PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **22** (2013), 195–206

PLISKA
STUDIA MATHEMATICA
BULGARICA

FRACTAL ANALYSIS FOR CANCER RESEARCH: CASE STUDY AND SIMULATION OF FRACTALS

Milan Stehlík*, Fabian Wartner
Mária Minárová†

Abstract. This paper discusses the possibilities of application of fractal geometry for cancer research. Fractal geometry is a new tool that can be extremely useful for many problems in almost every scientific field. The studies recently done in medicine show fractals can be applied for cancer detection and the description of pathological architecture of tumors. This fact is not surprising, as due to the irregular structure, cancerous cells can be interpreted as fractals. Cancer diagnosis can be done via determination of fractal dimension. A likelihood ratio test for the Hausdorff dimension is employed in [7]. We empirically checked the obtained tests on Sierpinski Carpet and on cancer data. However, several issues arisen, especially those related to simulation of fractals which may mimic tissues. These are discussed in the present paper.

1. Introduction. The term fractal came from the Latin adjective fractus meaning fragmented or broken and was originally given to highly irregular sets by the French mathematician Benoit Mandelbrot in 1975 [4]. He characterized fractal as "a rough or fragmented geometric shape that can be split into parts, each

*First author acknowledges partial support of Aktion Project 58p5 and WTZ Project Nr. IN 11/2011. The authors are thankful for the support of the Editor and constructive comments of the anonymous Referee.

†Third author acknowledges support of APVV-0351-07 and VEGA 1/0269/09

2010 *Mathematics Subject Classification:* 65D18.

Key words: cancer research, Fractal, genetic algorithm, simulation of fractal.

of which is (at least approximately) a reduced-size copy of the whole". Mandelbrot introduced fractals as a new class of mathematical objects which represent nature. Many areas of science, such as physics, chemistry and biology, soon recognized how powerful his ideas were. In mathematics, a new area, namely fractal geometry, came up and developed very quickly on the base of geometric measure theory, harmonic analysis, dynamical systems and ergodic theory. Fractal geometry can be viewed as an extension of the classical Euclidean geometry. Fractal dimension is a number associated with a fractal that can be used to compare one fractal to another. It shows how densely the fractal occupies the metric space in which it is situated. Fractal dimensions are of great importance, because they can be attached to real-world data as well. Every fractal, including fractals found in nature, has its own dimension and this dimension can be measured by means of experiments. There are many ways of defining fractal dimension. Such terms, as the Hausdorff, the Sandbox, the box-counting and the packing dimensions are widely used. Hausdorff dimension plays an important role in developing general mathematical theory. On the other hand, the box-counting dimension can be easily implemented in practice, that is why it is so widely used. Box-counting dimension is probably the simplest dimension in use.

[1] discussed planar tissue preparations in mice which has a remarkably consistent scaling exponents (fractal dimensions) for tumor vasculature even among tumor lines that have quite different vascular densities and growth characteristics. An extensive study of cancer risk assessment on simulated and real data and fractal based cancer is given in [7]. Both non-random and random carpets are modelling the cancer growth. On the other hand, in previous investigations, it has been shown that the texture of mammary tissue, as seen at low magnification, may be characterized quantitatively in terms of stereology (see [5] and references therein). In [6], the images of the mammary cases were reexamined (20 cases of mastopathy and 20 cases of mammary cancer, each with 10 images). In [8] a construction of a statistical test is given, which enables to distinguish between the two groups.

The paper is organized as follows. Case study discriminating the mammary tissues motivates fractal analysis for cancer research in section 2. In section 3 we illustrate issues which we met by simulation of fractals. Such simulations are useful for validation of statistical procedures on complex sets.

2. Case study. Data has been taken from [6] and are collected from histological examination whose result was a grey level image with a resolution of 512x512 pixels. These 400 histological images of mammary tissues were transmitted to black-and-white binary images with a black-and-white CCD camera.

These binary images consisted two phases only. We have used ImageJ software (see Appendix) to determine the box-counting dimension of binary images.

