Minimum-Description-Length Criterion for Image Interpretation and Data Analysis in Spatial Informatics

He-Ping Pan

Cooperative Research Centre for Sensor Signal and Information Processing
SPRI Building, Technology Park Adelaide, The Levels
SA 5095, Australia
Email: heping@cssip.edu.au

Abstract

Spatial Informatics typically involves interpretation of remotely sensed images and analysis of multi-sources of data. Random perturbation in the observations and diversity of hypothesized models give rise to the uncertainty and difficulty of image interpretation and data analysis. The Minimum-Description-Length (MDL) principle is a best established criterion, which selects the best model with the minimal length of jointly encoding the data and the model. Although in terms of probability, MDL criterion is equivalent to the Maximum Aposterior Probability (MAP) criterion, it is advantageous at combining different data types and different model structures in a uniform measure - the total number of bits. It is more realistic to computerized information processing, because everything is discrete with limited resolution. This paper clarifies the formulation of the MDL criterion, its relationships to information theory, stochastic complexity, and Bayesian decision strategy. To sufficiently demonstrate the applicability of this criterion, a number of cases where this criterion can be and has been applied are described, including global interpretation of remotely sensed images for landuse mapping, line generalization, digital terrain modelling, spatial indexing in GIS, and unsupervised clustering. The emphasis is on showing how each of these classic problems can be reformulated under this new criterion. The reformulations are likely to lead to breakthroughs or significant progresses in the fields.

I. INTRODUCTION

Spatial Informatics refers to a major branch of the general information science which studies the acquisition, representation, processing, analysis, and use of the information of *spatial phenomena*. If we only consider the geographical sphere of the earth, spatial informatics is instantiated explicitly to *geomatics* typically so called in North America, or *geoinformatics* so called in Europe. These terms are used to refer to a new interdiscipline which is a marriage of *geography* and *informatics*. Here geography is meant as a general umbrella covering natural and cultural geography, geodesy and cartography. Informatics typically so called in Europe is also called *computer science* in North America. It studies the general mechanisms of information representation and processing. All the discovered formalisms and mechanisms are realized and integrated in a machine — the computer. There are three parts that constitute an application domain of informatics: the information, the computer, and the user. The constitution of these three parts, in general, takes the form of an information system. In geomatics, we are concerned only with *geographical (or spatial) information systems (GIS or SIS)*.

What is really special in spatial infomation systems in comparison with all other information systems like office automation, bank accounting, economic management, library archiving, and so on ? Some major specialties of problems and issues in spatial informatics may be identified as follows:

1. *Spatial dimensionality:* The prerequisite for starting thinking on any problems and issues in spatial informatics and acquiring observations and existing data is that we must presume the existence of the space. In GIS, we can ignore the Einstein space, but only work with the Euclidean space. The order of the data in spatial information systems cannot be arbitrated. Instead, all the spatial data must be indexed in a spatial reference system. In general, three-dimensional space and one-dimensional and uni-directional time constitute the four dimensions of such a spatial reference, which provides the very basis for any other transformed spatial indexing.

1082-4006/95/0101-42$3.00

2. *Geometric isomorphism*: The mapping between the real spatial objects and their models in the spatial information systems should preserve the isomorphism in geometry. Without this isomorphism, the spatial data cannot be analyzed on an objective ground. By geometry, we mean the relative position, orientation, shape, size of spatial objects and other geometric relationships among them. On the basis of geometry, the more flexible topology among objects can then be defined.

3. *Stability and randomness*: The complexity of spatial phenomena is several orders higher than other phenomena. In most cases, the acquired data describing some spatial phenomena contains comparatively small amount of information that cannot encode the full complexity of the phenomena. Therefore prior knowledge must be used. The most general prior knowledge takes the form of implicit assumptions. Disregarding application domains the first assumption is that the world is not made up of chaotic phenomena, but it has its own structures. The structure means the stability of the world in space, time, and resolution scale of observation, e.g. imaging or measurement. Attached with the stability of the world, randomness is ubiquitous. Spatial data are often acquired (observed) at a scale of resolution either above or below some ideal level, thus must have a random distribution. This randomness is usually supposed to be either uniform, Gaussian, fractional, or waveletal. In fact, randomness of spatial phenomena so considered by human beings is a reflection of over-complexity of the nature which exceeds our capability of mathematical analysis and modeling.

4. *Multiple sources of information:* Geography is an umbrella covering many aspects of the spatial phenomena on the earth surface, such as topography, geomorphology, hydrology, forestry, agriculture, urban planning and administration, traffic network, and so on. All these aspects produce spatially indexed data. Information from multiple sources are initially encoded in different forms. A proper indexing and fusion of multiple sources of spatial information is an obvious difficulty which is not significantly present in other information systems. Even in the spatial information systems, new information may be produced through interaction of different components of the system.

