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

MAŁGORZATA CHARYTANOWICZ[*], HENRYK CZACHOR[**], JERZY NIEWCZAS[***]

NONPARAMETRIC REGRESSION APPROACH: APPLICATIONS IN AGRICULTURAL SCIENCE

ZASTOSOWANIE REGRESJI NIEPARAMETRYCZNEJ W NAUKACH ROLNICZYCH

A b s t r a c t

In this paper, a method for determining the soil pore size distribution, constituting the subject of the presented investigations, is proposed. A research study was conducted using image analysis algorithms, and in turn, nonparametric statistical techniques. The results and further work will be discussed in section four. The purpose of this investigation is to discover the relationship between the pore size and volume of the corresponding pores. The algorithm presented here is based on the theory of statistical kernel estimators. This frees it of assumptions in regard to the form of regression function. The approach is universal, and can be successfully applied for many tasks in data mining, where arbitrary assumptions concerning the form of regression function are not recommended.

Keywords: nonparametric regression, kernel estimators, morphological image processing, closing procedure, pore size distribution, pore space, total porosity, soil structure, aggregation, soil compaction

S t r e s z c z e n i e

Celem niniejszego artykułu jest zaprezentowanie procedury wyznaczania rozkładu wielkości porów w agregatach glebowych. Do scharakteryzowania zależności pomiędzy badanymi zmiennymi wykorzystana zostanie funkcja regresji. W przeprowadzonych badaniach zastosowano algorytmy analizy obrazów cyfrowych oraz metodykę statystycznych estymatorów jądrowych. Przedstawiona metoda umożliwia uzyskanie właściwej charakterystyki rozkładu wielkości porów i może stanowić efektywne narzędzie stosowane w wielu zagadnieniach eksploracji danych. Jako model nieparametryczny, nie wymaga założeń dotyczących kształtu zależności funkcyjnej między rozpatrywanymi zmiennymi.

Słowa kluczowe: regresja nieparametryczna, estymatory jądrowe, przekształcenia morfologiczne, operacja zamknięcia, rozkład wielkości porów, porowatość ogólna gleby, struktura gleby, agregacja, gęstość gleby

[*] Małgorzata Charytanowicz, Ph.D., Institute of Mathematics and Computer Science, The John Paul II Catholic University of Lublin; System Research Institute, Polish Academy of Sciences.

[**] Henryk Czachor, D.Sc., Ph.D., Bohdan Dobrzański, Institute of Agrophysics, Polish Academy of Sciences.

[***] Jerzy Niewczas, D.Sc., Ph.D., Institute of Mathematics and Computer Science, The John Paul II Catholic University of Lublin.

1. Introduction

Pore size distribution is one of many physical measurements characterizing soil structure as far as plant growth is concerned. A number of scientists have reported studies of pore space as a general method for defining soil properties. In this respect, a complete analysis of the soil pore size distribution is used for predicting the gas diffusivity, water infiltration rates, water availability to plants, water-storage capacity and movement of water.

The most common measure characterizing the fraction of the pore space within a solid is the total porosity, defined by the ratio:

$$\varphi = \frac{V_P}{V_T} \tag{1}$$

where:

V_p – the volume of void-space,

V_T – the total volume of soil material, including the solid and void components.

Porosity is a dimensionless quantity and can be reported either as a decimal fraction or as a percentage. Being simply a fraction of total volume, it can range between 0 and 1, typically falling between 0.3 and 0.7 for most soils. This provides a more useful physical description of an particular soil, such as providing an estimate of compaction and the maximum space available for water. Moreover, a number of scientists have reported that studies of pore size distribution are useful as a general method for defining the soil structure. Pore sizes usually have traditionally been divided into macropores and micropores, with the division between the two being arbitrary. Because of the relationship between the pore properties and the reaction of chemicals in the soil, a more detailed pore-fraction analysis seems warranted [7, 15].

The purpose of this investigation is to elaborate a method of measuring the pore size distribution in the soil. Internal difference in porosity within an aggregate can be visualized by microtomography scanning of the air-dried samples. Once scanned, the computed tomography information allows the non-destructive visualization of slices, arbitrary sectional views and pseudo-color representations [14].

