ON STATISTICAL DESCRIPTION OF RANDOM STRUCTURES

Tomáš Pospíšil*

Mathematical modeling of fibre composite materials is very difficult because of their random values of the coefficient describing mechanical properties of their separate phases. For the computational reasons, the real materials, i.e. materials with non-periodic structure are replaced by 'equivalent' structures having almost the same mechanical properties. To the implementation of this, the various algorithms were developed for generating an 'equivalent' structures, which will be similar to the real one as much as possible. Therefore some simple methodology for a statistical comparing of different structures developed by different algorithms is needed.

Keywords: *non-periodic structures, spatial randomness*

1. Introduction

The use of the homogenization theory in a mathematical modeling of composites, which is suitable for numerical computations, see e.g. [1], [2] or [6], assumes a periodic structure of the considered medium, which is not often true in reality. Therefore the algorithms for generating similar structures are developed. To the comparing the results obtained by different methods we need some simple tool, which will be able to intercept deviations from the statistical viewpoint. In other words, we introduce methods for statistical comparing of random structures.

The content of the paper is as follows. We start with a characterization of two-phase fiber composite materials together with its brief description. In Section 3 we give a statistical description of the structures including descriptive statistics and spatial randomness. The next three sections are devoted to the separate branches of the comparisons: the quadrat tests of randomness, second order methods and distance methods. In the last section we arrive to the conclusion.

2. Random structures

The essential requisition for developing a methodology for comparing structures of fiber composite material is to keep real samples(i.e. photos) of such material at disposition. We used the data obtained from the Czech Technical University in Prague, Klokner Institute, Department of Engineering Mechanics. Those are bitmaps of a dimension 1144×1144. From the image analysis of these data, we chose normal distribution of the fibre diameters with expected value of the diameter 71.87 and standard deviation 4.58. The average amount of the fibres in selected rectangular area was 164.60 and the average volume fraction 48.69 %.

* T. Pospíšil, Department of Mathematics, Brno University of Technology, Technická 2896/2, Brno 616 69, Czech Republic

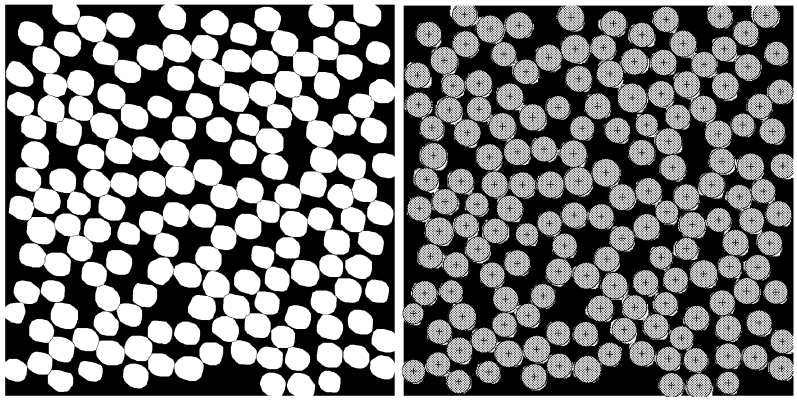

Fig.1: Real sample (left) and its correction for computation (right)

For further information about the samples, see [7]. Next figures represent an example of the obtained data and corrected ones used to the experiments.

Now, having the set of bitmaps of the real or simulated composite media at disposal, it is suitable to compare it.

3. Statistical description

In this section we introduce some statistical indicators used to the correct description of differences between separate instances of a composite media. These instances can be taken from the real material or they can be simulated computationally.