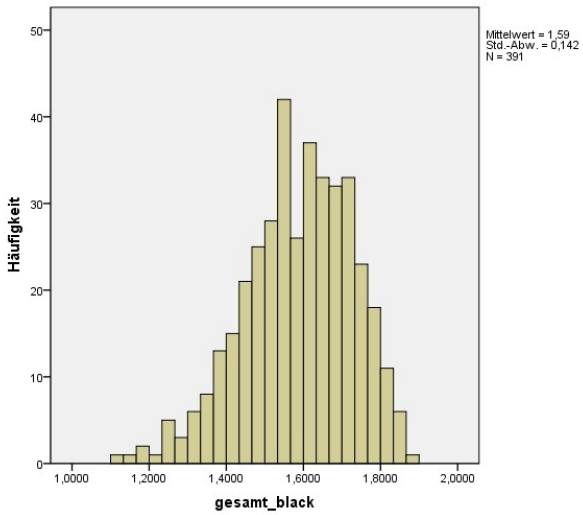

Fig. 1 Histogram of empirical box-counting dimensions

2.1. Goodness of fit. Naturally, the box-counting dimensions of these 391 binary images are between 1 and 2. The mean is 1.590 and standard deviation is 0.142. The lowest dimension is 1.104 and the highest dimension is 1.872. The histogram plotted in Figure 1 shows a skewed to the left distribution. To fit a distribution to the data it was assumed that it could be a gamma distribution or a Weibull distribution. Estimation of Gamma distribution gives shape parameter $\hat{\lambda} = 125.379$ and scale $\hat{\beta} = 78.98$. For Weibull distribution we obtained (by method of moments) shape parameter $\hat{\lambda} = 13.68$ and scale $\hat{\beta} = 1.648$.

To test goodness of fit we applied a Kolmogorov-Smirnov test. For testing the statistical software *R* was used, more precisely the procedure "ks.test" with the estimated parameter values above. Based on Kolmogorov-Smirnov test the gamma distribution is not acceptable for the fit. The Kolmogorov-Smirnov test for Weibull distribution shows a test statistic d=0.0489 and a corresponding p-value $p_V = 0.3069$. For a given $\alpha = 0.05$ the null hypothesis cannot be rejected, which means that data may be considered to follow a Weibull distribution with parameter values given above. Same results show the QQ-plots for gamma and Weibull distribution, plotted on Figure 2. The points in QQ-plot of Weibull distribution are nearly on a 45 degree line; in contrast the points of QQ-plot of gamma distribution show a slight curve and are above a 45 degree line. The