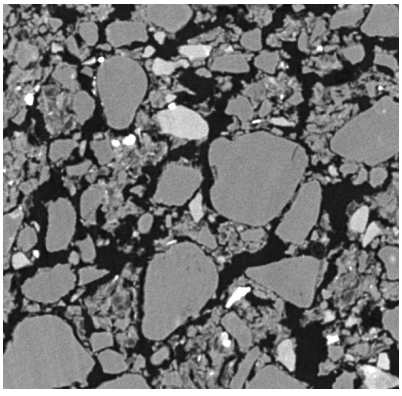

Fig. 1. Cross section of a typical soil with pore space in black

Rys. 1. Przekrój agregatu glebowego, pory zostały wyróżnione kolorem czarnym

For image processing, the authors used a program for computer image analysis package. After preprocessing methods, morphological operation closing, consisting of a dilation followed by erosion, was used to fill in holes and small gaps without changing the size and original boundary shape [5, 6, 12, 17]. Subsequently, digital images were segmented into pore space and solid. The data derived automatically from images was then statistically examined, as nonparametric statistical regression allowed for the determination of the soil pore distribution.

2. Material and Methods

These studies were conducted using soil aggregates from the cultivated soil layer explored at the Institute of Agrophysics, of the Polish Academy of Sciences in Lublin. The direct and nondestructive analyses of internal soil aggregate structures were detected using computed tomography equipment Nanotom 800, with the voxel-resolution of 2.5 microns per volume pixel [4, 8]. Three types of aggregates, different in terms of fertilization, denoted as aggregates 0 – without fertilization, NPK – mineral fertilization, and OB – pig manure, were studied. Tomography sections were processed using macros writing in the Aphelion 4.0.1 package.

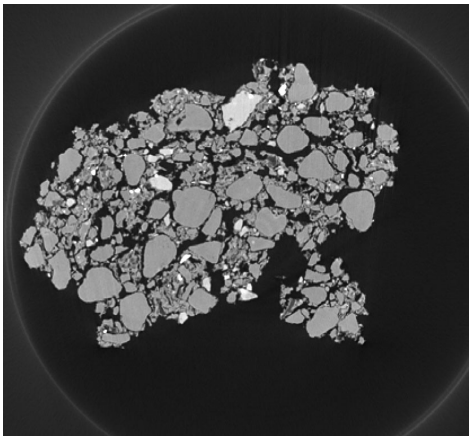

Fig. 2. The soil aggregate microtomographic image

Rys. 2. Obraz tomograficzny agregatu glebowego

The entire procedure was composed of the following steps. Firstly, grayscale images were preprocessed by removing ring artifacts using the ROI method. After selecting the region of interest, the automatic Otsu binarization method was then employed.

Subsequently, binary morphological closing with increasing size of square structuring element (starting with size 2), was processed. In each step, the source image was subtracted from the target image, and the result volumes were listed, giving soil aggregate pore distribution. 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 sample soil pores. Finally, the pore fractions analysis was performed using the regression approach [3].

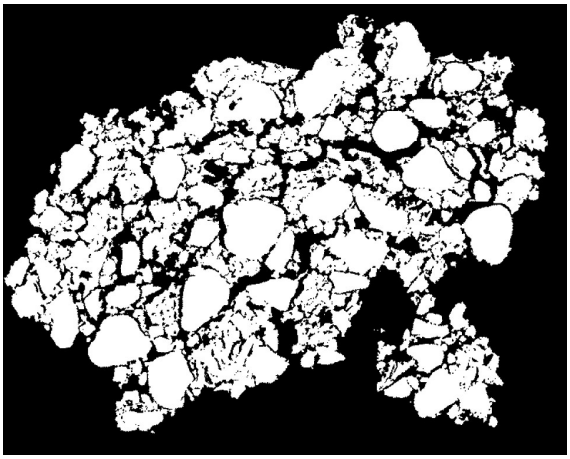

Fig. 3. The soil aggregate image after selecting ROI and binarization through using Otsu method

Rys. 3. Obraz agregatu glebowego po zastosowaniu selekcji ROI i binaryzacji metodą Otsu

The procedure for determining the soil pore size distribution
1. X-ray microtomographic image analysis of aggregates.
2. Ring artifact removal using – the ROI method.
3. Image binarization – the OTSU method.
4. Determining the pore size distribution – binary morphological closing.
5. Pore fractions analysis – regression approach.

In next section, the regression approach is shortly described. A simple linear regression model is not appropriate for the data. Therefore nonlinear estimation and nonparametric methods were concerned.

3. The regression analysis

The validation of the efficiency and accuracy of regression technique was explored by comparing results on a variety of real datasets. Both a classical parametric regression model and several nonparametric methods were examined as far as the pore size distribution was concerned.