3.1. Descriptive statistics

In order to use methods of descriptive statistics, we have to create some kind of a set of parameters for each sample, which we use for next computations. The simplest way is to divide each sample by a regular $n \times n$ abstract rectangular grid. In our samples we choose $n = 10$, i.e. $10{\times}10$ grid. Then, in every cell c_i we compute an elementary volume fraction f_i (in percentages) and this obtained set of elementary volume fractions $\{f_i\}_{i=1}^{n\times n}$ serves us as a base for next computations. It is also very important to choose an optimal ratio between the size of n and average diameter of fibres. It is clear, that choosing n to be very large is meaningless, because many of f_i should have the value zero and some of them one. On the other hand, choosing n to be very small causes not taking into account the randomness

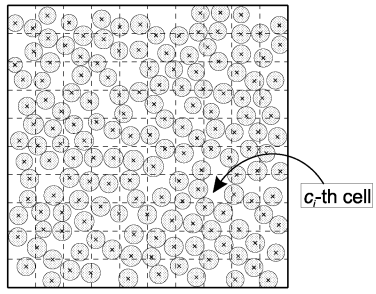

Fig.2: Dividing the sample by an abstract regular grid

of the distribution of the fibres in the matrix. Denote by the symbol f_i^j, $i = 1 \dots n \times n$, $j = 1 \dots 15$, the elementary volume fraction in the $i-$th cell of a realization j. The results of descriptive statistics for f_i^j are presented in Table 1.

	Mean	**Median**	**Min.**	**Max.**	**Range**	**Var.**	**Std. dev.**	**Kurt.**	**Skew.**
No. 1	51.63	53.35	17.38	92.39	75.01	215.32	14.67	2.97	−0.04
No. 2	51.13	54.21	0.00	86.84	86.84	295.23	17.18	2.99	−0.49
No. 3	47.93	48.53	0.28	86.17	85.90	326.10	18.06	2.85	−0.34
No. 4	44.32	46.51	0.00	74.44	74.44	277.27	16.65	2.91	−0.55
No. 5	53.79	54.44	15.86	93.25	77.39	242.45	15.57	2.60	0.05
No. 6	52.48	54.16	0.00	81.10	81.10	240.03	15.49	3.51	−0.62
No. 7	42.76	44.50	0.00	78.95	78.95	364.05	19.08	2.72	−0.55
No. 8	49.66	52.33	14.79	80.69	65.90	227.93	15.10	2.10	−0.09
No. 9	44.74	45.71	0.00	73.06	73.06	226.27	15.04	2.93	−0.37
No. 10	42.47	41.95	0.00	78.55	78.55	295.63	17.19	2.78	−0.09
No. 11	52.48	53.91	20.09	80.31	60.22	200.27	14.15	2.43	−0.32
No. 12	50.20	52.14	0.00	82.89	82.89	350.25	18.72	3.08	−0.70
No. 13	52.43	51.26	5.76	89.63	83.87	296.10	17.21	2.84	−0.24
No. 14	46.40	46.18	11.48	92.29	80.81	229.73	15.16	2.99	0.19
No. 15	47.90	47.36	0.00	86.14	86.14	275.41	16.60	2.81	−0.15
Average	48.69	49.77	5.71	83.78	78.07	270.80	16.39	2.83	−0.29

Tab.1: Computed values of descriptive statistics of all volume fractions for all samples and their averages

Next, we present descriptive statistics for the amount of fibres in the real samples, see the table 2.

	Mean	**Median**	**Min.**	**Max.**	**Range**	**Var.**	**Std. dev.**	**Kurt.**	**Skew.**
Real	164.60	164	145	189	44	167.40	12.94	2.04	0.14

Tab.2: Computed values of descriptive statistics for a total amount of fibres for all samples

3.2. Spatial randomness

The *complete spatial randomness (CSR)*, see e.g. [4] or [3] for its definition, is of limited scientific interest in itself in the theory of composites. The reason is due to the real physical aspects(e.g. an impossibility of overlapping of the particular fibres), see e.g. [4]. But on the other hand there are several good reasons why we might begin an analysis with a test of the 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 the CSR. Of course, CSR operates as a dividing hypothesis between regular and clustered (aggregated) patterns, see [4].

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 the informative descriptives statistics that can help elucidate models that might be used for the simulating of the real structures.

Randomness tests of CSR are commonly based on the following three branches of the 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.

