TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
4-AC/2013

MAŁGORZATA CHARYTANOWICZ*

NONPARAMETRIC ESTIMATION FOR SOIL PORE SIZE DISTRIBUTION

NIEPARAMETRYCZNA ESTYMACJA ROZKŁADU WIELKOŚCI PORÓW GLEBOWYCH

Abstract

The study is concerned with the nonparametric kernel estimation to determine the soil porosity and pore size distribution. The kernel density estimation, the kernel estimation of cumulative distribution function, and the kernel estimator of quantile are considered. After a short description of the method, practical aspects and applications in agricultural science are presented. The nonparametric kernel estimation does not require a priori assumptions relating to the choice of the density function shape. Moreover, its natural interpretation together with its suitable properties makes them an adequate tool among others in estimation methods.

Keywords: *nonparametric estimation, kernel estimators, cumulative distribution function, kernel estimator of quantile, pore size distribution, pore space, total porosity*

Streszczenie

Przedmiotem niniejszego artykułu jest zastosowanie nieparametrycznej estymacji jądrowej do scharakteryzowania rozkładu wielkości porów glebowych. W artykule przedstawiono jądrowy estymator gęstości i dystrybuanty oraz opisano algorytm wyznaczania jądrowego estymatora kwantyla, istotne ze względu na badanie porowatości agregatów glebowych. Zagadnienia te zostały zilustrowane przykładowymi zastosowaniami w naukach rolniczych. Nieparametryczna estymacja jądrowa nie wymaga *a priori* założeń dotyczących kształtu funkcji gęstości rozkładu prawdopodobieństwa i jest uzasadniona w sytuacji braku znajomości jej teoretycznego modelu. Ze względu na swobodę w doborze jądra oraz procedur wyznaczania parametrów estymatora możliwe jest dostosowanie jego własności do uwarunkowań konkretnego problemu.

Słowa kluczowe: *estymacja nieparametryczna, estymatory jądrowe, agregaty glebowe, rozkład wielkości porów, porowatość gleby*

* Ph.D. Małgorzata Charytanowicz, e-mail: malgorzata.charytanowicz@ibspan.waw.pl, Institute of Mathematics and Computer Science, The John Paul II Catholic University of Lublin; Systems Research Institute, Polish Academy of Sciences, Warsaw.

1. Introduction

Along with the development of statistical methods, it is evident that classical parametric techniques are widely used in empirical studies [5, 10, 14]. This approach to density estimation assumes that the data are drawn from one of a known parametric family of distributions, determined by their parameters. The density underlying the data could then be estimated by finding from the data estimates of unknown parameters, using, for example, maximum likelihood estimation and substituting these estimates into the formula for the chosen density [3]. Such attitude requires performing goodness-of-fit tests on the data, in which the null hypothesis states that our data follows a specific distribution. When performing a statistical test, one never concludes that the null hypothesis is accepted. Whenever it is plausible that the data are consistent with the null hypothesis, the reached conclusion is not significant and does not mean that this hypothesis is true. The value of the test is statistically significant however, only in a negative case, when a decision of rejecting the tested hypothesis has been made. This leads to some subjectivism in the presented attitude, increasing rapidly in the multidimensional case, in which normal distribution is most commonly used. Moreover, for skewed distributions, a mathematical transformation, for example, logarithmic, tending the data to normal distribution, is recommended. Some data cannot be transformed satisfactorily and some should not to be. Additionally, this proceeding hinders the interpretation of the results obtained. These difficulties of the parametric approach tend to cause researchers to look for other estimation methods.

The rigidity of parametric models can be overcome by removing the restriction that the density belongs to a parametric family and assumes no pre-specified functional form for a density function. This approach leads to a nonparametric estimation method [5, 15]. Some of these methods have the advantage of being very intuitive and relatively simple to analyze mathematically. The oldest and most widely used nonparametric density estimator is the histogram. The histogram has several problems – like estimating density by a step function, the extension to multivariate settings, and finally, not using the data efficiently. Kernel smoothing provides a simple method of finding structures in data sets without these deficiencies. This technique is most relevant in many practical tasks [6, 7, 13], for example, as an effective tool for quantile estimators, especially when the underlying distribution is skewed. Moreover, because of ongoing research into the computer implementation of the algorithm, it is worth noting that all parameters appearing in the model can be effectively calculated using convenient numerical procedures based on optimizing criteria.