The best fit to the data in the family of nonlinear estimation was revealed for two models: polynomial with the power of three; and logarithmic, for which the determination coefficients were in range 80–90%. The proposed regression functions did not well discover the properties of the pore size distributions, especially on the left side, because of their positively skewed and unimodal character. Therefore, the nonparametric kernel method was recommended [1, 9, 10].

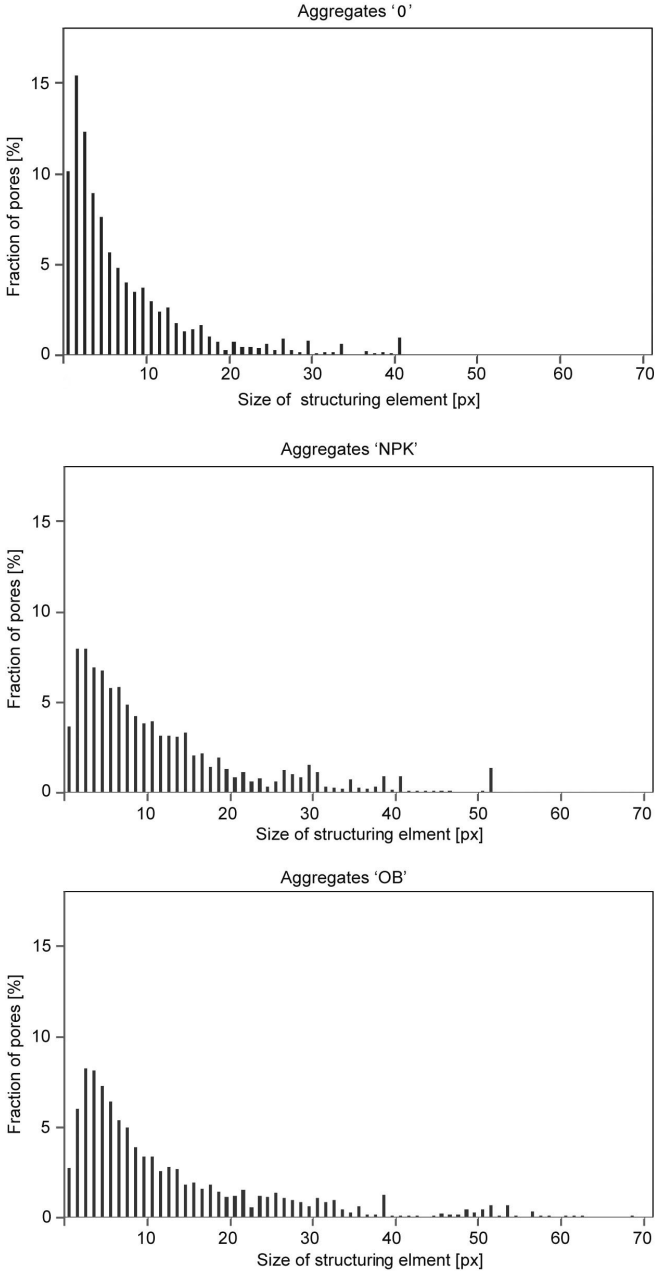

Fig. 4. The pore size distribution: the data presents three types of aggregates different in terms of fertilization, denoted as aggregates: 0 – without fertilization, NPK – mineral fertilization, and OB – pig manure

Rys. 4. Rozkłady porów, wykresy przedstawiają trzy typy agregatów: 0 – bez nawożenia, NPK – nawóz mineralny, OB – nawóz naturalny

Classical parametric methods of determining an appropriate functional relationship between the two variables impose arbitrary assumptions concerning the functional form of the regression function. Moreover, the choice of parametric model depends very much on the situation. If a chosen parametric family is not of appropriate form, then there is a danger of reaching incorrect conclusions in the regression analysis. This also makes it difficult to take into account the whole of the accessible information. However, the rigidity of this regression can be overcome by removing the restriction that the model is parametric. This approach leads to nonparametric regression. This lets the data decide which function fits them best. In this study, a class of kernel–type regression estimators called 'local polynomial kernel estimators' is presented.