5. *Multiple categories of modeling:* Geography is a science of modeling the geographical phenomena. Different aspects of spatial phenomena should be modelled into different categories. For example, a model of urban traffic network may be quite different from a model of hydrological network, and should be completely different from a model of forest canopy distribution. In remote sensing, the basic categories of models are geometric and radiometric models. These two categories of models are direct related to image observations. With the presence of uncertainty and noise, multiple models of different categories may need to be estimated from a common set of observations and existing data. This is so-called problem of information fusion in GIS.

Modeling of spatial phenomena from observations is a central issue in spatial informatics. With the ubiquitous presence of randomness in the data, the science of probability and statistics must be used, which provides a basis for any acceptable scientifically sound mathematical approaches. However, two facts must be considered. Firstly, a simultaneous solution for models of multiple categories in terms of Maximum Aposterior Probability (MAP) may not be formulatable, because in some models some probabilities may be unknown or not estimable, or because some models of different types are simply not comparable. Secondly, some formulations of solutions may be in terms of continuous mathematics with variables of infinite resolution; however everything in computer must be represented as discrete values, so we must take into account the limitation of finite resolution. For many reasons like these, we should use a new criterion called the Minimum-Description-Length (MDL) principle for model selection and robust estimation.

In the next section, we provide a brief introduction to and a formal formulation of the MDL principle. We will point out the advantages of the MDL criterion over the classical MAP criterion. In section III and IV, we show how this new criterion can be used in image interpretation and data analysis. Practical examples for MDL-based image interpretation include (1) a general paradigm for interpretation of remotely sensed images for landuse mapping, (2) an objective recursive mechanism for line generalization. Examples for data analysis in GIS include (1) digital terrain modeling, (2) spatial indexing. Finally, a new criterion for unsupervised clustering is proposed as an application of the MDL criterion in general pattern recognition.

What is important in this work is how to reformulate each of these classical problems in a new way under a new criterion, which is very likely to lead to breakthroughs or significant progresses in the fields. Once a somewhat fuzzy problem is completely formulated into mathematical expressions, solving the problem is reduced to a pure mathematical matter or an experimental effort, which is then beyond the scope of this paper.

II. FORMULATION OF THE MDL CRITERION

Statistical modeling can be regarded as a problem of understanding and explaining a given set of observed data which often appear quite chaotic at the first glance. By understanding and explanation, we naturally presume that there should be some regularities underlying the data, and some redundancy as complement of regularities inherent in the data. A thorough understanding of the data set means that we can give a description of these regularities and redundancy which in turn can determine the data set completely. Our purpose in general is only to have a perfect nonredundant description of the data by removing all redundancy. In this sense, there is only one criterion to be used in model selection and estimation, namely to consider the total number of binary digits with which the data set and the model can be written down completely. This number is called the total description length of the data including its explaining model. The shorter this description length is, the better the model is. The best model out of all alternative models will be the one with the shortest total description length. This is the intuitive formulation of the Minimum Description Length (MDL) principle.

The root of the MDL principle goes back to the algorithmic notion of information by Solomonoff [16], Kolmogorov [5], and Chaitin [1], which defines the stochastic complexity of a binary string to be the length of the shortest program needed to generate this string in a universal computer. This principle was first formally introduced by Rissanen [11-14] as a new criterion for statistical modeling. Later on, it was formulated independently by Georgeff and Wallace [4] as a general theory selection criterion for inductive inference. Fua and Hanson [3] and Leclerc [6] belong to those who first applied the MDL principle to the image analysis problem, specfically for object delineation from aerial images and general image segmentation. Pan and Forstner [9] first applied this principle in the field of neural network research for automatic architecturing of pattern recognition neural networks.

The following formulation of the MDL criterion integrates various discontiguous aspects of the previous theoretical discoveries and practical constructs into a consistent description with the clarity and ease required for direct application in spatial informatics.

A. Data, Model, and Language

Let us consider the problem: given a set of observed data D, we are seeking a good *model* M which can explain the data D. If alternative models may be hypothesized, we are seeking the best model. In order to hypothesize models, we need a language which provides the necessary terms, syntax, and semantics with which a model M and the data D with or without a model can be described. Thus there are at least three categories: data, model, and language. We will call this language the description language of the given problem domain, denoted by $\mathcal{L}$. In fact, such a language represents all our prior knowledge about the problem domain. Without any prior knowledge, we even have no words and terms to describe any problem in the domain. We naturally presume that the syntax and semantics of every term to be used are well understood, and every relationship and procedure associated with terms are also well defined. In other words, any terms, syntax, and semantics of this description language need not to be encoded, because they are known *a priori*.

