INFORMATICA, 2016, Vol. 27, No. 2, 257–281

DOI: http://dx.doi.org/10.15388/Informatica.2016.84

Fractal-Based Methods as a Technique for Estimating the Intrinsic Dimensionality of High-Dimensional Data: A Survey

Rasa KARBAUSKAITĖ*, Gintautas DZEMYDA
Institute of Mathematics and Informatics, Vilnius University
Akademijos 4, LT-08663, Vilnius, Lithuania
e-mail: rasa.karbauskaite@mii.vu.lt, gintautas.dzemyda@mii.vu.lt

Received: December 2015; accepted: April 2016

Abstract. The estimation of intrinsic dimensionality of high-dimensional data still remains a challenging issue. Various approaches to interpret and estimate the intrinsic dimensionality are developed. Referring to the following two classifications of estimators of the intrinsic dimensionality – local/global estimators and projection techniques/geometric approaches – we focus on the fractal-based methods that are assigned to the global estimators and geometric approaches. The computational aspects of estimating the intrinsic dimensionality of high-dimensional data are the core issue in this paper. The advantages and disadvantages of the fractal-based methods are disclosed and applications of these methods are presented briefly.

Key words: high-dimensional data, intrinsic dimensionality, topological dimension, fractal dimension, fractal-based methods, box-counting dimension, information dimension, correlation dimension, packing dimension.

1. Introduction

In real applications, we confront with data that are of a very high dimensionality. For example, in image analysis, each image is described by a large number of pixels of different colour. The analysis of DNA microarray data (Kriukienė *et al.*, 2013) deals with a high dimensionality, too. The analysis of high-dimensional data is usually challenging. The dimensionality reduction and visualization methods allow the human-based decisions in discovering knowledge hidden in multidimensional data sets (Borg and Groenen, 2005; Žilinskas and Žilinskas, 2009; Dzemyda *et al.*, 2013). Although data are considered in a high-dimensional space, in fact they are often either points of a nonlinear manifold of some lower dimensionality or points close to that manifold. An easily understandable example of such a manifold is a plane in a three-dimensional space. Thus, one of the major problems is to find the exact dimensionality of the manifold. In the example above, this dimensionality is equal to two. It is reasonable to reduce the dimensionality of the data

*Corresponding author.

set to that of a manifold. Therefore, the problem is to disclose the manifold dimensionality, i.e. the intrinsic dimensionality of the analysed data. The intrinsic dimensionality of a data set is the minimum number of free variables (features) needed to represent the data without information loss (Camastra, 2003). In this field, the terms 'dimensionality' and 'dimension' are often used as synonyms.

Recently, a lot of manifold learning methods have been proposed to solve the problem of nonlinear dimensionality reduction. Important manifold learning algorithms include isometric feature mapping (ISOMAP) (Tenenbaum *et al.*, 2000), locally linear embedding (LLE) (Roweis and Saul, 2000; Saul and Roweis, 2003), Laplacian eigenmaps (LE) (Belkin and Niyogi, 2003), Hessian LLE (Donoho and Grimes, 2005), etc. They all assume data to be distributed on an intrinsically low-dimensional manifold and reduce the dimensionality of data by investigating the intrinsic structure of data. However, all manifold learning algorithms require the intrinsic dimensionality of data as a key parameter for implementation. In recent years, ISOMAP and LLE have become of great interest. They avoid a nonlinear optimization and are simple to implement. However, both ISOMAP and LLE methods require the intrinsic dimensionality d of the data set and the neighbourhood parameter k of data points. The ways of selecting the value of the parameter k are proposed and investigated in Kouropteva *et al.* (2002), Karbauskaitė *et al.* (2007, 2008, 2010), Karbauskaitė and Dzemyda (2009). If the value of d is set larger than the intrinsic dimensionality really is, much redundant information will also be preserved; if it is set smaller, some useful information of the data could be lost during the dimensionality reduction (Fan *et al.*, 2009). The discussion above has highlighted one of many reasons why the intrinsic dimensionality estimation is very important for dimensionality reduction.

According to the statistical learning theory (Vapnik, 1998), the generalization capability of the classifiers depends on the intrinsic dimensionality: classification performance may be improved when using the data points of smaller dimensions. When using an autoassociative neural network for a nonlinear feature extraction (Kirby, 2001), the intrinsic dimensionality can propose a proper number of hidden neurons. For reliable predictions, the model order in a time series may be fixed by intrinsic dimensionality.

Due to increased interest in dimensionality reduction and manifold learning, a lot of techniques have been proposed in order to estimate the intrinsic dimensionality of a data set (Camastra, 2003; Brand, 2003; Costa and Hero, 2004; Kégl, 2003; Hein and Audibert, 2005; Levina and Bickel, 2005; Weinberger and Saul, 2006; Fan *et al.*, 2009, 2013; Yata and Aoshima, 2010; Mo and Huang, 2012; Einbeck and Kalantan, 2013; He *et al.*, 2014).

After passing in review a plenty of articles (van der Maaten, 2007; Fan *et al.*, 2009, 2013; Yata and Aoshima, 2010; Einbeck and Kalantan, 2013), two classifications of estimators of the intrinsic dimensionality are determined:

1. Local/global estimators (van der Maaten, 2007; Einbeck and Kalantan, 2013),
2. Projection techniques/geometric approaches (Fan *et al.*, 2009, 2013; Yata and Aoshima, 2010).

As mentioned above, the intrinsic dimensionality of a data set is usually defined as some integer number of features. However, fractional measures of the intrinsic dimension-

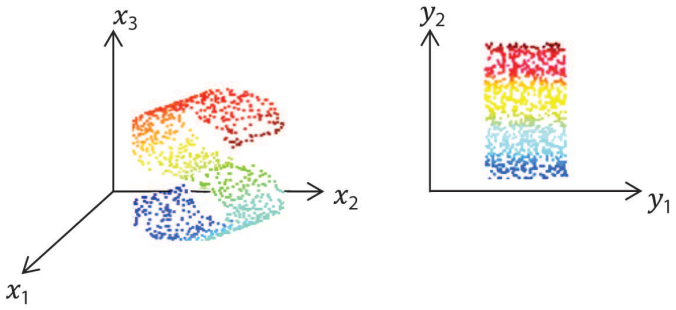

Fig. 1. The two-dimensional manifold (b) embedded in the three-dimensional space (a).

ality find interest and are employed. A group of fractal-based methods that estimate the intrinsic dimensionality by fractional values is developed (Grassberger and Procaccia, 1983; Camastra, 2003; Kégl, 2003). These fractional values may be applied to describe the complexity of the analysed object as well as to reduce the initial dimensionality of data. It is the main reason that we concentrate in this paper namely on the fractal-based methods that are assigned to the global estimators and geometric approaches following the classifications above. The computational aspects of estimating the intrinsic dimensionality of high-dimensional data are on the focus.

2. Intrinsic Dimensionality: Concept and Phenomenon

The dimension of an object is a topological measure of the size of its covering properties. It may be interpreted as the number of coordinates needed to specify a point on the object. For example, a line is a one-dimensional object, a rectangle is two-dimensional, while a cube is three-dimensional. In general, the object is described by some data set – a set of points, consisting of n coordinates, i.e. the object is n-dimensional. However, this fact does not necessarily imply that its actual dimension is n. Here, the necessity of the concept of the intrinsic dimension (dimensionality) appears.

The *intrinsic dimensionality* of a data set is usually defined as the minimal number of features or latent variables necessary to describe the data (Lee and Verleysen, 2007). Latent variables are still often called as degrees of freedom of a data set (Tenenbaum *et al.*, 2000; Lee and Verleysen, 2007). Let the dimensionality of the analysed data be n. High-dimensional data sets can have meaningful low-dimensional structures hidden in the observation space, i.e. the data are of a much lower intrinsic dimensionality d ($d \ll n$). In more general terms, following Fukunaga (1982), a data set $X \subset R^n$ is said to have the intrinsic dimensionality equal to d, if its elements lie entirely within a d-dimensional subspace of R^n (where $d < n$).

Both definitions of the intrinsic dimensionality are quite of a general nature and are not exact. For clarity, let us analyse two examples that are often used in the intrinsic dimensionality research (Karbauskaitė *et al.*, 2011; Karbauskaitė and Dzemyda, 2014, 2015). A simple example is given in Fig. 1. The data set consists of three-dimensional

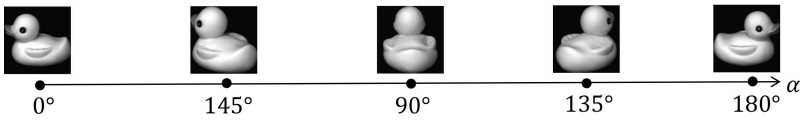

Fig. 2. High-dimensional data, corresponding to the greyscale pictures of a rotated duckling, mapped in the one-dimensional space.

data points ($n = 3$) that lie on a nonlinear two-dimensional S-shaped manifold which is a two-dimensional subspace of R^3 (Fig. 1(a)). Therefore, the intrinsic dimensionality is equal to two and the data points could be transformed exactly into a two-dimensional space (plane) (Fig. 1(b)). A more complex example is related to the set of images of a moving object. Usually, the image is represented by a high-dimensional point, the dimensionality of which depends on the number of pixels in the image. So, the dimensionality of these data is very large. Let us analyse high-dimensional data, obtained from the set of greyscale pictures of a rotated duckling ($n = 16\,384$) (Nene *et al.*, 1996). Since a duckling was gradually rotated at a certain angle on the same plane, i.e. without turning the object itself, the angle may be the key feature that describes a particular picture of the set (Fig. 2). Therefore, the object in the set of pictures has only one degree of freedom, i.e. the intrinsic dimensionality of these data may be equal to one. Thus, high-dimensional data points can be represented in the one-dimensional space and such a representation will be useful in making a decision on the content of pictures (Karbauskaitė *et al.*, 2011; Karbauskaitė and Dzemyda, 2014, 2015). From the examples above we can conclude that the intrinsic dimensionality may be interpreted differently.