The main aim of this paper is to elaborate on a nonparametric estimation method, based on statistical kernel estimators, which can be successfully applied for determining the soil pore size distribution. The studies of employing pore space have been reported as a general method for defining soil structures. The most common measure characterizing the pore space within a solid is the total porosity defined as a fraction of the total pore volume that is taken up by the pore space. On the other hand, soils may be nearly uniform in regard to their total porosity but differ in pore size distribution [12, 20]. Their individual properties have different influences on fluid retention and conduction within the soil. Therefore, besides the total pore volume, their size and distribution is very important with respect to soil structure and soil fertility. These characteristics impact the majority of physical and physicochemical soil parameters, such as water retention, water conductivity, aeration, erosion susceptibility and gas diffusivity. They vary significantly due to several factors, including mutual interactions between soil fauna,

soil macro-organisms, roots, inorganic agents and environment factors. Weather conditions, vegetation, fertilization and soil tillage operations can cause both soil loosening or compaction. As far as water-air soil properties are concerned, pores are usually divided into three groups: micropores, mesopores and macropores, with the division between them being arbitrary. Macropores, relevant to coarse-grained soils, drain freely by gravity and allow easy movement of water and air. They provide a habitat for soil organisms and plant roots, which can grow into them. On the other hand, macropores cannot hold water under tension induced by gravity when allowed to drain after saturation. Inverse properties are found in fine-grained soils, containing a major amount of micropores. These soils retain large amounts of water, due to the fact that this water would be considered unavailable to plants. Mesopores, i.e. medium-size pores, are essential for capillary water distribution. They provide water storage sites which retain water useful to plants. Soils with a predominance of mesopores and a moderate system of micro- and macropores possess most favorable physical properties relevant to plant growth.

A more detailed analysis can be obtained by developing a pore-size distribution curve. Soil pore size distribution is often determined by mercury intrusion porosimetry or low temperature nitrogen adsorption. However these methods do not provide the all the information about pore size and shape, as they are not appropriate for pore measurements. Recent advances in computed tomography and digital image processing algorithms [4, 11, 17, 19] provide technologically advanced measurement tools for studying the internal structures of soil aggregates.

In elaborate investigations, the internal structure of an aggregate was visualized by microtomography scanning. A research study was conducted using image analysis algorithms appropriate for pore measurements, and in turn, kernel estimation techniques. After a short description of the method, practical aspects and applications in agricultural science are presented.

2. Statistical Kernel Estimators

Let (Ω, Σ, P) be a probability space. Let a real random variable $X : \Omega \to R$, whose distribution has the density function f, also be given. The corresponding kernel estimator $\widehat{f} : R \to [0, \infty)$, calculated using experimentally obtained values for the m-element random sample $x_1, x_2, ..., x_m$, in its basic form is defined by:

$$\widehat{f}(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h} \right) \tag{1}$$

where $m \in \mathbf{N}/\{0\}$, the positive coefficient h is known as a smoothing parameter, whereas the measurable function $K : R \to [0, \infty)$ of unit integral, symmetrical with respect to zero and having a weak global maximum at this point, takes the name of a kernel. The influence of the smoothing parameter on particular kernels is the same for the basic definition of the kernel estimator (1). Advantageous results are obtained thanks to the individualization of this effect, achieved through a so-called modification of the smoothing parameter [6, 15, 18]. It relies on mapping the positive modifying parameters $s_1, s_2, ..., s_m$ on particular kernels, described as:

$$s_i = \left(\frac{\widehat{f}(x_i)}{\overline{s}} \right)^{-c} \tag{2}$$

where $c \in [0, \infty)$, $\widehat{f}$ denotes the kernel estimator without modification, and $\overline{s}$ is the geometrical mean of the numbers $\widehat{f}(x_1), \widehat{f}(x_2), ..., \widehat{f}(x_m)$. The parameter c stands for the intensity of the modification procedure and based on indications for the criterion of the integrated mean square error, the value $c = 0.5$ can be suggested. Finally, the kernel estimator with the smoothing parameter modification is defined in the following formula:

$$\widehat{f}(x) = \frac{1}{mh} \sum_{i=1}^{m} \frac{1}{s_i} K\left(\frac{x - x_i}{hs_i} \right) \tag{3}$$