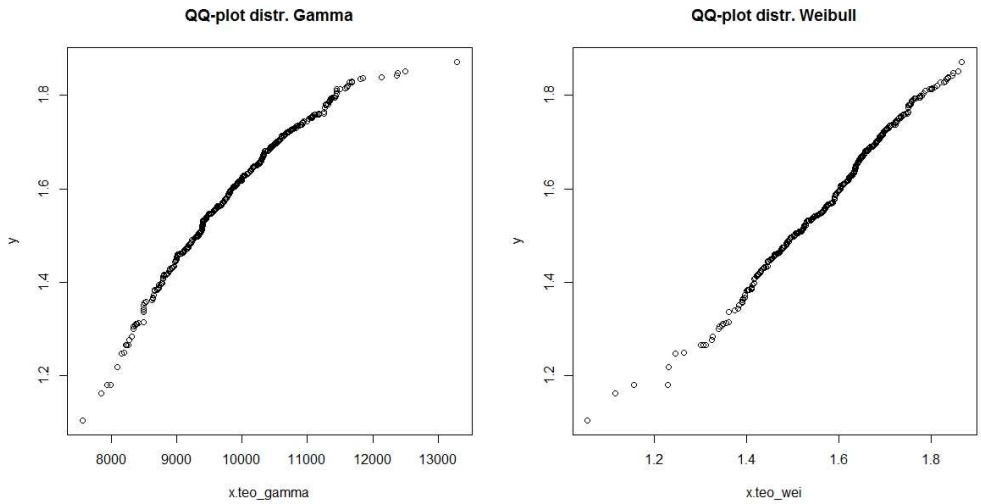

Fig. 2. QQ-plot for gamma and Weibull distribution

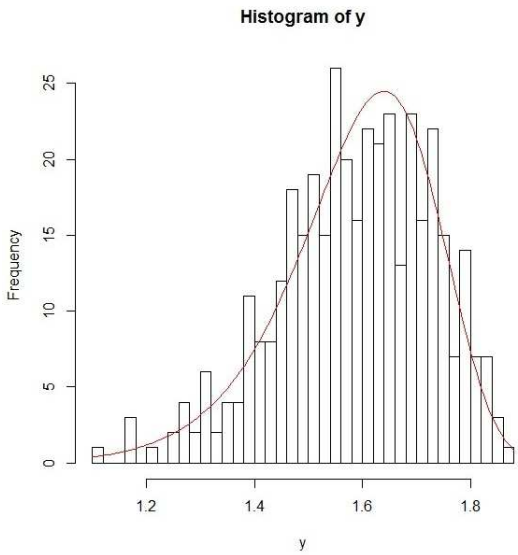

Fig. 3. Goodness of fit of Weibull distribution

histogram (plotted at Figure 3) demonstrates fitting of Weibull distribution with the underlying data, which is quite good.

Table 1. Lower percentiles of empirical distribution

Lower percentiles	Value	Percentage of cancer
10th percentile:	1.3977	74.36
9th percentile:	1.3868	74.29
8th percentile:	1.3824	70.97
7th percentile:	1.3656	70.37
6th percentile:	1.3504	78.26
5th percentile:	1.3139	73.68
4th percentile:	1.3072	86.67
3rd percentile:	1.2750	90.91
2nd percentile:	1.2486	85.71
1st percentile:	1.1791	100.00

Table 2. Upper percentiles of empirical distribution

Upper percentiles	Value	Percentage of cancer
90th percentile:	1.7598	10.00
91st percentile:	1.7744	8.33
92nd percentile:	1.7802	9.38
93rd percentile:	1.7897	10.71
94th percentile:	1.7934	12.50
95th percentile:	1.7982	15.00
96th percentile:	1.8135	6.25
97th percentile:	1.8173	8.33
98th percentile:	1.8294	0.00
99th percentile:	1.8424	0.00

2.2. Discriminating between tissues based on order statistic. Together with rank statistics, order statistics are among the most fundamental tools in non-parametric statistics and inference. For order statistic data is sorted ascending. Then important cases are sample percentiles. When using probability theory to analyze order statistic these percentiles from a continuous distribution are important. We have fitted empirical cumulative distribution function to the data via percentages in Excel. For Weibull distribution certain percentiles are used. Results of discrimination between masthophatic and cancer tissues are plotted in Tables 1–4.

In both cases the percentage of cancer is high in lower percentiles and low in higher percentiles. By employing a generic results observed in cancer research (see e.g. [1]), we may conclude that in our study for images with fractal dimension of 1.75 and higher, the risk of is higher than for images with fractal dimension 1.39 and lower. By the cancer risk we mean the quantitative assessment (by means

Table 3. Lower percentiles of Weibull distribution

Lower percentiles	Value	Percentage of cancer
10th percentile:	1.3987	75.00
9th percentile:	1.3875	74.29
8th percentile:	1.3750	72.41
7th percentile:	1.3611	76.00
6th percentile:	1.3454	77.27
5th percentile:	1.3271	73.68
4th percentile:	1.3051	84.62
3rd percentile:	1.2775	90.91
2nd percentile:	1.2397	100.00
1st percentile:	1.1780	100.00

Table 4. Upper percentiles of Weibull distribution

Upper percentiles	Value	Percentage of cancer
90th percentile:	1.7524	13.04
91st percentile:	1.7581	11.63
92nd percentile:	1.7643	8.11
93rd percentile:	1.7709	8.33
94th percentile:	1.7783	8.57
95th percentile:	1.7864	10.34
96th percentile:	1.7958	13.04
97th percentile:	1.8071	5.88
98th percentile:	1.8216	0.00
99th percentile:	1.8435	0.00

of fractal dimension analysis) of the degree of abnormality and aggressiveness of cancer tumors obtained through biopsy. The fractal dimension is an objective and reproducible measure of the complexity of the tissue architecture of the biopsy specimen. The higher the number, the more abnormal the tissue is.