The concept of intrinsic dimensionality is closely related to the theory of topological spaces (Buskes and van Rooij, 1997). A topological space is a set of points, along with a set of neighbourhoods for each point, that satisfy a set of axioms relating points and neighbourhoods. The definition of a topological space relies only upon the set theory and is the most general notion of a mathematical space that allows for the definition of concepts such as continuity, connectedness, and convergence. Other spaces, such as manifolds and metric spaces, are specializations of topological spaces with extra structures or constraints.

The intrinsic dimensionality of a topological space is called the *Lebesgue covering dimension*, also known simply as a *topological dimension* (Weisstein, 2003).

The definition of the topological dimension given in Kégl (2003), Lee and Verleysen (2007) requires some additional notions. Given a topological space X, the *covering* of a subset S is a collection C of open subsets in X whose union contains S. A *refinement* of the covering C of S is another covering $\acute{C}$ such that each set in $\acute{C}$ is contained in some set in C. The following definition is based on the observation that an n-dimensional set can be covered by open balls such that each point belongs to maximum $(n + 1)$ open balls. Taking into account the above notions, the topological dimension may be defined as follows. A subset S of a topological space X has the *topological dimension* d_T (also known as the Lebesgue covering dimension), if every covering C of S has a refinement $\acute{C}$ in which every point of S belongs to at most $(d_T + 1)$ sets in $\acute{C}$, and d_T is the smallest such integer (Kégl, 2003; Lee and Verleysen, 2007). From this definition, a particular case follows: the Lebesgue covering dimension of the usual Euclidean space R^n is n.

Let us analyse the examples. Consider some arbitrary open cover (covering) of the unit circle. This covering has a refinement consisting of a collection of open arcs. The circle has dimension 1, by this definition, because any such covering can be further refined to the stage where a given point x of the circle is contained in 2 arcs at most. That is, whatever collection of arcs we begin with, some can be discarded, so that the remainder still covers the circle, but with simple overlaps. Similarly, consider a unit disk in the two-dimensional plane. It is not difficult to visualize that any covering can be refined so that any point of the disk is contained in no more than three sets. A solid cube has the topological dimension of three because in any decomposition of the cube into smaller bricks all points belong to at least four $(3 + 1)$ bricks, and it is possible to construct such a decomposition, where all points belong exactly to four bricks.

A more comprehensible definition of the *topological dimension* with explanatory examples is presented in Broomhead (1985). The topological dimension d_T requires only that the continuity have some meaning on the set. The definition is a recursive one: the topological dimension of X is $d_T = 1 + \acute{d}_T$, where $\acute{d}_T$ is the topological dimension of a set the removal of which would divide X. A point is considered to have $d_T = 0$. It follows from this definition that a line has $d_T = 1$, since it is divided by the removal of a point. Similarly, a surface has $d_T = 2$, since it is divided by the removal of a line. These examples illustrate that d_T is coincident with an intuitively reasonable idea of dimension, in particular, note that it is always an integer.

Another definition of *topological dimension* was given by Brouwer in 1913 (Heyting and Freudenthal, 1975). A topological dimension is the basis dimension of the local linear approximation of the hypersurface where the data reside, i.e. the tangent space. For example, if the data set lies on a d-dimensional manifold, then it has a d-dimensional tangent space at every point in the set and the topological dimension, according to Brouwer, is d. For instance, let us analyse a surface of a ball. It is a sphere that can be realized in three dimensions, i.e. its points are embedded in R^3. Such a sphere has a two-dimensional tangent space at every point and may be viewed as a two-dimensional manifold. So, its topological dimension is two: only two coordinates, i.e. longitude and latitude, are necessary to define any point of the sphere. But the intrinsic dimensionality of the full sphere is three.

Fractional measures are not allowed in view of a topological dimension. It was impossible to define the dimensionality of strange geometric objects such as space-filling curves using the concept of ordinary topological dimension. An example can be the Hilbert curve (Fig. 3). Although a topological dimension of the Hilbert curve (as well as of any other curve) is one, a topological dimension of the filled square is two. Filling of the square by the Hilbert curve converges to the full-filled square, when the number of iterations of the Hilbert curve grows, i.e. the space-filling curve is a one-dimensional object that evolves iteratively and progressively fills a square – a two-dimensional object (see Fig. 3). Therefore, a new type of dimension, i.e. a *fractal dimension*, was introduced. While a topological dimension always yields an integer value, the so-called fractal dimension does not have to be an integer and it is often a real number (Lee and Verleysen, 2007).

The discussion above indicates that the concept of the intrinsic dimensionality should not be equalized to the topological dimension only. Various approaches are possible to

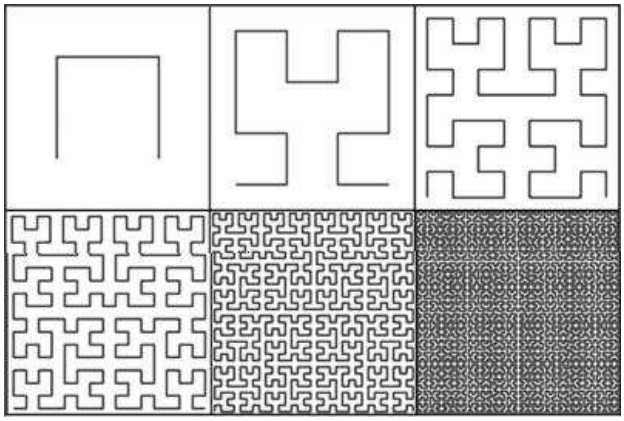

Fig. 3. Six iterations of the Hilbert curve construction.

interpret and estimate the intrinsic dimensionality. These approaches combine ideas of the topological and fractal dimensions, but there are approximate methods for estimating the intrinsic dimensionality that interpret this dimension quite differently as compared with the topological and fractal ones. We review the approaches for estimating the intrinsic dimensionality and show their advantages and disadvantages briefly below.

There are attempts at the direct estimation of the topological dimension of a manifold (Broomhead *et al.*, 1987). However, the direct estimation encounters some essential troubles:

- It is computationally difficult to estimate the topological dimension, if only a finite set of points is available (Lee and Verleysen, 2007). It is a common case in the exploratory data analysis.
- If a data set consists of points of a certain manifold, then its intrinsic dimensionality is an integer number that is coincident with the topological dimension. In the general case, when a data set does not belong to some manifold, the intrinsic dimension may not be coincident with the topological one and may take even the fractional values.
- A topological dimension does not provide details on the form of the object. For instance, the topological dimension of a straight line and a crooked line is the same, i.e. equal to 1. This leads us to the necessity to widen the conception of the intrinsic dimension and to its extension, e.g. to the fractal dimension, allowing more possibilities to analyse the shapes.
- Given some covering C of S, the search through all possible refinements $\acute{C}$ of C is a daunting and infinite task.

Hence, practical methods use various other notions of the intrinsic dimensionality (Lee and Verleysen, 2007; Camastra, 2003). The most usual ones are related to the fractal dimension, the estimators of which (fractal-based methods) are explored in Section 4. The fractal dimension is a parameter that characterizes how densely a fractal fills the space. However, the fractal dimension has a number of different interpretations: capacity (box-counting) dimension (Ott, 1993), correlation dimension (Grassberger and Procaccia, 1983), packing dimension (Kégl, 2003), etc. Several estimators of the intrinsic

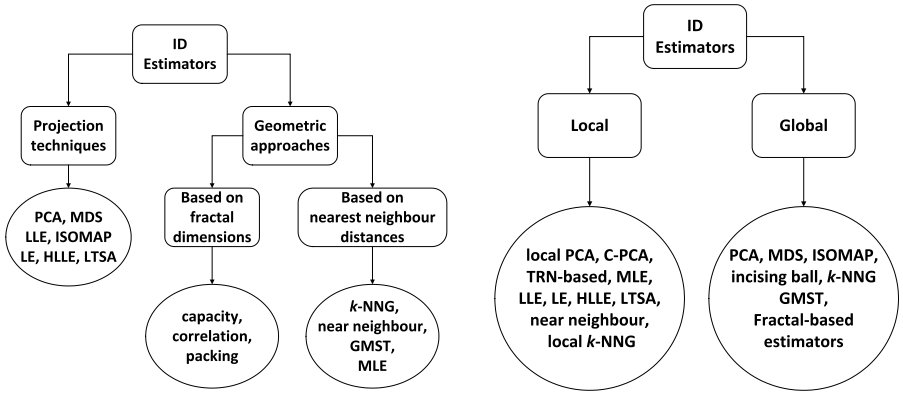

Fig. 4. Classifications of the intrinsic dimensionality (ID) estimators: (a) projection/geometric; (b) local/global.

dimensionality refer to the distances between the nearest neighbours (Camastra, 2003; Costa and Hero, 2004, 2005; Levina and Bickel, 2005; Fan *et al.*, 2009). Other notions of the intrinsic dimensionality are based on dimensionality reduction methods. These estimators are related to the principal component analysis (PCA) (Jolliffe, 1986) and various PCA modifications (Fukunaga and Olsen, 1971; Hastie and Stuetzle, 1988; Fan *et al.*, 2013) or are based on the trial-and-error approach, e.g. using multidimensional scaling (MDS) (Cox and Cox, 2001; Borg and Groenen, 2005; Dzemyda *et al.*, 2013) or nonlinear manifold learning methods such as locally linear embedding (LLE) (Roweis and Saul, 2000; Saul and Roweis, 2003), isometric feature mapping (ISOMAP) (Tenenbaum *et al.*, 2000), Laplacian eigenmaps (LE) (Belkin and Niyogi, 2003), Hessian LLE (HLLE) (Donoho and Grimes, 2005), local tangent space analysis (LTSA) (Zhang and Zha, 2004), etc. (Lee and Verleysen, 2007).