Specifying the kernel estimator of a density function f, gives a natural description of the distribution of X, and allows the estimator of the distribution function, denoted hereinafter as $\widehat{F} : R \to [0, 1]$, to be found from the relation:

$$\widehat{F}(x) = \int_{-\infty}^{x} \widehat{f}(u)\,\mathrm{d}u \tag{4}$$

where $\widehat{f}$ denotes the kernel density estimator (3). Denoting the primitive of a kernel K as $K : R \to [0, 1]$, that is:

$$I(x) = \int_{-\infty}^{x} K(u)\,\mathrm{d}u \tag{5}$$

the kernel estimator of the distribution function can be expressed as:

$$\widehat{F}(x) = \frac{1}{m} \sum_{i=1}^{m} I\left(\frac{x - x_i}{hs_i} \right) \tag{6}$$

If for the estimator (3), one uses a kernel with positive values, then the function I, and thus $\widehat{F}$ are strictly increasing.

In the case when the kernel estimator of the distribution function $\widehat{F}$ is used, the kernel estimator of the quantile of order r denoted as $\widehat{g} \in R$, may be uniquely defined by the solution of the equation:

$$\widehat{F}(x) = r \tag{7}$$

The equation (7) can be expressed equivalently in a form:

$$\widehat{F}(x) - r = 0 \tag{8}$$

If the left side of the equation (8) is denoted by $L(x) = \widehat{F}(x) - r$, then $L'(x) = \widehat{f}(x)$, and the kernel estimator of the quantile can be effectively calculated on the basis of Newton's algorithm [2, 8, 16] as the limit of the sequence $\left\{\hat{q}_k\right\}_{k=0}^{\infty}$ defined by:

$$\hat{q}_0 = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{9}$$

$$\hat{q}_{k+1} = \hat{q}_k - \frac{\widehat{F}(\hat{q}_k) - r}{\widehat{f}(\hat{q}_k)} \quad \text{for} \quad k = 0, 1, \ldots \tag{10}$$

The choice of the kernel K form and the calculation of the smoothing parameter h is made most often with the criterion of the mean integrated square error [1, 9]. From a statistical point of view, the choice of the kernel form has no practical meaning and thanks to this, it becomes possible to take into account primarily properties of the estimator obtained or calculation aspects, advantageous from the viewpoint of the application problem under investigation. The standard normal kernel given by:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{11}$$

is used most often. It is differentiable up to any order and assumes positive values in the whole domain.

The fixing of the smoothing parameter h has significant meaning for the quality of estimation. A smoothing parameter controls the tradeoff between bias and variance in the result. A large bandwidth leads to a very smooth density distribution, whereas a small bandwidth leads to a ragged density distribution. A frequently used bandwidth selection technique, called the 'cross-validation method', chooses h to minimize the function $g : R \rightarrow R$ defined as:

$$g(h) = \frac{1}{m^2 h}\sum_{i=1}^{m}\sum_{j=1}^{m}\tilde{K}\left(\frac{x_j - x_i}{h}\right) + \frac{2}{mh}K(0) \tag{12}$$

where:

$$\tilde{K}(x) = K^{*2}(x) - 2K(x) \tag{13}$$

whilst K^{*2} denotes convolution function of k, i.e.:

$$K^{*2}(x) = \int_R K(u)K(x-u)\,\mathrm{d}u \tag{14}$$

The tasks concerning the choice of the kernel form, as well as additional procedures improving the quality of the estimator obtained, and all rules needed for calculating the smoothing parameter, are found in [6, 15, 18]. The utility of kernel estimation has been investigated in the context of determining the soil pore size distribution.

3. Material and Methods

The investigated material was sampled from the cultivated soil layer classified as silty loam (WRB Mollic Gleysols), explored at the Institute of Agrophysics, at the Polish Academy of Sciences in Lublin. The proportion of each particle size group in the soil was as follows: sand – 46%, silt – 28%, clay – 26%, pH was: H_2O – 5.9, KCl – 5.4. On the experimental fields, a long-term fertilization trial was executed. The adopted crop rotation from 1955 to 1989 was a cycle of potato – barley – rye, and from 1990 – a cycle of sugar beat – barley – rape – wheat. Three treatments concerning fertilization: control group – plant residues only, mineral fertilization – according to plant needs; pig manure – 80 ton per ha; were studied. Aggregate soil organic matter was measured by the Multi N/C 3100 Autoanalyser (Analitic Jena, Germany). The total organic carbon and total nitrogen contents for three fertilizations (pig manure, mineral fertilizers, control) were respectively: 21.50, 14.89; 13.54 g/kg; 2.10; 1.51; 1.35 g/kg. The total organic carbon shows the same tendency as total nitrogen, i.e. increasing in the same order: the lowest – control, middle – mineral fertilization, the highest – pig manure.