4. 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 of fibers) 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 a quadrat method. The comparison will be as follows: For each sample we compute Pearson's test statistic

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

where m is the total number of fibres (centers) in the sample, n_i denotes the number of centers in a cell c_i and $\bar{n}$ is the mean of n_i. In our cases we chose $n = 10$, i.e. the 10×10 grid (an assumption $n^2 > 6$ should be fulfilled, see [11] for the explanation). The results are in Table 4. It holds, under CSR, the Pearson's test statistic has χ^2- distribution with $f = n^2 - 1 = 10^2 - 1 = 99$ degrees of freedom, see e.g. [3].

No.	**Pears**	**No.**	**Pears.**	**No.**	**Pears.**
1	35.45	**6**	26.96	**11**	23.54
2	34.69	**7**	35.53	**12**	34.35
3	34.20	**8**	32.38	**13**	36.75
4	39.97	**9**	36.89	**14**	30.39
5	26.70	**10**	38.95	**15**	39.09

Tab.3: Values of the Pearson's statistics Q

If the value for Q is less than the $100\,\alpha/2$ percentile of the χ^2 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 [8], 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 for 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 $\chi^2_{99}(0.975) = 73.36$ and $\chi^2_{99}(0.025) = 128.42$. Since in our case, all values of Pearson's test statistic Q are smaller than 73.36, it indicates significant departure from the CSR.

5. 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.

5.1. Tests based on Ripley's K function

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

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

where λ is the intensity (number of fibres per unit area) of events, see e.g. [11]. So, $K_\lambda(t)$ describes characteristics of the point process at many distances scales.

$K_\lambda(t)$ does not uniquely define the point process in the sense that the two different processes can have the same $K_\lambda(t)$ function. Also, processes with the same $K_\lambda(t)$ function can be different.

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

$$K_\lambda(t) = \pi\, t^2 \ .$$

Values of $K_\lambda(t)$ for a process are often compared with those for the Poisson process. Values larger or smaller than $\pi\, t^2$ respectively indicate a more clustered or more regular process than the Poisson process. In [5] are presented $K_\lambda(t)$ functions for various types of process in details.

5.1.1. Estimating $K_\lambda(t)$

Given the locations of all events within a defined study area, $K_\lambda(t)$ 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. [8], [4], [3] or [9].

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

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

CSR is then $L(t) = t$. $\hat{K}(t)$ is an estimation of $K_\lambda(t)$. 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. 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 [4] (Lotwick & Silverman) as

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

where

$$a_1(t) = \frac{0.21\,P\,t^3 + 1.3\,t^4}{|A|^2}\ , \qquad a_2(t) = \frac{0.24\,P\,t^5 + 2.62\,t^6}{|A|^3}\ ,$$

$$b(t) = \frac{\pi\,t^2}{|A|}\left(1 - \frac{\pi\,t^2}{|A|}\right) + \frac{1.0716\,P\,t^3 + 2.2375\,t^4}{|A|^2}\ ,$$

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. As suggested in [4], $\pm 2\sqrt{var_{LS}(t)}$ 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 [4] 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. Whether $\hat{D}(t)$ is smaller than the lower bound, the pattern tends to regularity; if $\hat{D}(t)$ is bigger than the upper bound, the pattern tends to cluster; otherwise, the CSR assumption becomes applicable.

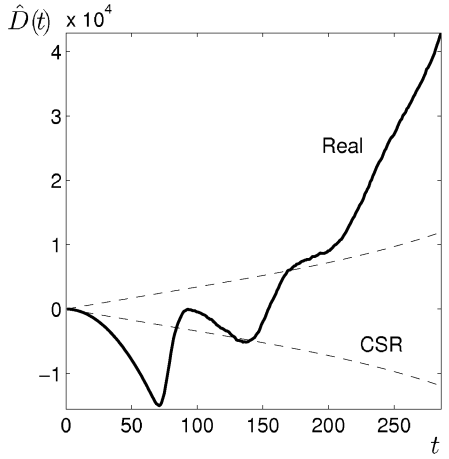

Fig.3 Comparison of D-functions

6. Distance methods

Distance methods, also known as plotless sampling techniques, were introduced because of the practical difficulties caused by quadrat sampling sometimes. 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.