After passing in review plenty of articles (Fan *et al.*, 2009, 2013; Yata and Aoshima, 2010), it is possible to categorize intrinsic dimensionality estimating methods into two classes: projection techniques and geometric approaches (Fig. 4(a)). Projection techniques project the data into a low-dimensional space. The intrinsic dimensionality may be estimated by comparing the projections to the space of various dimensions with the initial data set. Such methods are: PCA and its various modifications (e.g. algorithm of Fukunaga and Olsen, 1971, and topology representing network based methods of Bruske and Sommer, 1998; Frisone *et al.*, 1995), MDS, nonlinear manifold learning methods like LLE, ISOMAP, LE, HLLE, LTSA, etc. Geometric techniques find the intrinsic dimensionality by investigating the geometric structure of the data. The geometric methods are mostly based on fractal dimensions (capacity dimension, correlation dimension, packing dimension, etc.) or nearest neighbour distances: the near neighbour algorithm (Pettis *et al.*, 1979; Verveer and Duin, 1995), incising ball method (Fan *et al.*, 2009), k-nearest neighbour graphs (k-NNG) method (Costa and Hero, 2003, 2005), geodesic minimal spanning tree (GMST) method (Costa and Hero, 2004), the maximum likelihood estimator (MLE) (Levina and Bickel, 2005), etc.

The estimators of the intrinsic dimensionality may be classified in the other way (Fig. 4(b)): local and global methods (van der Maaten, 2007; Einbeck and Kalantan, 2013). Local methods estimate the dimensionality at each data point from its local neighbourhood and then compute the average over the local estimates of intrinsic dimensionality. Local methods are as follows: Fukunaga–Olsen's algorithm, the near neighbour algorithm, topology representing network based methods, the maximum likelihood estimator (MLE), nonlinear manifold learning methods like LLE, LE, HLLE, LTSA, etc. Global methods estimate the dimensionality using the whole data set, assuming that the data set has the same dimension throughout. Among global methods, the most popular ones are PCA, MDS, ISOMAP, the incising ball method, the k-nearest neighbour graphs (k-NNG) method, the geodesic minimal spanning tree (GMST) method, the fractal-based methods: the correlation dimension, capacity dimension, and packing dimension estimators, etc.

The local intrinsic dimensionality may be estimated by adopting the global estimators. Fukunaga–Olsen's algorithm is one of the attempts, based on PCA, and it is called a local PCA method. In Fan *et al.* (2013), another PCA-based method (C-PCA) is developed for local intrinsic dimension estimation. This method works first by finding a minimal cover of the data set, then performing PCA locally on each subset in the cover to obtain local intrinsic dimension estimations and finally giving the estimation result as the average of the local estimations. In Costa and Hero (2005), Carter *et al.* (2010), a method to estimate the local dimensionality associated with each point in a data set is proposed. This method uses a global dimensionality estimator, based on k-NNG, together with an algorithm for computing neighbourhoods in data with similar topological properties.

3. Fractal Dimension

The terms *fractal* and *fractal dimension* were first introduced by mathematician Benoit Mandelbrot in 1975 (Mandelbrot, 1975, 1977, 1983). He has noticed that the key features of fractals are: self-similarity which implies that the object looks similar to its zoomed part, symmetry, irregularity locally and globally that is not easily described in the traditional Euclidean geometry. One often cited description that Mandelbrot suggested to describe geometric fractals is a rough or fragmented geometric shape that can be split into parts, each of which is (at least approximately) a reduced-size copy of the whole (Mandelbrot, 1983). A fractal is a never-ending pattern. Fractals are infinitely complex patterns that are self-similar across different scales. They are created by repeating a simple process over and over in an ongoing feedback loop.

A fundamental characteristic of fractal objects is that their measured metric properties, such as length or area, are a function of the scale of measurement (Lopes and Betrouni, 2009). A classical example to illustrate this property is the 'length' of a coastline (Mandelbrot, 1967). Figures 5 and 6 illustrate the coastline of the Koch island. As shown in Fig. 6, it is built by starting from an equilateral triangle, removing the inner third of each side, replacing it by two edges of a three-times-smaller equilateral triangle, and then repeating the process indefinitely (Lee and Verleysen, 2007). As pointed out by Mandelbrot,

Fig. 5. Construction of the Koch curve.

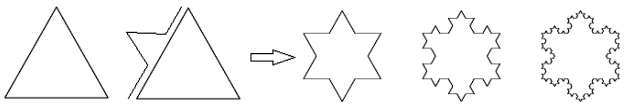

Fig. 6. Construction of the Koch island (snowflake) from the triangle – the Koch island consists of three Koch curves.

the length of such a coastline is different depending on the length ruler used to measure it. This paradox is known as the coastline paradox: the shorter the ruler, the longer the length measured is (Lee and Verleysen, 2007).

Fractals are found in many places in nature, including ferns, mountains, bacteria, snowflakes, clouds, and coastlines. Some parts of the human body, such as the lungs and trabecular bone, also appear to grow in the form of fractals. Other elements of the body, such as brain tissues or tumours, may also exhibit fractal characteristics (Zook and Iftekharuddin, 2005).

The traditional Euclidean geometry may analyse only smooth lines and surfaces such as circles, ellipses, spheres, ellipsoids, etc. A fractal geometry is a development (extension) of the classical Euclidean geometry. It enables us to create exact mathematical models of physical structures (from ferns to galaxy) (Valantinas, 1999). So, the main attraction of the fractal geometry stems from its ability to describe the irregular or fragmented shape of natural features as well as other complex objects that the traditional Euclidean geometry fails to analyse (Lopes and Betrouni, 2009).

In 1975, Benoit Mandelbrot introduced a fractional dimension refusing the conception of a topological dimension. While a topological dimension always yields an integer value, a fractal dimension must not be an integer and it is often a real number (Zook and Iftekharuddin, 2005; Lee and Verleysen, 2007). A fractal dimension characterizes how densely a fractal fills the space. For sets describing ordinary geometric shapes, the theoretical fractal dimension equals the set's topological dimension. Mandelbrot (1975) defined a fractal set as a set for which the fractal dimension (also called as Hausdorff or Hausdorff–Besicovitch dimension) is greater than its topological dimension (d_T). The *Hausdorff dimension* d_H is defined as:

$$d_H = \frac{\log N}{\log r}, \tag{1}$$

where N is the number of self-similar objects created from the original object when it is divided by r, i.e. each object is r times smaller than the original one (r is a magnification factor).

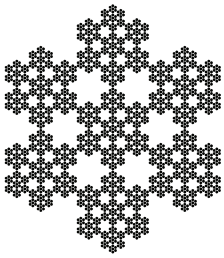

Fig. 7. Hexaflake.

Example. Let us calculate the fractal dimension of the Koch curve (see Fig. 5). The number of line segments in the Koch curve is 4, and each line segment is replaced by a copy of the original reduced by a scale of 1/3. Thus, the theoretical fractal dimension of the Koch curve is $\frac{\log 4}{\log 3} = 1.2619$.

A topological dimension does not provide details about the form of the object. For instance, the topological dimension of a straight line and a crooked line is the same, i.e. equal to 1 (Lopes and Betrouni, 2009). The Koch curve (Fig. 5), the fractal dimension of which is near to 1, i.e. 1.2619, behaves quite so as an ordinary line, but a hexaflake (Fig. 7), the fractal dimension of which is 1.7712 (Lai, 2012), winds convolutedly through a space nearly like a surface. An approximate fractal dimension of the surface of human brain is 2.79, and that of human lung surface is 2.97 (Vrobel, 2011).

The example with the Koch curve was chosen for clarity, and the scaling unit and ratios were known ahead of time. In practice, however, fractal dimensions can be determined using approximate techniques, because the Hausdorff dimension (1) is not computable in this form in most cases. Each of the approximate techniques has its own theoretical basis and uses different algorithms to estimate the parameter N (Lopes and Betrouni, 2009). Therefore, different types of fractal dimension are known, such as capacity (box-counting) dimension, information dimension, correlation dimension, generalized Rényi dimension, packing dimension, Higuchi dimension, uncertainty exponent, etc. Although for some classic fractals all these dimensions are coincident, in general they are not equivalent.

The techniques to estimate a fractal dimension are called fractal-based methods. They have been discussed in Section 4.

4. Fractal-Based Methods

The foundation for a definition of the fractal dimension is laid by Hausdorff (1919). Therefore, the fractal dimension is also called as the *Hausdorff dimension* (Mandelbrot, 1983). The Hausdorff dimension is also used as a generic name for different mathematical definitions of fractal dimension (Eckmann and Ruelle, 1985).