Soil samples were air dried in room conditions, divided into smaller amounts, and gently sieved through 2 and 10 mm sieves. Soil aggregates remaining at 2 mm sieve and ranging from 2 to 10 mm were then detected by means of *X*-ray computational tomography. The direct and nondestructive analysis of internal soil aggregate structure was detected using a GE Nanotom S device, with the voxel-resolution of 2.5 microns per volume pixel. Three 2D sections uniformly located within each aggregate were performed to characterize the aggregate structure. Next, tomography sections were processed using the Aphelion 4.0.1 package. Thus pore size distribution of a particular aggregate was determined by means of image processing techniques. In the initial step, the ROI (region of interest) selection from the original grayscale image was performed. All of the ROI's were selected by hand, around the aggregate, removing the ring artifacts, and next they were saved as a bitmap format. The automatic Otsu binarization method was then employed to separate pores within the sampled aggregate. Subsequently, binary morphological closing with increasing size of square structuring element was processed. The operation was repeated until all pores were filled. Subtraction of the transformed image from the original image gives the total pore volume in the examined aggregate. Pore identification and the determination of pore radius were done automatically using morphological operations, clustering and splitting procedures. Each pore was individually identified and approximated by a circle of radius r, calculated from the surface area.

Finally, the pore size distribution was presented in the form of pore radius distribution curves using kernel density estimation (as described in section 2). A more detailed analysis was obtained by kernel estimation of the distribution function (6) and kernel estimation of the quantile (7).

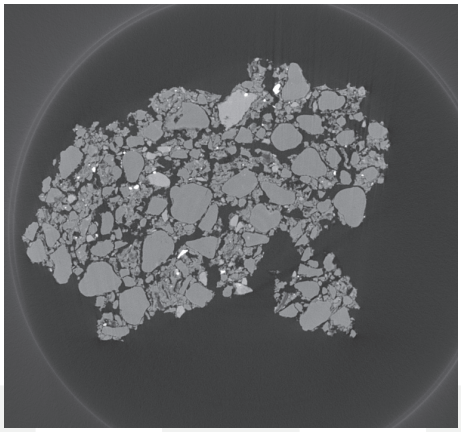

Fig. 1. The soil aggregate microtomographic image

4. Estimation of the Pore Size Distribution

The examined soil samples comprised aggregates belonging to three treatments, different in terms of fertilization: selected for the experiment control (without fertilization); mineral fertilization, and pig manure. High quality visualization of the internal soil structure was detected using a non-destructive soft *X*-ray technique. Tomography sections were processed using the image processing methods described in section 3. The total porosity of the investigated aggregates, calculated as the average of three sections, are as follows: control group – 14.2%, mineral fertilization – 22.95%, manure fertilization – 33.28%. The fraction of the total aggregate volume occupied by soil pores was significantly greater when manure fertilization was used. Moreover the total porosity was higher for both the manure and the mineral fertilization than for the control group.

For each fertilization, the examined group constitutes a one-dimensional sample containing 510 measurements of pore radii ranging from 5 μm to 0.24 mm. Kernel density estimates (3) for soil pore size, based on the data, were constructed using the Cauchy kernel given by the rule:

$$K(x) = \frac{2}{\pi} \frac{1}{(1+x^2)^2} \tag{15}$$

Its primitive has a form convenient for further calculations:

$$I(x) = \frac{1}{\pi}\left(\frac{x}{1+x^2} + \operatorname{arctg}(x) + \frac{\pi}{2} \right) \tag{16}$$

The smoothing parameters h detected by the *cross-validation* method equal to 1.06, 2.96, and 1.49 for the samples belonging to the control group, mineral and manure fertilized aggregates

respectively. The soil pore size distributions evaluated from the databases of examined groups, constructed by means of the kernel density estimates, are shown in Fig. 2.

