching with Minutiae Quality Score *Lecture Notes in Computer Science*, Volume 4642, 2007.

[7] Cliff, A.D., Ord, J.K.: *Spatial autocorrelation.* Pion, London, 1973.

[8] Cressie, N.A.C.: *Statistics for Spatial Data.* John Wiley & Sons, New York, 1993.

[9] Diggle, P.J.: *Statistical Analysis of Spatial Point Patterns.* Oxford University Press Inc., New York, 2003.

[10] Dixon, P:M.: Ripley's K-function. *Encyklopedia of Environmetrics*, **3**:1796–1803, 2002.

[11] Fitzgibbon, A., Pilu, M., Fisher, R.B.: Direct least square fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**(5):476–480, 1999.

[12] Gajdošík, J.: Quantitative analysis of fiber composite microstructure. Master's thesis, Czech Technical University in Prague, 2004.

[13] Ghanem, R.G., Spanos, P.D.: *Stochastic Finite Elements: A Spectral Approach.* Springer-Verlag, New York, 1991.

[14] Hátle, J., Likeš, J.: *Základy počtu pravděpodobnosti a matematické statistiky.* SNTL, ALFA, Praha, 1974.

[15] Hebák, P., Hustopecký, J. & kol.: *Vícerozměrné statistické metody [1], [2], [3].* INFORMATORIUM, Praha, 2005.

[16] Hendl, J: *Přehled statistických metod zpracování dat.* Portál, Praha, 2004.

[17] Holden, H., Øksendal, B., Ubøe, J., Zhang, T.: *Stochastic Partial Differential Equations. A modelling, White Noise Functional Approach.* Birkhäuser Boston, Springer-Verlag New York, New York, 1996.

[18] Islam, K.: *Transformed Tests for Homogeneity of Variances and Means.* PhD thesis, Graduate College of Bowling Green State University, Ohio, 2006.

[19] Kloeden, P.E., Platen, E., Schurz, H.: *Numerical Solution of SDE Through Computer Experiments.* Springer-Verlag, Berlin, 1997.

[20] Mansilla, D. Trias: *Analysis and Simulation of Transverse Random Fracture of Long Fibre Reinforced Composites.* PhD thesis, Universitat de Girona, 2005.

[21] Meloun, M., Militký, J.: *Statistická analýza experimentálních dat.* Academia, Praha, 2004.

[22] Meloun, M., Militký, J., Hill, M.: *Počítačová analýza vícerozměrných dat v příkladech.* Academia, Praha, 2005.

[23] Mikosch, T.: *Elementary Stochastic Calculus.* World Scientific Publishing Co. Pte. Ltd., Singapore, 1998.

[24] Militký, J.: *Prostorová statistika v matlabu.* Sborník příspěvků z 8. konference, MATLAB 2000, HUMUSOFT Praha.

[25] Ohser, S., Mücklich, F.: *Statistical Analysis of Microstructures in Material Science.* John Wiley & Sons, Chichester, 2000.

[26] Okabe, A., Boots, B., Sugihara, K., Chiu, S.N.: *Spatial Tesselations.* John Wiley & Sons, Chichester, 1999.

[27] Øksendal, B.: *Stochastic Differential Equations.* Springer-Verlag Berlin Heidelberg New York, Germany, 2000.

[28] Pospíšil, T.: Mathematical modelling of composite material with random structure. Master's thesis, Brno University of Technology, Dept. of Mathematics, 2003.

[29] Rektorys, K. & kol.: *Přehled užité matematiky I, II.* Prometheus, Praha, 2000.

[30] Ripley, B.D.: *Spatial Statistics.* John Wiley & Sons, New Jersey, 2004.

[31] Ryan, T.A., Joiner, B.L.: Normal probability plots and tests for normality, 1976.

[32] Sofia Mucharreira de Azeredo Lopes: *Statistical Analysis of Particle Distributions in Composite Materials.* PhD thesis, University of Sheffield, 2000.

[33] Škrášek, J., Tichý, Z.: *Základy aplikované matematiky I,II,III.* SNTL, Praha, 1990.

[34] Smith, T.E.: *Notebook On Spatial Data Analysis.* `http://www.seas.upenn.edu/~ese502/`.

[35] Statistics Toolbox for use with Matlab. User's Guide, 2001.

[36] Štecha, J.: *Optimalizační rozhodování a řízení.* Skriptum FEL ČVUT, Praha, 2000.

[37] Torquato, S.: *Random Heterogeneous Materials.* Springer-Verlag, New-York, 2002.

[38] Zeman, J.: *Analysis of Composite Materials with Random Microstructure.* PhD thesis, Czech Technical University in Prague, 2003.

[39] Zeman, J.: *Analysis of mechanical properties of fiber-reinforced composites with random microstructure.* Master's thesis, Czech Technical University in Prague, 1999.

List of Author's Publications

- Pospíšil T.: *Matematické modelování kompozitního materiálu s náhodnou strukturou,* Práce SVOČ 2003, ÚM FSI VUT v Brně, 36 stran.
- Pospíšil T.: *Mathematical modeling of composite materials,* 2nd International Workshop 2003, FAST VUT v Brně.
- Pospíšil T.: *Modelování pomocí stochastických diferenciálních rovnic,* Sborník z 13. semináře Moderní matematické metody v inženýrství, VŠB-TU Ostrava 2004, str. 180-183.
- Pospíšil T.: *Mathematical Modelling of Composite Materials,* Proceedings of the International Interdisciplinary Honeywell EMI 2005.
- Franců J., Nechvátal L., Pospíšil T.: *Mathematical Problems of Modelling of Real Composite Materials,* SEMINÁŘ Z APLIKOVANÉ MATEMATIKY U příležitosti 100. výročí narození prof. Františka Vyčichlo, Stavební fakulta ČVUT Praha 2005.
- Pospíšil T.: *Generating of Random Structures for Mathematical Modelling of Composite Materials,* 5th International Workshop 2006, FAST VUT v Brně.
- Pospíšil T.: *On Mathematical Modelling of Composites,* 6th International Workshop 2007, FAST VUT v Brně.
- Pospíšil T.: *On Mathematical Modeling of Composite Material with Random Structure,* preprint.
- Pospíšil T.: *On Statistical Description of Random Structures,* (accepted for publication in Mechanical Engineering).
- Pospíšil T.: *Generating Non-Periodic Microstructures of Fiber Composites,* accepted for publication in Mechanical Engineering).

Others

- 2005 – Successful research worker of the doctoral grant project BUT FME 2005, No. BD 135 3004 called *"Simulation of Random Structures of Composites"* (2^{nd} overall position in the section *Applied Sciences* in the faculty)
- 2003 – Dean´s Award
- 2003 – Josef Hlávka Prize

Brno, 29th August 2010 Ing. Tomáš Pospíšil