In order to define the Hausdorff dimension d_H of a set $X \subset R^n$, it is necessary to introduce the quantity:

$$\Gamma_H^d(r) = \inf_{\{s_i\}} \sum_i (r_i)^d, \tag{2}$$

where the set X is covered by sets s_i with a variable diameter r_i (the largest distance between any two points in s_i) and all diameters satisfy $0 < r_i \leqslant r$ (Camastra, 2003). Denote $\{s_i\}$ as the cover of the set X by sets s_i: $X \subset \bigcup_i s_i$.

In the geometric sense, the sum $\sum_i (r_i)^d$ is the 'volume' of the number of d-dimensional cubes needed to cover X. That is, we look for such a collection $\{s_i\}$ of covering sets s_i with diameters less than or equal to r that minimizes the sum in (2) and we denote the minimized one by a sum $\Gamma_H^d(r)$. The *d-dimensional Hausdorff measure* is then defined as:

$$\Gamma_H^d = \lim_{r \to 0} \Gamma_H^d(r). \tag{3}$$

The Hausdorff measure Γ_H^d means the 'volume' of X, if it is in a d-dimensional space. Γ_H^0 is the number of points of the set, Γ_H^1 is the length of a curve, i.e. one-dimensional measure, Γ_H^2 is the area of a surface, i.e. two-dimensional measure, Γ_H^3 is the volume of a body, i.e. three-dimensional measure, and generalizing Γ_H^d is the d-dimensional measure of a set X.

Hausdorff proved that for every set X,

$$\Gamma_H^d = \begin{cases} +\infty, & \text{if } d < d_H; \\ 0, & \text{if } d > d_H. \end{cases}$$

This critical value $d = d_H$ is called the *Hausdorff dimension* of the set.

For example, if we tried to cover a two-dimensional square with one-dimensional lines, we would need an infinity of lines. Then one-dimensional 'volume' of a square (the total length of lines) will be ∞. If we tried to cover a two-dimensional square with three-dimensional cubes, then its three-dimensional 'volume' (the total volume of cubes) will be 0. In other words, if the set X is a smooth surface of finite area situated in a three-dimensional space, then Γ_H^2 is the area of the set, while $\Gamma_H^d = +\infty$ for $d < 2$, and $\Gamma_H^d = 0$ as $d > 2$.

Despite the origin of a fractal dimension from the Hausdorff dimension, various definitions of a fractal dimension have been proposed or derived from the Hausdorff dimension (Eckmann and Ruelle, 1985). The reason is that it is difficult to evaluate the Hausdorff dimension numerically because of the necessity to find the infimum over all coverings in (2) (Theiler, 1990). In practical applications, it is substituted by other fractal dimensions, i.e. the box-counting dimension (also known as the Minkowski–Bouligand dimension or capacity dimension), information dimension, correlation dimension, generalized Rényi dimension, packing dimension, Liapunov dimension, Higuchi dimension, uncertainty exponent, etc.

A generalized Rényi dimension is presented below. Then the box-counting, information, and correlation dimensions are introduced, because they follow from the Rényi dimension.

4.1. *Rényi Dimensions*

The Rényi dimension, also called the generalized dimension or D_q dimension, was introduced by Rényi (1960, 1961) in 1960 as a tool to analyse various problems in information theory (Olsen, 2007). In fact, it is a family of dimensions.

The generalized dimension of order q is defined by the formula:

$$D_q = \frac{1}{q-1} \lim_{r \to 0} \frac{\log \sum_{i=1}^{N(r)} (p_i)^q}{\log r}, \tag{4}$$

where $N(r)$ is the number of 'boxes' of size r (hypercubes with the edge length r) needed to cover the data set X, and $p_1, p_2, \ldots, p_{N(r)}$ are probability measures of these 'boxes'.

The generalized dimension takes into account the number of points of X in the 'box'. Let β_i denote the ith 'box', and p_i be a normalized measure of this 'box', then the probability for a randomly chosen point of X is in the ith 'box' β_i:

$$\sum_{i=1}^{N(r)} p_i = 1. \tag{5}$$

The probability p_i is usually estimated by counting the number of points that are in β_i and dividing by the total number of points (Theiler, 1990).

Grassberger and Procaccia (1983) have proved analytically that the box-counting d_{box}, information d_{inf}, and correlation d_{cor} dimensions are special cases of the generalised dimension D_q, when $q = 0$, 1, and 2, respectively (see Maggi, 2002, and Lee and Verleysen, 2007, for details). The generalized dimension D_q is monotonically decreasing or at least monotonically non-increasing while q increases. As a consequence, it follows that $d_0 \geqslant d_1 \geqslant d_2$, i.e. $d_{\text{box}} \geqslant d_{\text{inf}} \geqslant d_{\text{cor}}$.

The box-counting and correlation dimensions are most popular among all the fractal dimensions. However, the information dimension is usually employed by physicists as well as in information theory. All these three dimensions are described in detail below.

4.2. *The Box-Counting Dimension*

The box-counting dimension (also known as the capacity dimension, Minkowski–Bouligand dimension or Kolmogorov capacity/dimension) was proposed by Kolmogorov in 1958. The interpretation of the box-counting dimension may be as follows. The basic idea arises when considering the length, area, and volume of Euclidean objects such as a line, plane, and cube. In one dimension, let us consider a curve and a ruler of the length r. If one counts the number of rulers $N(r)$, required to cover the curve as r decreases, then the relationship will be as follows: $N(r)$ is inversely proportional to r, i.e. $N(r)$ is proportional to $1/r^1$. Similarly in two dimensions, if one counts the number of squares $N(r)$ of a side length r, required to cover a surface, then the relationship will be as follows: $N(r)$ is proportional to $1/r^2$. In three dimensions, if one counts how many cubes $N(r)$ of the side length r are required to fill the volume, $N(r)$ is proportional to $1/r^3$.

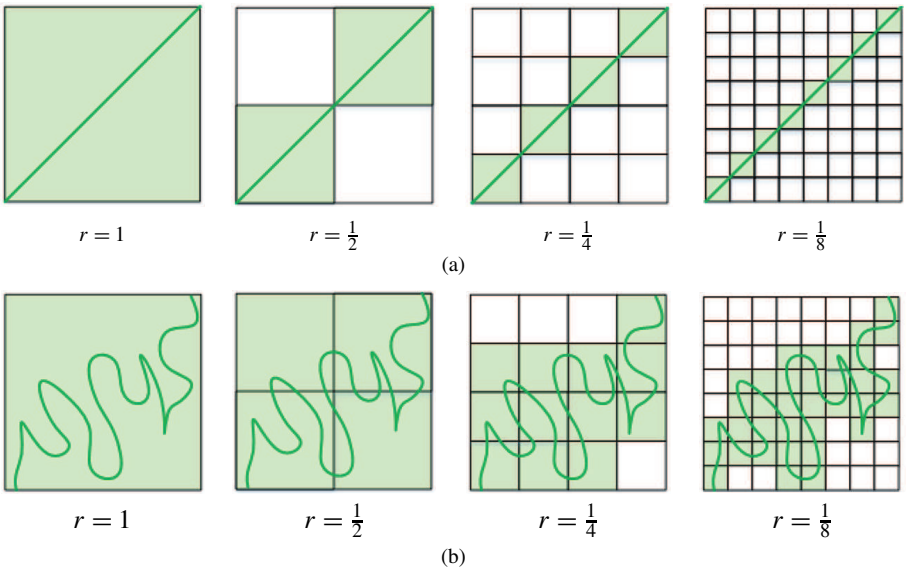

Fig. 8. Box-counting on a plane: (a) linear segment, (b) curve.

However, we often deal with objects that are embedded into the spaces of higher dimensionality. For example, a linear segment or a curve is drawn on the plane (Fig. 8). We need to cover this linear segment/curve with $N(r)$ squares of the side length r. $N(r)$ is proportional to $1/r^d$. A question arises as to what the values of d are in these cases. It is obvious from Fig. 8(a) that the number of squares $N(r)$ is inversely proportional to the side length r, i.e. $N(r)$ is proportional to $1/r^1$. So, in the case of a linear segment, $d = 1$. Figure 8(b) shows that $N(r)$ is not proportional to $1/r^1$ or $1/r^2$, i.e. as $d = 1$ or $d = 2$. It means that, in this case, d is larger than 1 and smaller than 2, i.e. $1 < d < 2$ and the value of d depends on the curve form.