6.1. Skellam's Statistic

To make ideas of nearest-neighbor distances precise, we have to determine the probability distribution of a nearest neighbor distance under CSR and compare the observed nearest neighbor distances 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 nearest neighbor

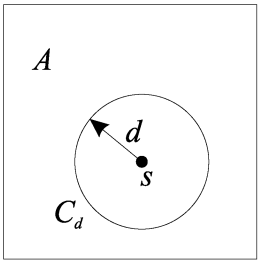

Fig.4: Cell of radius d

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 4. Then, according to the picture, the probability that D is at least equal to d is precisely the probability that there are no other points in C_d. Hence, it can be shown, see [10], that this probability is given by

$$\mathbf{P}(D > d) = \mathrm{e}^{-\lambda\,\pi\,d^2} \tag{1}$$

and that's why we finally obtain the distribution function of D

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

As we can see, this is an instance of the *Rayleigh distribution*. Next, for a random sample of m nearest-neighbor distances $(D_1, \ldots, D_m)$ from this distribution, the scaled sum (Skellam's statistics)

$$S_d = 2\,\lambda\,\pi\,\sum_{i=1}^{m} D_i^2 \tag{3}$$

is χ^2 *distributed* with $2\,n$ degrees of freedom, see [10]. So, finally, this statistic provides a test of the CSR hypothesis based on nearest neighbors. If we choose a significant level $\alpha = 0.05$ and approximately $n = 165$ 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 exceptions are samples 4, 7 and 10.

No. 1	445.96	**No. 6**	459.40	**No. 11**	496.99
No. 2	468.99	**No. 7**	364.61	**No. 12**	427.88
No. 3	390.72	**No. 8**	541.02	**No. 13**	523.05
No. 4	362.57	**No. 9**	489.81	**No. 14**	404.41
No. 5	481.53	**No. 10**	366.78	**No. 15**	435.36

Tab.4: Extremes of the Skellam's statistic for all samples

6.2. Clark-Evans test

The Clark-Evans test, see e.g. [10], 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. It was introduced by Clark and Evans (1954). These authors stated that the distance from a point to its nearest neighbor, irrespective of a direction, provides the basis for a measure of spacing.

Let us denote for a random sample set of independent nearest-neighbor distances $\{D_1, \ldots, D_m\}$ (more information about choosing an independent set of samples, see e.g. [3], [4], [9] or [10]). It follows from the *central limit theorem*, that independent sums of identically distributed random variables are approximately *normally distributed*. Hence, the most common test of the CSR hypothesis based on nearest neighbors involves a normal approximation to the sample mean of D_i, $i = 1 \ldots m$, as defined by

$$\bar{D}_m = \frac{1}{m} \sum_{i=1}^{m} D_i \ . \tag{4}$$

It can be shown, see [10], that mean and variance of this distribution are given respectively by

$$\mathbf{E}\left[D\right] = \frac{1}{2\sqrt{\lambda}} \ , \qquad \mathbf{D}\left[D\right] = \frac{4-\pi}{4\,\lambda\,\pi} \ . \tag{5}$$

Next we observe from the properties of *iid* random samples that for the sample mean $\overline{D}_m$ in (4) it holds

$$\mathbf{E}\left[\bar{D}_m\right] = \frac{1}{m} \sum_{i=1}^{m} \mathbf{E}\left[D_i\right] = \frac{1}{m} \left(m\,\mathbf{E}\left[D_1\right]\right) = \mathbf{E}\left[D_1\right] = \frac{1}{2\sqrt{\lambda}} \tag{6}$$

and similarly

$$\mathbf{D}\left[\bar{D}_m\right] = \left(\frac{1}{m}\right)^2 \sum_{i=1}^{m} \mathbf{D}\left[D_i\right] = \frac{1}{m^2} \left(m\,\mathbf{D}\left[D_1\right]\right) = \frac{4-\pi}{m\,(4\,\lambda\,\pi)} \ . \tag{7}$$