3. On simulation of carpets. In this section we introduce the Hausdorff measure calculation both on fractals with random choice and on real tissues. First we gauge the Sierpinski gasket with edge division $r = 2$, i.e. compression ratio $=1/2$ with the random choice loaded. The diameter stipulation of such a kind of set R within the text is based on the genetic algorithm. Considering a metric as the Euclidean distance between two points of a set R, we have well defined diameter of the set.

It is easy to see that the diameter of Sierpinski gasket without random choice, see Figure 4, is equal to the length of the source triangle.

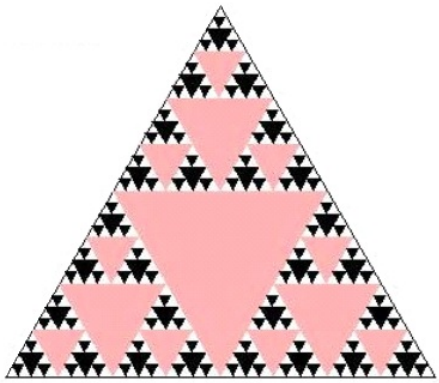

Fig. 4. Sierpinski gasket; the 5th ramification with 121 triangles

3.1. What is a genetic algorithm. The genetic algorithm, which is a self adaptive and global optimizing probability search algorithm and is inspired by evolution, was introduced and investigated by John Holland in [3] and by some of his students.

Let us denote F_i the Sierpinski gasket of the i-th ramification, F_0 being an equilateral triangle in the Euclidean plane with the edge length $l = 1$.

It is clear that $F_0 \supset F_1 \supset F_2 \supset \ldots \supset F_n \supset \ldots$ Moreover, the nonempty set $\cap F_i$ is the Sierpinski carpet of the Hausdorff dimension $\log 3/\log 2 \approx 1.58496$.

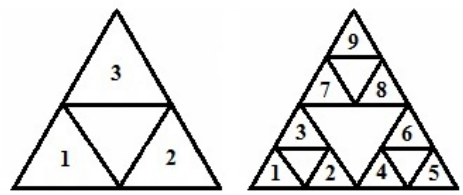

Fig. 5. Numbering of triangles within the ramification (counterclockwise); left) numbering of triangles in the first ramification, right) in the second ramification

The particular numbers are allocated to the triangles in the counterclockwise within the ramification, see Figure 5. Encoding and decoding of the position within the gasket of arbitrary ramification level is done by using so called position mappings. S_1, S_2, S_3, pointing to the position of the smaller gasket triangle within the greater one (of the previous iteration) can be written as follows:

$$S_1(x,y) = (x,y), S_2(x,y) = (x,y) + (l/2, 0), S_3(x,y) = (x,y) + (l/4, l\sqrt{3}/4).$$

So the mapping S_1 points to the lower left triangle, the mapping S_2 points to the lower right triangle, and the mapping S_3, points to the upper triangle of the particular ramification. Accordingly, the position chain of the small triangle in the gasket is given by the mapping composition, i.e. the triangle No. 8 on the right part of Figure 5 can be reached by the $S_3 \circ S_2$, triangle No. 4 by the $S_2 \circ S_1$, triangle No. 5 by the $S_2 \circ S_2$, etc. The number of the mappings in the composition is equal to the ramification iteration.