Let us consider some object, the dimensionality d of which is less than n ($d < n$) and which is embedded in an n-dimensional space. If one counts the number $N(r)$ of n-dimensional hypercubes of the side length r, required to cover the object, then $N(r)$ is proportional to $1/r^d$:

$$N(r) \propto \frac{1}{r^d}. \tag{6}$$

From (6), it follows:

$$d \propto \frac{\log N(r)}{\log \frac{1}{r}}. \tag{7}$$

In the general case, suppose we have a set X in an n-dimensional space. Imagine that we cover the space with equal n-dimensional hypercubes with a side length r, and count how many hypercubes contain points of the set, say $N(r)$. Then the box-counting

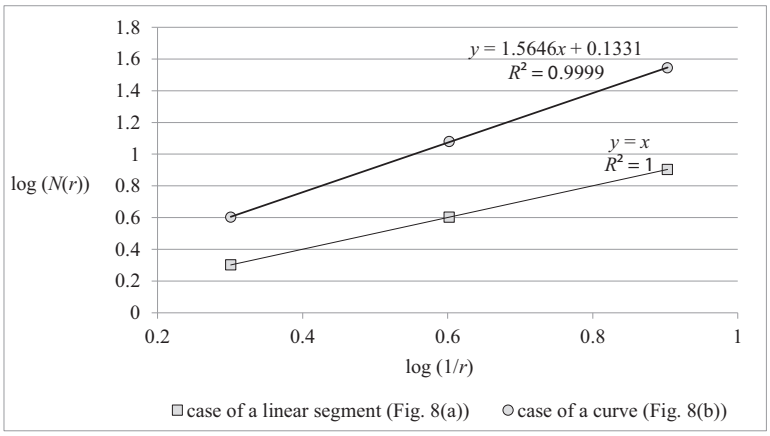

Fig. 9. Linear regression lines $y = y(x)$ approximating the points $\log N(r)$ versus $\log(1/r)$.

dimension of a data set X is defined as follows:

$$d_{\text{box}} = \lim_{r\to 0} \frac{\log N(r)}{\log \frac{1}{r}} = -\lim_{r\to 0} \frac{\log N(r)}{\log r}. \tag{8}$$

The difference of the box-counting dimension from the Hausdorff one is that the set X is covered by sets with a variable diameter in the case of the Hausdorff dimension.

There are several ways to estimate the box-counting dimension. The theoretical one is based on formula (8) and can be used when a continuous data set is analysed, for example, a straight line, etc. Suppose that the length of a linear segment, given in Fig. 8(a), is L and this set is put on the evenly-spaced grid of size r. Then we need $N(r) = L/(r\sqrt{2})$ boxes to cover the entire segment. The box-counting dimension is calculated as follows:

$$d_{\text{box}} = \lim_{r\to 0} \frac{\log N(r)}{\log \frac{1}{r}} = \lim_{r\to 0} \frac{\log \frac{L}{r\sqrt{2}}}{\log \frac{1}{r}} = \lim_{r\to 0} \frac{\log L - \log r - \log\sqrt{2}}{-\log r} = 1.$$

Thus, a linear segment is one-dimensional.

Since formula (8) for the box-counting dimension estimation includes a limit when a side of hypercubes tends to zero, this theoretical estimation is clearly impossible in practice, because the limit in (8) will not be achieved when a data set with a finite number of points is analysed. There are two ways to estimate the box-counting dimension in practice.

In the first way, one obtains the values of $N(r)$ for a variety of r and analyses the dependence of $\log N(r)$ on $\log(1/r)$. There should be a linear relationship between $\log N(r)$ and $\log(1/r)$. If one draws the linear regression line of best fit of $\log N(r)$ versus $\log(1/r)$, then the slope of that line is the estimate of the box-counting dimension. An example of a box-counting dimension estimation is presented in Fig. 9. The cases of a linear segment (Fig. 8(a)) and a curve (Fig. 8(b)) are analysed as $r = \frac{1}{2}$, $r = \frac{1}{4}$, and $r = \frac{1}{8}$. For a linear segment, the slope is equal to 1, thus $d_{\text{box}} = 1$, and, in the case of a curve, the slope is ap-

Table 1
Values of $\hat{d}_{\text{box}}(r_1, r_2)$ in the case of the curve (Fig. 8(b)).

r_1	r_2			
	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
1		2.0000	1.7925	1.7098
$\frac{1}{2}$	2.0000		1.5850	1.5646
$\frac{1}{4}$	1.7925	1.5850		1.5443
$\frac{1}{8}$	1.7098	1.5646	1.5443	

proximately equal to 1.5646, thus $d_{\text{box}} \approx 1.5646$. It is an illustrative example. For a more precise evaluation of the box-counting dimension, r values should be much smaller.

Another way to overcome the problem of a limit in (8) is to define the scale-dependent box-counting dimension (Lee and Verleysen, 2007):

$$\hat{d}_{\text{box}}(r_1, r_2) = \frac{\log N(r_2) - \log N(r_1)}{\log \frac{1}{r_2} - \log \frac{1}{r_1}} = -\frac{\log N(r_2) - \log N(r_1)}{\log r_2 - \log r_1}, \tag{9}$$

where $N(r_1)$ and $N(r_2)$ are the numbers of n-dimensional hypercubes of a side length r_1 and r_2, respectively, required to cover the data set X. Smaller r_1 and r_2 are better, because r should vanish in the estimation of d_{box} (8). Since $N(r)$ versus $1/r$ is an exact straight line in a log–log plot between two points $(\log(1/r_1), \log N(r_1))$ and $(\log(1/r_2), \log N(r_2))$, $\hat{d}_{\text{box}}(r_1, r_2)$ is the slope of this straight line.

Suppose r_1 and r_2 are equal to 1, $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$. Let us apply formula (9) to analyse the linear segment (Fig. 8(a)) and the curve (Fig. 8(b)). Then $\hat{d}_{\text{box}}(r_1, r_2) = 1$ for the linear segment. In the case of the curve, the values of $\hat{d}_{\text{box}}(r_1, r_2)$ are presented in Table 1. Combination of smaller values of r_1 and r_2 leads to the smaller values of $\hat{d}_{\text{box}}(r_1, r_2)$ and to the more exact value of the box-counting dimension.

There are several practical realizations/modifications of the box-counting method (Grassberger, 1990; Tolle *et al.*, 2003; Zook and Iftekharuddin, 2005; Lopes and Betrouni, 2009; Chaudhuri and Sarkar, 1995; Sandau and Kurz, 1997), etc. The differential box-counting (Chaudhuri and Sarkar, 1995), extended counting (Sandau and Kurz, 1997), and piecewise modified box counting (Zook and Iftekharuddin, 2005) are used in medical data mining.

Despite that efficient algorithms have been proposed, the box-counting dimension can be computed only for low-dimensional sets, because the algorithmic complexity grows exponentially with the set dimensionality (Camastra, 2003). Moreover, the amount of points in each box does not influence the calculation of d_{box}. Therefore, in order to better characterize data sets with heterogeneous structures, the information and correlation dimensions are used (Monteiro, 2013).

4.3. *Information Dimension*

The information dimension is mentioned here just for the sake of completeness. The term 'information dimension' reflects the information-theoretic origins of the concept (Rényi,

1970). It is calculated as follows:

$$d_{\inf} = \lim_{q\to 1} D_q = \lim_{r\to 0} \frac{\sum_{i=1}^{N(r)} p_i \log p_i}{\log r}. \tag{10}$$

Usually, the points of X are not spread out uniformly. Therefore, there are regions that are more often visited than others. As far as the probability p_i for a randomly chosen point of X to be in the ith box is seldom known when dealing with a finite number of samples, its evaluation remains difficult, except when all p_i are assumed to be equal, meaning that all occupied boxes have the same probability to be visited: $\forall i,\ p_i = 1/N(r)$. In this case, the information dimension becomes the box-counting dimension, i.e. $d_{\inf} = d_{\text{box}}$.

Let us consider a simple example to understand the meaning of the information dimension. Consider a curve (Fig. 8(b)). In (10), $p_i = L_i/L$ is the probability, where L_i is the length of the curve that falls into the ith box, and L is the total length of the curve. As $r = \frac{1}{4}$, $d_{\inf} \approx 1.7606$; as $r = \frac{1}{8}$, $d_{\inf} \approx 1.6973$.

4.4. *Correlation Dimension*

Due to the computational simplicity, the correlation dimension, introduced by Grassberger and Procaccia (1983), is successfully used to replace the box-counting dimension (Camastra, 2003; Fan *et al.*, 2009). Furthermore, it can be evaluated for smaller values of r (Einbeck and Kalantan, 2013). By fixing $q = 2$ in (4), we get the correlation dimension:

$$d_{\text{cor}} = \lim_{r\to 0} \frac{\log \sum_{i=1}^{N(r)} (p_i)^2}{\log r}. \tag{11}$$

Let us consider a curve (Fig. 8(b)) to understand the meaning of the correlation dimension. In (11), $p_i = L_i/L$ is the probability, where L_i is the length of the curve that falls into the ith box, and L is the total length of the curve. $d_{\text{cor}} \approx 1.7380$, as $r = \frac{1}{4}$; $d_{\text{cor}} \approx 1.6857$, as $r = \frac{1}{8}$.

However, for a discrete case, i.e. when the object is only known by a countable set of points, a slightly different definition of d_{cor} exists in terms of a correlation integral (Grassberger and Procaccia, 1983; Ding *et al.*, 1993). Grassberger and Procaccia (1983) suggest to measure the distance between every pair of points and then compute a correlation integral. The correlation integral $C(r)$ is defined to be the probability that a pair of points chosen randomly is separated by a distance less than or equal to r in a data set (Ding *et al.*, 1993). Let a data set X consist of m n-dimensional points $X_i = (x_{i1}, \ldots, x_{in})$, $i = \overline{1, m}$ ($X_i \in R^n$). The correlation integral $C(r)$ is defined as:

$$C(r) = \lim_{m\to\infty} \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} I\big(\|X_i - X_j\| \leqslant r\big), \tag{12}$$

where I is an indicator function: $I(\lambda)$ is 1, if condition λ holds, 0 otherwise. $\|X_i - X_j\|$ denotes the Euclidean distance between the data points X_i and X_j. So, summations count

the number of pairs of points the distance between which is shorter or equal to r. As the number of points m tends to infinity, and the distance r between them tends to zero, the correlation integral for small values of r is:

$$C(r) \propto r^d, \tag{13}$$

$$d \propto \frac{\log C(r)}{\log r}. \tag{14}$$

If the number of points is sufficiently large and evenly distributed, the value of d represents the correlation dimension (Lantos and Márton, 2011), i.e. the correlation dimension d_{cor} of a data set X is defined as:

$$d_{\text{cor}} = \lim_{r \to 0} \frac{\log C(r)}{\log r}. \tag{15}$$

The equivalence of (11) and (15) is shown in Grassberger and Procaccia (1983), Ding *et al.* (1993).