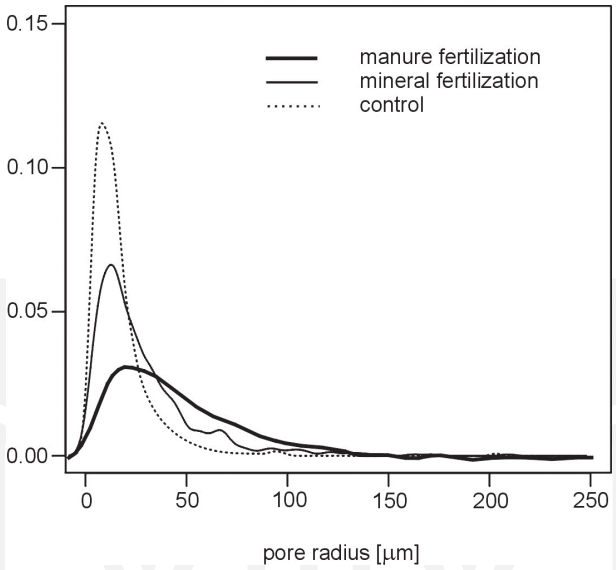

Fig. 2. Kernel density estimates for soil pore size

As shown, both unimodal and right-skewed distributions are well developed for all three data samples. It should also be noted that the region between 0 to 50 μm is the most crucial representation of the pore size interval of the control group, whereas the region between 0 to 100 μm is the most representative pore size interval of the fertilized samples. Moreover, the first considered region is somewhat narrowed compared with the two regions of the fertilized samples. Furthermore, the highest and sharpest peak was found for the control group, followed by the mineral and manure fertilizations. This peak comes about as a result of the substantial contribution of narrow pores evident within the samples, and the control group was seen to contain more narrow pores than the fertilized (both mineral and manure) samples. In our data, upon analysis it was evident that the second and third peaks are moved to the right side. This, we feel, indicates a greater proportionality of less narrow pores. This characteristic feature complies well with the major aim of the study – to present a reliable density estimation method for evaluating pore size distribution patterns.

In order to perform further analysis of micro, meso and macropores the kernel estimator of the distribution function (6) was constructed. The limits of mesopores, taken arbitrarily, were between 30 and 75 μm. Table 1 contains fractions of micro, meso and macropores for each type of fertilization.

Calculated fractions confirm the results obtained by means of kernel density estimates, as given in Figure 2. The largest fraction of mesopores occurs in the soil fertilized with pig manure, this fraction represents 39% of the total pore area. Moreover, it contains percentages of 26% macropores and 35% micropores. This creates the most favorable conditions for plant growth. The largest fraction of micropores, equaling 87%, was observed in the soil without

fertilization. This soil has a small amount of macropores, equaling 1%, and a small amount of mesopores, equaling 12%. This creates the least favorable conditions for plant growth. The soil with mineral fertilization contains 65% of micropores, 28% of mesopores and 7% of macropores.

Table 1

Fractions of micro-, meso- and macropores

Management fertilization	Fraction of micropores	Fraction of mesopores	Fraction of macropores
Pig manure	0.35	0.39	0.26
Mineral fertilization	0.65	0.28	0.07
Control group	0.87	0.12	0.01

Table 2 shows quantile estimators of order 0.25, 0.5, and 0.75, calculated using rules (9)–(10).

Table 2

Quantile estimators of order 0.25, 0.5, and 0.75

Management fertilization	Quantile of order 0.25 [μm]	Quantile of order 0.5 [μm]	Quantile of order 0.75 [μm]
Pig manure	22.63	43.50	76.50
Mineral fertilization	12.83	22.35	38.98
Control group	7.45	13.00	20.90

The largest fraction of large pores occurs in the soil fertilized with pig manure, 50% of its pores have a radius between 22.63 and 76.50 microns. The soil with mineral fertilization incorporates 50% of its pores having a radius between 12.83 and 38.98 microns. The largest fraction of small pores occurs in the soil without fertilization, 50% of its pores has a radius between 7.45 and 20.90 microns.

This study has shown the possibility to determine soil pore size distribution using the nonparametric kernel estimation theory. Generally, the effect of fertilization is to increase amounts of meso- and macropores in relation to the control group. The greater increase is for pig manure fertilization. Soils without fertilization contain significantly more micropores. These soils possess less favorable properties for plant growth.