Let therefore, m elements $(x_i, y_i) \in R \times R$, $i = 1, 2,..., m$ be given, where values x_i may designate some non-random numbers or realizations of the one-dimensional random variable X, whereas y_i designate realizations of the one-dimensional random variable Y. Assuming the existence of the function $f: R \rightarrow R$ having a continuous first derivative that:

$$y_i = f(x_i) + \varepsilon_i \tag{2}$$

where ε_i are independent random variables with zero mean and unit finite variance. Let then $p \in N$ be the degree of the polynomial being fit. The kernel regression estimator $\hat{f}: R \rightarrow R$, obtained by using weighted least squares with kernel weights, is given by the formula:

$$\hat{f}(x) = e(X^T W X)^{-1} X^T W y \tag{3}$$

where:

$$y = [y_1, y_2, \ldots, y_m]^T \tag{4}$$

is the vector of responses,

$$X = \begin{bmatrix} 1 & x_1 - x & (x_1 - x)^2 & \cdots & (x_1 - x)^p \\ 1 & x_2 - x & (x_2 - x)^2 & \cdots & (x_2 - x)^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m - x & (x_m - x)^2 & \cdots & (x_m - x)^p \end{bmatrix} \tag{5}$$

is an $m \times (p + 1)$ design matrix, and:

$$W = \operatorname{diag}\left(\frac{1}{h} K\left(\frac{x_1 - x}{h} \right), \frac{1}{h} K\left(\frac{x_2 - x}{h} \right), \ldots, \frac{1}{h} K\left(\frac{x_m - x}{h} \right) \right) \tag{6}$$

is an $m \times m$ diagonal matrix of kernel weights, while:

$$e = [1, 0, \ldots, 0]^T \tag{7}$$

is the 1 × (p + 1) vector having 1 in the first entry and zero elsewhere. The coefficient $h > 0$ is called 'a bandwidth', while the measurable function $K : R \to [0, \infty)$ of unit integral, symmetrical with respect to zero, and having a weak global maximum in this place, takes the name of the kernel.

An important problem is the choice of the parameter p. For sufficiently smooth regression functions, the asymptotic performance of $\hat{f}$ improves for higher values of p. However, for higher p, the variance of the estimator becomes larger, and in practice, a very large sample may be required. On the other hand, the even degree polynomial kernel estimator has a more complicated bias expression which does not lend itself to simple interpretation. These facts suggests the use of either $p = 1$ or $p = 3$. Moreover, for $p = 1$, the convenient explicit formulae exists:

$$\hat{f}(x) = \frac{1}{mh} \sum_{i=1}^{m} \frac{\left(\hat{s}_2(x) - \hat{s}_1(x)(x_i - x)\right) y_i K\left(\frac{x - x_i}{h}\right)}{\hat{s}_2(x)\hat{s}_0(x) - \left(\hat{s}_1(x)\right)^2} \tag{8}$$

where:

$$\hat{s}_r(x) = \frac{1}{mh} \sum_{i=1}^{m} (x_i - x)^r K\left(\frac{x_i - x}{h}\right) \text{ for } r = 0, 1, 2 \tag{9}$$

Therefore, except in more advanced statistical applications, $p = 1$ is preferred.

The choice of the kernel form has no practical meaning and thanks to this, it is possible to take into account the primarily properties of the estimator obtained. Most often, the standard normal kernel is expressed by the convenient analytical formula:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{10}$$

is used.

The practical implementation of the kernel regression estimators requires a good choice of bandwidth. If h is too small, a spiky rough kernel estimate is obtained, and if h is too large, it results in a flat kernel estimate. A frequently used bandwidth selection technique is the 'cross-validation method' introduced by Clark [2], which chooses h to minimize:

$$CV(h) = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \hat{f}_{-i}(x_i)\right)^2 \tag{11}$$

where:

$\hat{f}_{-i}(x_i)$ – the leave-one-out kernel regression estimator based on data $(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_m, y_m)$.

A typical regression curve obtained by the kernel regression method for m = 200 data points generated from function $f(x) = 0{,}1 \cdot \sin(5x - 2) + e^{-(x-1)^2/4}$ as defined on interval [0, 3] is demonstrated on Fig. 5. A standard normal kernel K given by rule (10) and bandwidth h calculated by the cross-validation method (11) were used.