Since formula (15) for estimating the correlation dimension includes a limit towards zero of the distance r, the theoretical estimation is clearly impossible in practice, because for a finite number of points, a limit towards zero cannot be achieved.

Like in the case of the box-counting dimension, there are two ways to estimate the correlation dimension in practice. In both ways, the estimation procedure of the correlation dimension is analogous to that of box-counting (see Section 4.2): only $\log C(r)$ replaces $\log N(r)$. The first estimation procedure consists of plotting $\log C(r)$ versus $\log r$ and measuring the slope of a linear part of the curve. The other way is to define the scale-dependent correlation dimension (Lee and Verleysen, 2007):

$$\hat{d}_{\text{cor}}(r_1, r_2) = \frac{\log C(r_2) - \log C(r_1)}{\log r_2 - \log r_1}, \tag{16}$$

which is computed as the average slope of the curve in a log-log plot of $C(r)$ versus r. The values of r_1 and r_2 are set between the minimal and maximal pairwise distances, measured in the analysed data set. The best estimate of $\hat{d}_{\text{cor}}$ is obtained in the largest region, where the slope of $C(r)$ is almost constant in the log-log plot.

Several other ways of estimating the correlation dimension are developed. In Einbeck and Kalantan (2013), three methods are proposed to approximate the correlation integral in the limit of r towards zero: intercept method, slope method, and polynomial method. A kernel version of the correlation dimension method was introduced in Hein and Audibert (2005). This method works by replacing the indicator function I in the correlation integral $C(r)$ with a generalized kernel function $K(x, y)$ (Fan *et al.*, 2009).

4.5. *Packing Dimension*

In practical approaches, the box-counting dimension d_{box} is usually discarded due to the high computational cost of its estimation. Finding the covering number $N(r)$ even of a

finite set of data points is computationally difficult. In Kégl (2003), the box-counting dimension is modified to the so-called packing dimension: packing numbers are proposed to be used instead of covering numbers.

Given a metric space R^n with the distance metric $d(\cdot,\cdot)$, the set $X \subset R^n$ is said to be *r-separated*, if $d(X_i, X_j) \geqslant r$ for all distinct $X_i, X_j \in X$. The *r-packing number* $M(r)$ of a set $X \subset R^n$ is defined as the maximum cardinality of an r-separated subset of X (Kégl, 2003).

The basic inequality between packing and covering numbers is as follows: $N(r) \leqslant M(r) \leqslant N(r/2)$.

Like in the box-counting dimension, the packing dimension of a set X has been suggested to be found by evaluating the limit:

$$d_{\text{pack}} = -\lim_{r\to 0} \frac{\log M(r)}{\log r}. \tag{17}$$

For the finite set X, the zero limit cannot be achieved. If we want to redefine the packing dimension in a scale-dependent manner, then the packing dimension of the finite data set X is estimated by the formula:

$$\hat{d}_{\text{pack}} = -\frac{\log M(r_2) - \log M(r_1)}{\log r_2 - \log r_1}. \tag{18}$$

In order to find the r-packing number $M(r)$ for the finite data set X, the approximation algorithm is used in Kégl (2003), Karbauskaitė and Dzemyda (2014).

In Karbauskaitė and Dzemyda (2014), a modification of the packing dimension estimator, that uses geodesic distances in order to improve the estimates of the intrinsic dimensionality, is proposed. It is shown that, in order to get true estimates, it is necessary to evaluate the geodesic distances between data points. If the Euclidean distances are used, one can get false estimates of the intrinsic dimensionality. The efficiency of modification of the packing dimension estimator is disclosed in the image analysis. The experiments with the sets of images of the moving object have showed that there are latent variables or features that characterize the motion of the object in the images. The number of latent variables (as well as the intrinsic dimensionality) is highly related to the number of degrees of freedom of a possible motion of the object.

5. Fractal Dimension in Applications

Applications of fractal-based methods appeared several decades ago, e.g. in astronomy (Scargle, 1990), ecology (Sugihara and May, 1990), meteorology (Houghton, 1991), earthquake analysis (Sahimi *et al.*, 1993), electroencephalogram analysis (Dvořák and Holden, 1991), physiological data analysis (Bernatavičienė *et al.*, 2007), etc. A comprehensive review of the most widespread fractal-based methods that are applied in medical image (signal) analysis is given in Lopes and Betrouni (2009). These methods are grouped

into three classes: box-counting methods (box-counting method, differential box-counting method, extended counting method), fractional Brownian motion methods (variogram method, the power spectrum), and area measurement methods (isarithm method, blanket method, triangular prism method).

Recently we have noticed the renascence of application of the fractal dimension: in geology (Nkono *et al.*, 2015), materials science (Lashgari *et al.*, 2015), novel pharmaceutics (Pippa and Demetzos, 2014, 2015), medicine (Nakatsuka *et al.*, 2015; Dedović *et al.*, 2015; Lennon *et al.*, 2015; Smitha *et al.*, 2015; Gokilavani and Vanitha, 2015; Smitha and Narayanan, 2015; Lawrence *et al.*, 2015), etc. Examples of such recent applications are described in detail below.

In geology, the relationship between the fractal dimension of orthopyroxene distribution and the temperature in mantle xenoliths has been disclosed. The fractal dimensions and their potential variations can be used to infer the physical conditions of rock formation at various scales of observation. The shape and distribution of orthopyroxene grains in ultramafic xenoliths are characterized quantitatively in terms of fractal dimensions.

In materials science, the fractal dimension is a significant factor that can be used to approximate the surface roughness, texture segmentation, and the image of compounds.

In novel pharmaceutics, the fractal dimension illustrates the self-assembly and morphological complexity of drug nanocarriers. The fractal dimension plays the key role in the elucidation of morphological characteristics, while the size and/or size distribution of drug nanocarriers did not change by changing the colloidal parameters, such as temperature and concentration.

In medicine, the fractal dimension is an important tool for the diagnosis of breast cancer (Dedović *et al.*, 2015), lung cancer (Lennon *et al.*, 2015), and brain cancer (Smitha *et al.*, 2015; Gokilavani and Vanitha, 2015). It is discovered in Nakatsuka *et al.* (2015) that the morphological change of midbrain, measured in the fractal dimension analysis, correlates a mild midbrain atrophy in patients with dementia with Lewy bodies. The fractal dimension has been used as a measure of complexity when analysing the effect of radiation on the electroencephalogram (EEG), while using a mobile phone (Smitha and Narayanan, 2015). The fractal dimension is one of the main parameters to reflect clot microstructure (Lawrence *et al.*, 2015).

Most applications above try to apply the discovered value of the fractal dimension in order to describe the complexity of the analysed object, but not to reduce the initial dimensionality of data. In the dimensionality reduction problems, the advantages of the fractal dimension arise when the analysed objects are described by a high number of features, i.e. when the data are high-dimensional and it is necessary to reduce the dimensionality. Such typical recent applications are presented e.g. in Ni *et al.* (2015), Zhang *et al.* (2015), Karbauskaitė and Dzemyda (2014).

A novel selective clustering ensemble algorithm, based on the fractal dimension and projection, is proposed for high-dimensional data clustering in Ni *et al.* (2015). In order to eliminate redundant and irrelevant attributes, at first the fractal dimension of a data set is calculated as the intrinsic dimension, and then the projection clustering algorithm is utilized to achieve the dimension reduction and clustering.

The fractal dimension may serve as a criterion in selecting the main features among a larger number of features describing the high-dimensional data. The optimally reduced dimensionality can be obtained by varying the number of features. In Zhang *et al.* (2015), a novel feature selection method, based on the multi-fractal dimension and harmony search algorithm, is proposed. Here, the multi-fractal dimension is adopted as the evaluation criterion of a feature subset that can determine the number of selected features.

One of the problems in the analysis of the set of images of a moving object is to evaluate the degree of freedom of motion of the object and the angle of its rotation in a separately taken image. Here the intrinsic dimensionality of multidimensional data, characterizing the set of images, can be used in order to reduce the dimensionality of data without losing much information. Usually, the image can be represented by a high-dimensional point the dimensionality of which depends on the number of pixels in the image and such a dimensionality is huge. In Karbauskaitė and Dzemyda (2014), it has been discovered that the intrinsic dimensionality, defined by a packing dimension, is highly related to the number of degrees of freedom of a possible motion of the object.

6. Conclusions

Real-life data, especially in image analysis, are often of a very high dimensionality. While analysing these data, we frequently have to reduce their dimensionality so that to preserve as much information on the analysed data set as possible. As usual, high-dimensional data can be efficiently summarized in a space of much lower dimensionality, i.e. on a nonlinear manifold, because high-dimensional data sets can have meaningful low-dimensional structures hidden in the observation space, i.e. the data are of a much lower intrinsic dimensionality. The concept of the intrinsic dimensionality should not be equalized to a topological dimension that is a classical measure of an object. The intrinsic dimension may be defined by a fractal dimension that describes natural objects and gives their degree of complexity. Fractals are objects with such a complexity that the classical means of measurement cannot be applied. While the topological dimension always yields an integer value, the fractal dimension is often a real number.