5. Summary

Recent advances in computed tomography and digital image processing provide non-destructive tools for studying the internal structures of soil aggregates. This seems very useful in characterizing the pore size distribution and in quantifying the differences in pore structures

from the different types of soil. A more detailed analysis may be obtained by deriving various methods to quantify the pore structure and to develop a pore size-distribution curve.

In this paper, an innovative method for characterizing the soil porosity and pore size distribution, based on computed tomography and nonparametric estimation, is proposed. The presented algorithm, based on image processing methods and the kernel estimators technique, is expected to be an effective procedure for this purpose. The presented approach is more objective than classical parametric methods, and can be successfully applied for many tasks in data mining, where arbitrary assumptions concerning the form of density function are not recommended.

This approach is also motivated by the current rapid growth in computational power. Improved real-time data processing and algorithm efficiency having important meaning due to the concurrent increase in the quantity and complexity of the data that are being collected. Historically, such data have been analyzed using classical methods. The presented approach is useful for determining the pore size distribution of any material for which a morphometric analysis is done, as it eliminates estimation subjectivity.

References

[1] Draper N.R., Smith H., *Applied regression analysis*, John Wiley and Sons, New York 1981.

[2] Kincaid D., Cheney W., *Numerical Analysis*, Brooks/Cole, Pacific Grove, 2002.

[3] Koronacki J., Mielniczuk J., *Statystyka dla studentów kierunków technicznych i przyrodniczych*, WNT, Warszawa 2001.

[4] Król A., Niewczas J., Charytanowicz M., Gonet S., Lichner L., Czachor H., Lamorski K., *Water-stable and non-stable soil aggregates and their pore size distributions*, 20th International Poster Day and Institute of Hydrology Open Day 'Transport of water, chemicals and energy in the soil – plant – atmosphere system', 2012, 870-871.

[5] Kruszewski D., *Nonparametric modeling of medical scheme data*, Technical Transactions,1-AC/2013, 93-117.

[6] Kulczycki P., *Estymatory jądrowe w analizie systemowej*, WNT, Warszawa 2005.

[7] Kulczycki P., Charytanowicz M., *Conditional Parameter Identification with Different Losses of Under- and Overestimation*, Applied Mathematical Modelling, 37 (4), 2013, 2166-2177.

[8] Kulczycki P., Dawidowicz A.L., *Kernel Estimator of Quantile*, Universitatis Jagiellonicae Acta Mathematica, 37, 2005, 101-112.

[9] Lange K., *Numerical analysis for statisticians*. *Statistics and Computing*, Springer, New York 2000.

[10] Motulsky H., *Intuitive Biostatistics*, Oxford University Press, New York 1995.

[11] Peth S., Nellesen J., Fischer G., Horn R., *Non-invasive 3D analysis of local soil deformation under mechanical and hydraulic stresses by μCT and digital image correlation*, Soil and Tillage Research, 111 (1), 2010, 3-18.

[12] Pires de Silva A., Imhoff S., Kay B., *Plant response to mechanical resistance and air-filled porosity of soils under conventional and no-tillage system*, Scientia Agricola, 61 (4), 2004, 451-456.

[13] Sheather S.J., Marron J.S., *Kernel quantile estimators*, Journal of the American Statistical Association, 85, 1990, 410-416.
[14] Siegel A.F., *Statistics and data analysis: an introduction*, Wiley and Sons, New York 1988.
[15] Silverman B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London 1986.
[16] Stoer J., Bulirsch R., *Wstęp do analizy numerycznej*, PWN, Warszawa 1987.
[17] Tadeusiewicz R., Korohoda P., *Komputerowa analiza i przetwarzanie obrazów*, Wydawnictwo Fundacji Postępu Telekomunikacji, Kraków 1997.
[18] Wand M.P., Jones M.C., *Kernel Smoothing*, Chapman and Hall, London 1994.
[19] Wojnar L., Majorek M., *Komputerowa analiza obrazu*, Computer Scanning System, Warszawa 1994.
[20] Zdravkov B., Cermak J., Sefara M., Janku J., *Pore classification in the characterization of porous materials*: *A perspective*, Central European Journal of Chemistry, 5 (2), 2007, 385-395.

PK