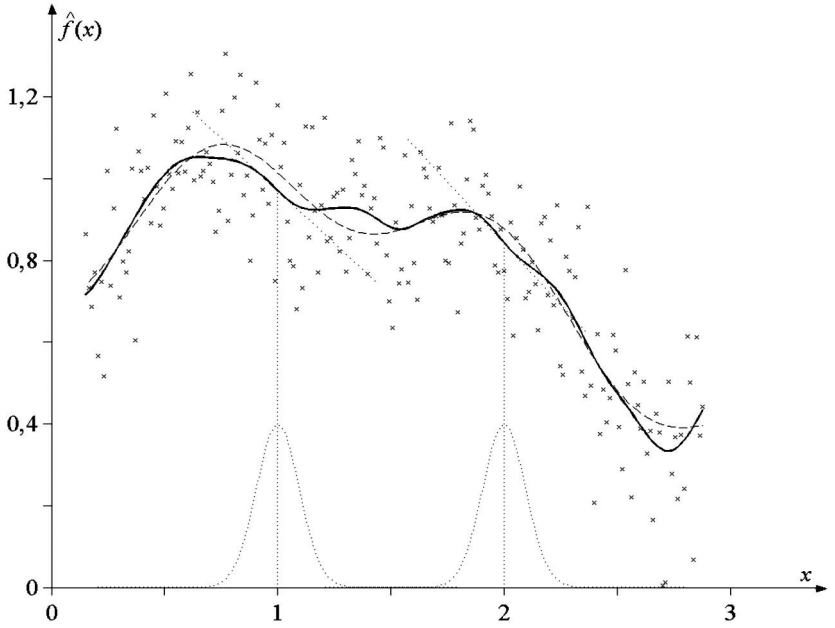

Fig. 5. The estimated regression: the data points are represented by a cross, the true function curve by a dashed line, the regression curve by a solid line, and kernels for arguments $x = 1$ and $x = 2$ by a dotted line

Rys. 5. Jądrowy estymator funkcji regresji: wartości próby losowej oznaczono krzyżykami, estymowaną funkcję linią przerywaną, jądrowy estymator funkcji regresji linią ciągłą, jądra dla argumentów $x = 1$ i $x = 2$ linią kropkowaną

The tasks concerning the choice of the kernel form, the bandwidth, as well as additional procedures improving the quality of the estimator obtained, are found in [13, 16]. The utility of local linear kernel estimators has been investigated in the context of some typical data derived from the soil aggregate images.

4. Results and discussion

The main aim of this research is to discover the relationship between the pore size and volume of the corresponding pores in soil aggregates. The data obtained by the proposed method, based on the image processing algorithms, allows the use of regression approach.

The kernel regression built upon a weighted local linear regression was used to analyze the soil pore size distribution. For ease of computation, the standard normal kernel (10) was used. The bandwidth was determined using the cross–validation method (11).

Thus, particle-size distribution was translated into an equivalent regression model, which in turn, is related to characterize water retention, crucial in any modeling study on water flow and solute transport in soil. This can be used for comparing and predicting some specific points of interest of the water retention characteristics.

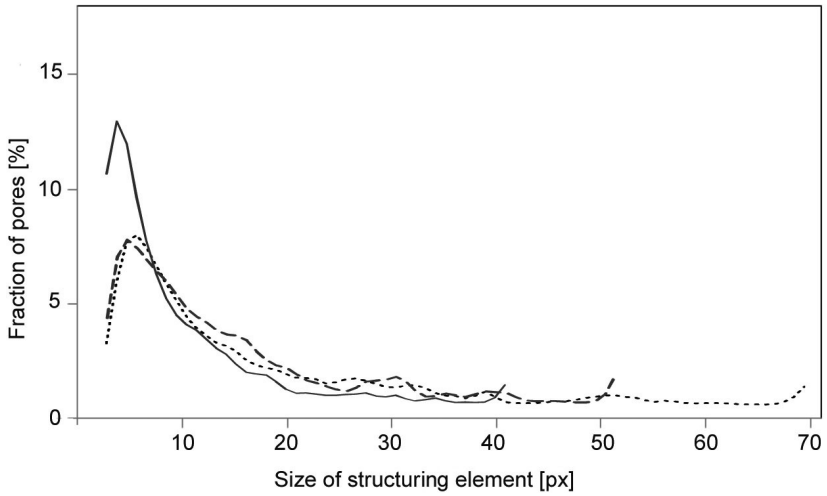

Fig. 6. The estimated regression: the data presents three types of aggregates different in terms of fertilization, denoted as aggregates 0 – without fertilization (a solid line), NPK – mineral fertilization (a dashed line), and OB – pig manure (a dotted line)

Rys. 6. Jądrowy estymator funkcji regresji dla trzech typów agregatów: 0 – bez nawożenia (linia ciągła), NPK – nawóz mineralny (linia przerywana), OB – nawóz naturalny (linia kropkowana)

Fraction of total aggregate volume occupied by pores was significantly greater in OB aggregates. There was also greater percentage of large pores in OB aggregates than in 0 or NPK aggregates. In the smallest range, however, porosity of 0 aggregates exceeded that of NPK and OB aggregates. Total porosities were higher for both NPK and OB aggregates (24% and 32%, correspondingly) than the 14% for 0 aggregates.