A lot of techniques have been proposed in order to estimate the intrinsic dimensionality of a data set. We review the approaches to estimate the intrinsic dimensionality and show their advantages and disadvantages. The stress is put on the fractal-based methods. Therefore, this study enables any researcher to learn the concept of the fractal dimension in essence and to choose the proper intrinsic dimensionality estimator with regard to the data set analysed and a problem in general.

References

Belkin, M., Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6), 1373–1396.

Bernatavičienė, J., Dzemyda, G., Kurasova, O., Marcinkevičius, V., Medvedev, V. (2007). The Problem of Visual Analysis of Multidimensional Medical Data. In: *Models and Algorithms for Global Optimization*, Vol. 4. Springer-Verlag, New York, pp. 277–298.

Borg, I., Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York.
Brand, M. (2003). Charting a manifold. In: Becker, S., Thrun, S., Obermayer, K. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 15. MIT Press, Cambridge, pp. 961–968.
Broomhead, D.S. (1985). Fractals. In: *5th International Symposium on Continuum Models of Discrete Systems, Nottingham, 14–20 July, 1985*, pp. 27–34.
Broomhead, D., Jones, R., King, G.P. (1987). Topological dimension and local coordinates from time series data. *Journal of Physics A: Mathematical and General*, 20(9), L563.
Bruske, J., Sommer, G. (1998). Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5), 572–575.
Buskes, G., van Rooij, A. (1997). *Topological Spaces From Distance to Neighborhood*. Springer-Verlag, New York.
Camastra, F. (2003). Data dimensionality estimation methods: a survey. *Pattern Recognition*, 36(12), 2945–2954.
Carter, K.M., Raich, R., Hero, A.O. (2010). On local intrinsic dimension estimation and its applications. *IEEE Transactions on Signal Processing*, 58(2), 650–663.
Chaudhuri, B., Sarkar, N. (1995). Texture segmentation using fractal dimension. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1), 72–77.
Costa, J.A., Hero, A.O. (2003). Entropic graphs for manifold learning. In: *Proceedings of the IEEE Asilomar Conference on Signals, Systems, and Computers*.
Costa, J.A., Hero, A.O. (2004). Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8), 2210–2221.
Costa, J.A., Hero, A.O. (2005). Estimating local intrinsic dimension with k-nearest neighbor graphs. *IEEE Transactions on Statistical Signal Processing*, 30(23), 1432–1436.
Cox, T.F., Cox, M.A. (2001). *Multidimensional Scaling*. Chapman and Hall, London.
Dedović, E., Gazibegović-Busuladžić, A., Beganović, A. (2015). Fractal analysis of digital mammograms. *Folia Medica Facultatis Medicinae Universitatis Saraeviensis*, 50(1), 55–58.
Ding, M., Grebogi, C., Ott, E., Sauer, T., Yorke, J.A. (1993). Estimating correlation dimension from a chaotic time series: when does plateau onset occur? *Physica D*, 69, 404–424.
Donoho, D.L., Grimes, C. (2005). Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 102(21), 7426–7431.
Dvořák, I., Holden, A.V. (1991). *Mathematical Approaches to Brain Functioning Diagnostics*. Manchester University Press, New York.
Dzemyda, G., Kurasova, O., Žilinskas, J. (2013). *Multidimensional Data Visualization: Methods and Applications*. *Springer Optimization and Its Applications*, Vol. 75. Springer, Berlin.
Eckmann, J.P., Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, 57, 617–659.
Einbeck, J., Kalantan, Z. (2013). Intrinsic dimensionality estimation for high-dimensional data sets: new approaches for the computation of correlation dimension. *Journal of Emerging Technologies in Web Intelligence*, 5(2), 91–97.
Fan, M., Qiao, H., Zhang, B. (2009). Intrinsic dimension estimation of manifolds by incising balls. *Pattern Recognition*, 42(5), 780–787.
Fan, M., Zhang, X., Chen, S., Bao, H., Maybank, S.J. (2013). Dimension estimation of image manifolds by minimal cover approximation. *Neurocomputing*, 105, 19–29.
Frisone, F., Firenze, F., Morasso, P., Ricciardiello, L. (1995). Application of topological-representing networks to the estimation of the intrinsic dimensionality of data. In: *Proceedings of International Conference on Artificial Neural Networks*, pp. 323–329.
Fukunaga, K. (1982). Intrinsic dimensionality extraction. In: Krishnaiah, P., Kanal, L. (Eds.), *Classification, Pattern Recognition and Reduction of Dimensionality, Handbook of Statistics*, Vol. 2. North-Holland, Amsterdam, pp. 347–362.
Fukunaga, K., Olsen, D. (1971). An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 20(2), 176–183.
Gokilavani, C., Vanitha, S. (2015). Fractional brownian motion and fractal analysis of brain mri images: a review. *International Journal of Applied Research*, 1(3), 21–24.
Grassberger, P. (1990). An optimized box-assisted algorithm for fractal dimension. *Physics Letters A*, 148, 63–68.
Grassberger, P., Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1–2), 189–208.

Hastie, T., Stuetzle, W. (1988). Principal curves. *Journal of the American Statistical Association*, 84, 502–516.
Hausdorff, F. (1919). Dimension und äusseres mass. *Mathematische Annalen*, 79, 157–179.
He, J., Ding, L., Jiang, L., Li, Z., Hu, Q. (2014). Intrinsic dimensionality estimation based on manifold assumption. *Journal of Visual Communication and Image Representation*, 25(5), 740–747.
Hein, M., Audibert, J. (2005). Intrinsic dimensionality estimation of submanifolds in R^d. In: *Proceedings of the 22nd International Conference on Machine Learning (ICML 2005), Bonn, Germany, August 7–11, 2005*. ACM, New York, pp. 289–296.
Heyting, A., Freudenthal, H. (1975). *Collected Works of L.E.J. Brouwer*. North-Holland, Elsevier.
Houghton, J. (1991). The bakerian lecture 1991: the predictability of weather and climates. *Philosophical Transactions of the Royal Society A*, 337, 521–572.
Jolliffe, I. (1986). *Principal Component Analysis*. Springer, Berlin.
Karbauskaitė, R., Dzemyda, G. (2009). Topology preservation measures in the visualization of manifold-type multidimensional data. *Informatica*, 20(2), 235–254.
Karbauskaitė, R., Dzemyda, G. (2014). Geodesic distances in the intrinsic dimensionality estimation using packing numbers. *Nonlinear Analysis: Modelling and Control*, 19(4), 578–591.
Karbauskaitė, R., Dzemyda, G. (2015). Optimization of the maximum likelihood estimator for determining the intrinsic dimensionality of high-dimensional data. *International Journal of Applied Mathematics and Computer Science*, 25(4).
Karbauskaitė, R., Kurasova, O., Dzemyda, G. (2007). Selection of the number of neighbours of each data point for the locally linear embedding algorithm. *Information Technology and Control*, 36(4), 359–364.
Karbauskaitė, R., Dzemyda, G., Marcinkevičius, V. (2008). Selecting a regularization parameter in the locally linear embedding algorithm. In: *Proceedings of the 20th International EURO Mini Conference "Continuous Optimization and Knowledge-Based Technologies" (EurOPT'2008)*, pp. 59–64.
Karbauskaitė, R., Dzemyda, G., Marcinkevičius, V. (2010). Dependence of locally linear embedding on the regularization parameter. *An Official Journal of the Spanish Society of Statistics and Operations Research (TOP)*, 18(2), 354–376.
Karbauskaitė, R., Dzemyda, G., Mazėtis, E. (2011). Geodesic distances in the maximum likelihood estimator of intrinsic dimensionality. *Nonlinear Analysis: Modelling and Control*, 16(4), 387–402.
Kégl, B. (2003). Intrinsic dimension estimation using packing numbers. In: Becker, S., Thrun, S., Obermayer, K. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 15. MIT Press, Cambridge, pp. 697–704.
Kirby, M. (2001). *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns*. Wiley, New York.
Kouropteva, O., Okun, O., Pietikäinen, M. (2002). Selection of the optimal parameter value for the locally linear embedding algorithm. In: *The 1st International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 359–363.
Kriukienė, E., Labrie, V., Khare, T., Urbanavičiūtė, G., Lapinaitė, A., Koncevičius, K., Li, D., Wang, T., Pai, S., Ptak, C., Gordevičius, J., Wang, S., Petronis, A., Klimašauskas, S. (2013). DNA unmethylome profiling by covalent capture of CpG sites. *Nature Communications*, 4, 2190.
Lai, Z.X. (2012). *Self similar optical fiber*. PhD thesis, Electrical Engineering and Computer Science – Dissertations. Paper 324.
Lantos, B., Márton, L. (2011). *Nonlinear Control of Vehicles and Robots*. *Advances in Industrial Control*. Springer, London.
Lashgari, A., Ghamami, S., Shahbazkhany, S., Salgado-Morán, G., Glossman-Mitnik, D. (2015). Fractal dimension calculation of a manganese-chromium bimetallic nanocomposite using image processing. *Journal of Nanomaterials*, 2015, 1–9.
Lawrence, M.J., Sabra, A., Thomas, P., Obaid, D.R., D'Silva, L.A., Morris, R.H.K., Hawkins, K., Brown, M.R., Williams, P.R., Davidson, S.J., Chase, A.J., Smith, D., Evans, P.A. (2015). Fractal dimension: a novel clot microstructure biomarker use in ST elevation myocardial infarction patients. *Atherosclerosis*, 240(2), 402–407.
Lee, J.A., Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Springer, New York.
Lennon, F.E., Cianci, G.C., Cipriani, N.A., Hensing, T.A., Zhang, H.J., Chen, C.T., Murgu, S.D., Vokes, E.E., Vannier, M.W., Salgia, R. (2015). Lung cancer – a fractal viewpoint. *Nature Reviews Clinical Oncology*.
Levina, E., Bickel, P.J. (2005). Maximum likelihood estimation of intrinsic dimension. In: Saul, L.K., Weiss, Y., Bottou, L. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 17. MIT Press, Cambridge, pp. 777–784.
Lopes, R., Betrouni, N. (2009). Fractal and multifractal analysis: a review. *Medical Image Analysis*, 13, 634–649.