Encoding/decoding: Finally, every small triangle of the m-th ramification either is or is not chosen in each random choice. This fact is represented by assigning 1 for chosen and 0 for non-chosen triangle (see Figure 7 for example of a tissue in 0-1 representation). Consequently, the entire code of the gasket evolved to the m-th ramification level consists of 3^m digits $= 3^m$ zeros or ones, belonging to the each related triangle. We denoted such a set by R. After the decoding procedure the diameter of R is calculated by using the formula

$$H(R) = \min_{R \in F_m^R} diam R/\mu(R),$$

where F_m^R is a set having elements R of the gaskets of m-th ramification with a random choice applied, $\mu(R)$ is the density of the R gasket (numbers of ones and numbers of all digits ratio).

Table 5. Hausdorff measures and diameters for the Sierpinski gasket (2nd–13rd ramification) with random choices of probability p

r.	CPU[s]	HM, p=1/2	d(U), p=1/2	HM, p=3/4	d(U), p=3/4	HM, p=7/8	d(U), p=7/8
2	0.8593	0.80925	0.875	0.80925	0.875	0.80925	0.875
3	0.9218	0.90277	0.9375	0.81906	0.881671	0.90277	0.9375
4	1.1406	0.90515	0.939061	0.88994	0.905151	0.95093	0.96875
5	1.4531	0.95151	0.969128	0.95313	0.903363	0.92673	0.953125
6	2.9063	0.93043	0.955523	0.91531	0.9457	0.93880	0.960938
7	7.9687	0.96922	0.980469	0.96344	0.976773	0.95701	0.972656
8	25.406	0.9846	0.990257	0.96631	0.978609	0.972429	0.982515
9	83.515	0.98767	0.9922	0.98018	0.98745	0.96928	0.980505
10	265.51	0.98457	0.990235	0.97842	0.986328	0.98	0.987334
11	862.75	0.98380	0.989747	0.98766	0.992195	0.98958	0.99341
12	3269.2	0.99440	0.99646	0.99749	0.998415	0.98822	0.992554
13	37145	0.995458	0.997132	0.996231	0.996231	0.98939	0.993291

There were 3^{m+1} measurements, i.e., 3^{m+1} random choices done on the each ramification from 1 to 13 for the sake of gaining appropriate amount of data, additional parameter p for probability of the 1 occurrence within the random choice

was involved to make the model more complex. The results for HM Hausdorff measure together with related CPU time values are completed in Table 5.

Remark. CPU time increases rapidly with the level of ramification. The computation on level 13 lasts about 10 hours.

3.2. Hausdorff measure calculation on the tissue. On Figure 6 there are ten mammary tissues in black & white color representation. The black regions represent holes; the white ones are the surroundings. The task is to count the