A combination of higher microporosity and higher percent of large pores in OB aggregates may generate more favorable conditions for microbial activity through a combination of high water-holding capacities, increased aeration and gas transport. Our current aim is comparing and relating specifics of internal pore structures in the aggregates from their water stability, composition and chemical properties. Water stable aggregates are much more porous in relation to the non stable aggregates.

Soil aggregation and its maintenance is very important for sustainable agriculture because it impacts majority of biological and physical soil properties and processes. This directly promotes better movement of air and of water in the soil, prevents runoff and soil water erosion, and indirectly determines plant growth. However soil aggregates are very sensitive against water and mechanical stresses. Discovering the main factors determining their water stability is a great challenge for worldwide agriculture which would help to prevent soil degradation and to increase a carbon sequestration in soils and their fertility.

The aim of future research is to discover the differences between the inter aggregate pore structure of water stable and non stable aggregates [8]. Additional physico-chemical parameters obtained by chromatography analysis are going to be used.

5. Conclusions

Recent advances in computed tomography and digital image processing algorithms provide technologically advanced measurement 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 of the aggregates from the different types of soil.

A more detailed analysis can be obtained by deriving various methods to quantify the pore structure and developing a pore size-distribution curve to predict retention properties. The proposed algorithm, based on image analysis and kernel estimator methodology, is expected to be an effective technique for various agricultural studies.

Nowadays, kernel regression is a common tool for empirical studies in many research areas. This is also a consequence of the fact that kernel regression techniques are provided by many software packages.

The kernel regression approach is also common in image processing and reconstruction methods. Quite recently, kernel regression has experienced a kind of revival in Earth sciences on estimating some characteristics of the distribution in nonparametric models [11].

References

[1] Charytanowicz M., Kulczycki P., *Nonparametric Regression for Analyzing Correlation between Medical Parameters*, [in:] Pietka E., Kawa J. (eds), *Advances in Soft Computing – Information Technologies in Biomedicine*, Springer-Verlag, Berlin, Heidelberg 2008.

[2] Clark R.M., *Non-Parametric Estimation of a Smooth Regression Function*, Journal of the Royal Statistical Society, Series B 39, 1977, 107-113.

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

[4] Gonet S., Czachor H., *Organic carbon and humic substances fractions in soil aggregates* 16th Meeting of the International Humic Substances Society, Functions of the Natural Organic Matter in Changing Environment, China, Hangzhou, 9-14.09.2012, 215-217, 2012.

[5] Gonzalez R.C., Woods R.E., *Digital Image Processing*, Prentice-Hall Inc., New Jersey 2002.

[6] Kowalski P., *Procedura ekstrakcji cech z obrazu twarzy dla potrzeb systemu biometrycznego*, Technical Transactions, vol. 1-AC/2012, Cracow University of Technology Press, 2012, 55-79.

[7] Kravchenko, A., Chun H.C., Mazer M., Wang W., Rose J. B., Smucker, A., Rivers M., *Relationships between intra-aggregate pore structures and distributions of Escherichia coli within soil macro-aggregates*, Applied Soil Ecology, vol. 63, 2013, 134-142.

[8] 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.

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

[10] Kulczycki P., *Kernel estimators in industrial applications*, [in:] Prasad B. (ed) Soft Computing Applications in Industry, Springer-Verlag, Berlin 2008.
[11] Perzanowski K.A., Wołoszyn-Gałęza A., Januszczak M., *Indicative factors for European bison refuges in the Bieszczady Mountains*, Ann. Zool. Fennici 45, 2008, 347-352.
[12] Pratt W.K., *Digital Image Processing*, John Wiley and Sons, New York 2001.
[13] Silverman B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London 1986.
[14] Smith J.R., Chang S.F., *VisualSeek: A fully automated content-based image query system*, Proc. ACM Multimedia Conf., 1996, 87-98.
[15] Van der Weerden T., Klein C., Kelliher F., *Influence of pore size distribution and soil water content on N_2O response curves*, 19th World Congress of Soil Science, Soil Solutions for a Changing World, 1–6 August 2010, Brisbane, Australia 2010.
[16] Wand M.P., Jones M.C., *Kernel Smoothing*, Chapman and Hall, London 1994.
[17] Wayne L.W.C., *Mathematical Morphology and Its Applications on Image Segmentation*, MS thesis, Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan 2000.