From the *central limit theorem* we obtain

$$\bar{D}_m \sim \mathrm{N}\left(\frac{1}{2\sqrt{\lambda}}, \frac{4-\pi}{4\,\lambda\,\pi\,m}\right) \tag{8}$$

and after standardization we can write

$$Z_m = \frac{\bar{D}_m - \mathbf{E}\left[\bar{D}_m\right]}{\sqrt{\mathbf{D}\left[\bar{D}_m\right]}} \sim \mathrm{N}(0,1) \ , \tag{9}$$

so Z_m has *standardized normal distribution*.

In the following table we can see the extremes of the Z-means of all samples.

	Minimum	**Maximum**
Real	10.316	14.067

Tab.5: Extremes of the mean values obtained by Monte-Carlo simulation of the Clark-Evans test

Similarly, as in the case of the Skellam's statistic, 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.

6.3. Conclusion

In this contribution we came from the well-known result that replacing random structure of two-phase fiber composite material leads to incorrect results on mathematical modeling. This fact can be e.g. demonstrated on a problem of a torsion of a bar that can be modeled by an elliptic PDE $-\operatorname{div}(a\,\nabla u) = f$ with a random coefficient a and Dirichlet boundary condition.

To obtain more accurate results using computational methods are being developed various algorithms generating structures similar to the real ones. Therefore, we summarized a simple collection of methods, which can be used for a statistical comparison of separate samples(real or simulated). The real samples(bitmaps) were obtained from the Czech Technical University in Prague, Klokner Institute. From such obtained samples the basic statistical descriptors were computed.

Very important question is about the complete spatial randomness(CSR). To find out this fact, Clark-Evans test and Skellam's 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. Also, quadrat test of randomness and distance methods were discussed.

Acknowledgments:

The paper was supported by research project from MŠMT of the Czech Republic No. 1M06047 'Center for Quality and Reliability of Production', research plan from MŠMT of the Czech Republic No. MSM0021630519 'Progressive reliable and durable structures', by grant from Grant Agency of the Czech Republic (Czech Science Foundation) Reg. No. 103/08/1658 'Advanced optimum design of composed concrete structures', and by grant Reg. No. 201/08/0874 from Grant Agency of the Czech Republic (Czech Science Foundation)

References

[1] Bensoussans A., Lions J.L., Papanicolaou G.: Asymptotic Analysis for Periodic Structures, North-Holland, Amsterdam (1978)
[2] Bourgat J.F., Lanchon H.: Application of the homogenization method to composite materials with periodic structure, Raport de Recherche No. 208, (1976), IRRIA, Paris
[3] Cressie N.A.C.: Statistics for Spatial Data, John Wiley & Sons, New York, (1993)
[4] Diggle P.J.: Statistical Analysis of Spatial Point Patterns Oxford University Press Inc., New York, (2003)
[5] Dixon P.M.: Ripley's K-function, Encyklopedia of Environmetrics 3 (2002), pp. 1796–1803
[6] Franců J.: Homogenization (in Czech) Proceedings of 6th seminar from P.D.E., Manětín 1981, JČSMF (1982), pp. 21–66
[7] Gajdošík J.: Quantitative Analysis of Fiber Composite Microstructure, Master Thesis, Czech Technical University in Prague, (2004)
[8] Sofia Mucharreira de Azeredo Lopes: Statistical Analysis of Particle Distributions in Composite Materials, Doctoral Thesis, University of Sheffield, (2000)

[9] Ripley B.D.: Spatial Statistics, John Wiley & Sons, New Jersey, (2004)
[10] Smith T.E.: Notebook On Spatial Data Analysis, http://www.seas.upenn.edu/∼ese502/
[11] Torquato S.: Random Heterogeneous Materials, Microstructure and Microscopic Properties, Springer-Verlag, (2002)

Received in editor's office: April 1, 2010
Approved for publishing: August 30, 2010

Note: This paper is an extended version of the contribution presented at the international conference *STOPTIMA 2007* in Brno.