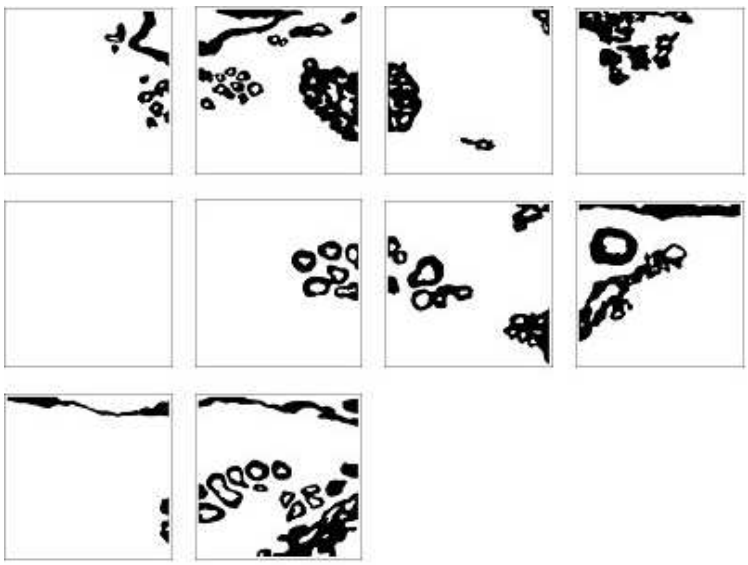

Fig. 6. Mastopatic tissue picture of which the Hausdorff measure is estimated

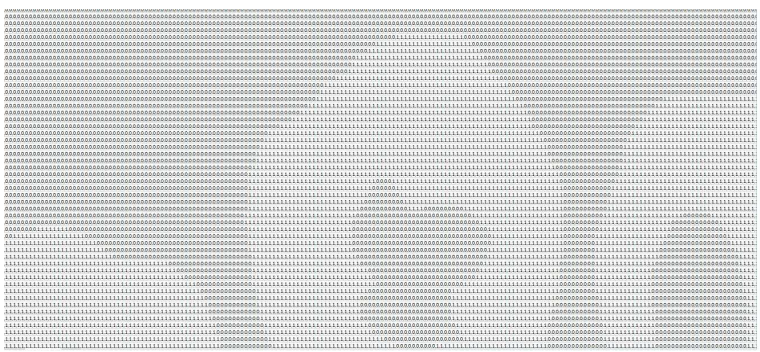

Fig. 7. Picture of a tissue is represented by zeros and ones

Table 6. Haudorff measure of mastopatic tissues

Pic. No.	No. of holes	diam(R)	(R)	HM	HM rescaled
1	12	192.005	6.584006E-02	2916.24	5.71811
2	33	359.878	0.233745	1539.62	3.01886
3	16	227.741	8.260284E-02	2757.06	5.40600
4	11	363.501	0.129592	2804.96	5.49991
5	—	—	—	—	—
6	11	158.597	6.919646E-02	2291.98	4.49408
7	19	198.479	0.147832	1342.60	2.63256
8	13	510.202	0.203264	2510.04	4.92166
9	2	512.884	5.026913E-02	10202.80	20.00540
10	29	515.248	0.263345	1956.55	3.83638

Table 7. Haudorff measure of mastopatic tissues

Pic. No.	No. of holes	diam(R)	(R)	HM	HM rescaled
1	103	217.975	0.360938	603.912	1.18414
2	94	187.579	0.335586	558.960	1.09600
3	63	220.565	0.267209	825.441	1.61851
4	62	265.006	0.210911	1256.48	2.46369
5	78	145	0.266090	544.928	1.06849
6	61	174.201	0.265483	656.168	1.28660
7	59	132.608	0.275898	480.643	0.942438
8	65	101.237	0.256071	395.349	0.775194
9	30	143.684	0.179466	800.620	1.56984
10	58	231.925	0.382088	606.993	1.19018

number and stipulate the diameter of the holes in tissue on the each slice. The diameter is given as the maximal distance between arbitrary two points of one hole. All holes on each image are measured in this way.

The density $\mu(R)$ of mastopatic tissue, as mentioned before, is the ratio of the black color to the whole picture (see Table 6). By the same procedure, the mammary tissue is treated, see Figure 8 and Table 7.

4. Conclusions and discussion. In the present paper we empirically model distribution of box-counting dimension from histological images of mammary tissues. We have shown that Weibull distribution provides reasonable fit. Moreover, discrimination between mammary cancer and masthopaty can be based on box-counting dimension percentiles, which justifies previous works in specific cancers (see e.g. [1] among others). In the second part of the paper we illustrate theoretical and practical issues in generation of Sierpinski gaskets and Hausdorff