Maggi, F. (2002). *Survey of the numerical characterisation of 2-D complex clusters*. Techical report 4-02, Faculty of Civil Engineering and Geosciences, Delft University of Technology.

Mandelbrot, B.B. (1967). How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science*, 156(3775), 636–638.

Mandelbrot, B.B. (1975). *Les Objets Fractals: Forme, Hasard et Dimension*. Flammarion, Paris.

Mandelbrot, B.B. (1977). *Fractals: Form, Chance, and Dimension*. Freeman, San Francisco.

Mandelbrot, B.B. (1983). *The Fractal Geometry of Nature*. Henry Holt and Company, New York.

Mo, D., Huang, S.H. (2012). Fractal-based intrinsic dimension estimation and its application in dimensionality reduction. *IEEE Transactions on Knowledge and Data Engineering*, 24(1), 59–71.

Monteiro, L.H. (2013). Overview of dynamical systems and chaos. In: *Chaotic Signals in Digital Communications*. CRC Press, Boca Raton, pp. 83–109.

Nakatsuka, T., Kudo, H., Kasai, R., Kasuya, S., Odashima, M., Inaoka, T., Terada, H. (2015). Morphological change of midbrain measured in fractal dimension analysis correlates midbrain atrophy in patients with dementia with Lewy bodies. *Poster presented at ECR 2015/C-0413*.

Nene, S.A., Nayar, S.K., Murase, H. (1996). *Columbia object image library (COIL-20)*. Technical report CUCS-005-96, Columbia University, New York.

Ni, Z., Wu, X., Ni, L., Tang, L., Xiao, H. (2015). The research on selective clustering ensemble algorithm based on fractal dimension and projection. *Journal of Computational Information Systems*, 11(11), 4025–4035.

Nkono, C., Féménias, O., Lesne, A., Mercier, J.C., Ngounouno, F.Y., Demaiffe, D. (2015). Relationship between the fractal dimension of orthopyroxene distribution and the temperature in mantle xenoliths. *Geological Journal*.

Olsen, L. (2007). Typical Rényi dimensions of measures. The cases: $q = 1$ and $q = \infty$. *Journal of Mathematical Analysis and Applications*, 331(2), 1425–1439.

Ott, E. (1993). *Chaos in Dynamical Systems*. Cambridge University Press, Cambridge.

Pettis, K., Bailey, T.A., Jain, A.K., Dubes, R.C. (1979). An intrinsic dimensionality estimator from near-neighbor information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, 25–37.

Pippa, N., Demetzos, C. (2014). Fractal analysis as a complementary approach to predict the stability of drug delivery nano systems in aqueous and biological media: a regulatory proposal or a dream? *International Journal of Pharmaceutics*, 473(1–2), 213–218.

Pippa, N., Demetzos, C. (2015). Fractal geometry as a new approach for proving nanosimilarity: a reflection note. *International Journal of Pharmaceutics*, 483(1–2), 1–5.

Rényi, A. (1960). Some fundamental questions of information theory. *MTA III, Osztályának. Közleményei*, 10, 251–282.

Rényi, A. (1961). On measures of entropy and information. In: *Proceedings 4th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 547–561.

Rényi, A. (1970). Introduction to information theory. In: *Probability Theory*. North-Holland, Amsterdam, pp. 540–616.

Roweis, S.T., Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2323–2326.

Sahimi, M., Robertson, M.C., Sammis, C.G. (1993). Fractal distribution of earthquake hypocenters and its relation to fault patterns and percolation. *Physical Review Letters*, 70(14), 2186–2189.

Sandau, K., Kurz, H. (1997). Measuring fractal dimension and complexity–an alternative approach with an application. *Journal of Microscopy*, 186(2), 164–176.

Saul, L.K., Roweis, S.T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4(2003), 119–155.

Scargle, J. (1990). Studies in astronomical time series analysis. IV. Modeling chaotic and random processes with linear filters. *Astrophysical Journal*, 359, 469–482.

Smitha, C.K., Narayanan, N.K. (2015). Analysis of fractal dimension of EEG signals under mobile phone radiation. In: *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, pp. 1–5.

Smitha, K.A., Gupta, A.K., Jayasree, R.S. (2015). Fractal analysis: fractal dimension and lacunarity from MR images for differentiating the grades of glioma. *Physics in Medicine and Biology*, 60(17), 6937–6947.

Sugihara, G., May, R.M. (1990). Applications of fractals in ecology. *Trends in Ecology and Evolution*, 5(3), 79–86.

Tenenbaum, J.B., de Silva, V., Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319–2323.

Theiler, J. (1990). Estimating fractal dimension. *Journal of the Optical Society of America A*, 7(6), 1055–1073.

Tolle, C.R., Junkin, T.R.M., Gorsich, D.J. (2003). Suboptimal minimum cluster volume cover-based method for measuring fractal dimension. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1), 32–41.

Valantinas, J. (1999). *Fraktalinė Geometrija*. Technologija, Kaunas.

van der Maaten, L.J.P. (2007). *An introduction to dimensionality reduction using MATLAB*. Technical report MICC 07-07, Maastricht University, Maastricht.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.

Verveer, P., Duin, R. (1995). An evaluation of intrinsic dimensionality estimators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1), 81–86.

Vrobel, S. (2011). *Fractal Time: Why a Watched Kettle Never Boils*. *Studies of Nonlinear Phenomena in Life Science*, Vol. 14. World Scientific, Singapore.

Weinberger, K.Q., Saul, L.K. (2006). Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1), 77–90.

Weisstein, E.W. (2003). *CRC Concise Encyclopedia of Mathematics*, second edition. Chapman and Hall/CRC, London.

Yata, K., Aoshima, M. (2010). Intrinsic dimensionality estimation of high-dimension, low sample size data with D-asymptotics. *Communications in Statistics – Theory and Methods*, 39:8–9, 1511–1521.

Zhang, C., Ni, Z., Ni, L., Tang, N. (2015). Feature selection method based on multi-fractal dimension and harmony search algorithm and its application. *International Journal of Systems Science*, 3476–3486.

Zhang, Z., Zha, H. (2004). Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment. *SIAM Journal of Scientific Computing*, 26(1), 313–338.

Zook, J.M., Iftekharuddin, K.M. (2005). Statistical analysis of fractal-based brain tumor detection algorithms. *Magnetic Resonance Imaging*, 23, 671–678.

Žilinskas, A., Žilinskas, J. (2009). Branch and bound algorithm for multidimensional scaling with city-block metric. *Journal of Global Optimization*, 43(2–3), 357–372.

R. Karbauskaitė is a researcher of System Analysis Department at Institute of Mathematics and Informatics of Vilnius University. She received a bachelor's degree in mathematics and informatics (2003) and a master's degree in informatics (2005) from Vilnius Pedagogical University, PhD in informatics from Vytautas Magnus University and Institute of Mathematics and Informatics (2010). Her research interests include multidimensional data visualization, estimation of the visualization quality, dimensionality reduction, estimation of the intrinsic dimensionality of high-dimensional data, and data clustering.

G. Dzemyda graduated from Kaunas University of Technology, Lithuania, in 1980, and in 1984 received there a doctoral degree in technical sciences (PhD) after postgraduate studies at the Institute of Mathematics and Informatics, Vilnius, Lithuania. In 1997 he received the degree of doctor habilius from Kaunas University of Technology. The title of a professor was conferred on him in 1998 at Kaunas University of Technology. He is the director of the Vilnius University Institute of Mathematics and Informatics and the head of the System Analysis Department of the institute. The areas of research are the theory, development and application of optimization, and the interaction of optimization and data analysis. The interests include visualization of multidimensional data, optimisation theory and applications, data mining in databases, multiple criteria decision support, neural networks, and parallel optimization.

Fraktalais grindžiamų metodų, skirtų daugiamačių duomenų vidinei dimensijai vertinti, apžvalga

Rasa KARBAUSKAITĖ, Gintautas DZEMYDA

Daugiamačių duomenų vidinės dimensijos vertinimas yra ypač aktualus uždavinys. Sukurta įvairių metodų, skirtų vidinei dimensijai interpretuoti ir įvertinti. Atsižvelgiant į dvi klasifikacijas – lokalus / globalus vertinimas ir projekcijos / geometriniai metodai – šiame straipsnyje susikoncentruojama į fraktalais grindžiamus metodus, kurie pagal pirmą klasifikaciją priskiriami globaliam vertinimui, o pagal antrą – geometriniams metodams. Labiausiai šiame straipsnyje akcentuojami daugiamačių duomenų vidinės dimensijos vertinimo skaičiuojamieji aspektai. Taip pat atskleisti fraktalais grindžiamų metodų privalumai ir trūkumai bei trumpai pristatyti šių metodų taikymai.