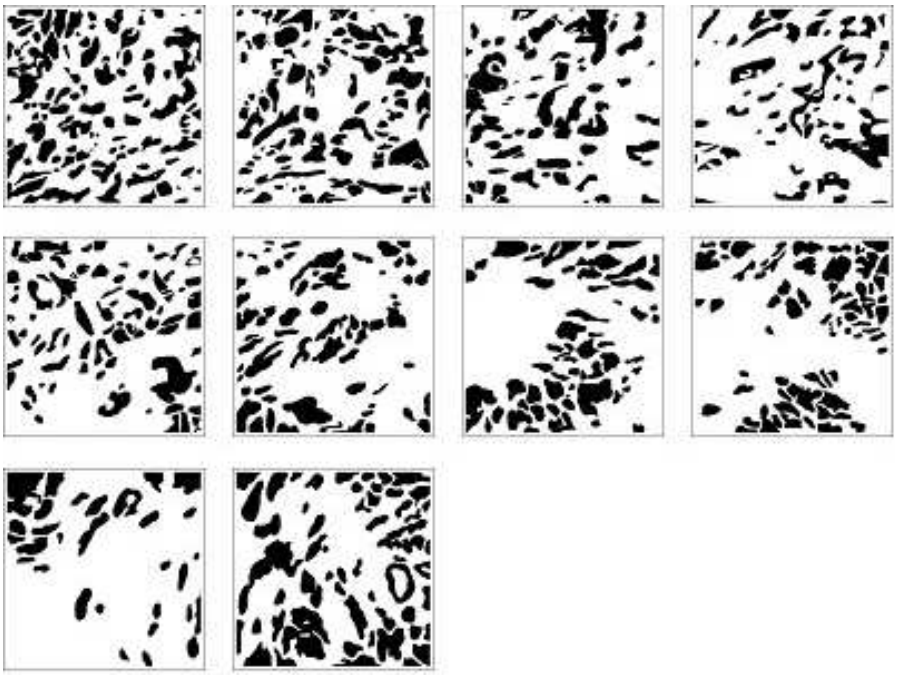

Fig. 8. Mamary tissue slices

measure calculations. Several open problems remains, e.g. what is the optimal way of computing non-Euclidean dimension for tissue? Furthermore, what is conceptual relation between mathematical model of non-Euclidean dimension and tissue growth?

5. Appendix.

5.1. ImageJ. To obtain box-counting dimension of binary images ImageJ is a useful program. ImageJ was developed at the National Institute of Health in the USA. ImageJ is a Java-based image processing program. It is public domain and designed with an open architecture which is expandable via Java plugins and recordable macros. The plugin FracLac is for Fractal Analysis. Within this plugin the option Standard Box Count is for estimating box-counting dimension. It is important to define background of binary images, because this procedure calculates dimension by examine white pixel in binary image, so the background is defined as black in this case. After these settings one picture or a folder to compute more pictures automatically can be selected. As a result you get box-counting dimension of binary images.

REFERENCES

[1] J. W. Baish, R. K. Jain. Fractals and cancer. *Cancer Research* **60** (2000), 3683–3688.

[2] J. Filus, L. Filus, M. Stehlík. Pseudoexponential modelling of cancer diagnostic testing. *Biometrie und Medizinische Informatik, Greifswalder Seminarberichte Heft* **15** (2009), 41–54.

[3] J. H. Holland. Adaptation in Natural and Artificial Systems. University of Michigan Press, 1975.

[4] B. B. Mandelbrot. Fractals: Form, Chance and Dimension. San Francisco, Freeman, 1977.

[5] T. Mattfeldt. Classification of binary spatial textures using stochastic geometry, non-linear deterministic analysis and artificial neural networks. *Int. J. Pattern Recogn. Artif. Intell.* **17** (2003), 275–300.

[6] T. Mrkvicka, T. Mattfeldt. Testing histological images of mammary tissues on compatibility with the boolean model of random sets. *Image Analysis and Stereology* **30** (2011), 11–18.

[7] M. Stehlík, S. Giebel, J. Prostakova, J. P. Schenk. Fractal based cancer risk assessment. Technical Report, IFAS, 2011.

[8] M. Stehlík, T. Mrkvicka, J. Filus, L. Filus, Recent development on testing in cancer risk: a fractal and stochastic geometry, *Journal of Reliability and Statistical Studies* **5** (2012), 83–95.

[9] Qili Xiao, Lifeng Xi. Application of Genetic Algorithm to Hausdorff Measure Estimation of Sierpinski Carpet.

Milan Stehlík, Fabian Wartner
Department of Applied Statistics
Johannes Kepler University in Linz
Freistädter Straße 315, 2. Stock
A-4040 Linz a. D., Austria
e-mail: Milan.Stehlik@jku.at
f_wartner@gmx.at

Mária Minárová
Department of Mathematics
Slovak University of Technology
Bratislava, Slovak Republic
e-mail: minarova@math.sk