B. The Description Language

Let us now consider the ensembles of data and models. The whole ensemble of all possible data sets form a data space $\mathcal{D}$. The whole ensemble of all possible models form a model space $\mathcal{M}$. In this sense, any observed data set D is a sample from the data space $\mathcal{D}$; and any possible model M is a sample from the model space $\mathcal{M}$ We shall use $\mathcal{L}(X)$ to denote the complete description of X in the description language $\mathcal{L}$., and $L(X)$ to denote the length (total number of bits) of $\mathcal{L}(X)$:

$$L(X) = |\mathcal{L}(X)| \qquad (1)$$

With these space concepts, we now can set up some principal criteria over the description language $\mathcal{L}$:

1. Completeness: The description language $\mathcal{L}$ must

be complete in two levels: (1) it must provide necessary terms and syntax to describe any data set D from the data space $\mathcal{D}$; but not limited to only one or some data sets; (2) all the descriptions of a given data set D must exactly determine this single data set.

2. Efficiency: For any given data set D, the length $L(D)$ of the description $\mathcal{L}(D)$ of the data D in this language $\mathcal{L}$ must be shorter than the total number of bits of the data set D in its original form.

3. Computability: The description $\mathcal{L}(D)$ must be representable and constructible by computer in a reasonable amount of time.

4. Stability: This criteria is a constraint on both the description language $\mathcal{L}$ and the definition of the best description $\mathcal{L}_{best}(D)$ of a given data set D. It is to say that $\mathcal{L}_{best}(D)$ should not change significantly when the data set D has some trivial changes.

The first three criteria represent the consensus from most MDL proponents, and the stability criterion was first pinpointed by Leclerc[6] for image segmentation problem.

C. The Alternative Models and the Best

It is generally assumed that the model space $\mathcal{M}$ is an enumerable set of models:

$$\mathcal{M} = \{M_i \ \mathrm{i} = 1,\ 2, \ldots\} \qquad (2)$$

If two alternative models M_i and M_j, $i \neq j$, have the same number of parameters, and two parameter vectors have the same semantic meaning, the two models should be considered as one model. Two models are said to be different if either the number of parameters in each model is different, or the semantic meaning of some parameters is different. Any structured model can be characterized by a number of primitives and relationships between primitives, and the primitives and relationships can always be represented by symbols which can be ordered properly into a vector of parameters. Therefore in general, we construct a model M_i as a vector of n_i parameters:

$$M_i = [\theta_{i1}\ \theta_{i2} \cdots \theta_{in_i}] \qquad (3)$$

When we fit this model into the data D its parameters will be estimated:

$$\hat{M}_i = [\hat{\theta}_{i1}\ \hat{\theta}_{i2} \cdots \hat{\theta}_{in_i}] \qquad (4)$$

The total description length of the data D which is explained by a model M_i is thus

$$\begin{aligned} L(D) &= |\mathcal{L}(D \mid M_i)| + |\mathcal{L}(M_i)| \\ &= L(D|M_i) + L(M_i) \qquad (5) \end{aligned}$$

Let $L(D) = |\mathcal{L}(D)|$ denote the description length of the data D without any model, a model M_i is said to be efficient if and only if

$$L(D, M_i) < L(D) \qquad (6)$$

This is in fact the necessary condition for accepting a model M_i. A model M_i is said to be better than another model M_j if and only if

$$L(D, M_i) < L(D, M_j) \qquad (7)$$

We now come to the point of formulating the MDL criterion.

The minimum-description-length criterion

A model M_i is said to be the best in the whole model space $\mathcal{M}$ if and only if

$$\begin{aligned} &L(D, M_i) < L(D), M_j) \\ &\qquad \forall M_j \in \mathcal{M} \ \ (j \neq i) \qquad (8) \end{aligned}$$

Notice that the conditional description length $L(D|M_i)$ varies with the different estimate $\hat{M}_i$ of the parameters of M_i, therefore the joint description length $L(D, M_i)$ should be detailed as

$$\begin{aligned} L(D, M_i) &= L(D|\hat{M}_i, M_i) \\ &\quad + L(\hat{M}_i|M_i) + L(M_i) \qquad (9) \end{aligned}$$

Because the description language $\mathcal{L}$ is supposed to be complete, so there must be a way for indexing each model M_i in the model space, so the syntax and semantics of each model M_i need not to be encoded. What we only need is an index i for M_i. Thus we can write

$$\begin{aligned} L(D, M_i) &= L(D|\hat{M}_i, M_i) + L(\hat{M}_i|M_i) + L(M_i) \\ &= L(D|\hat{M}_i, i) + L(\hat{M}_i|i) + L(i) \qquad (10) \end{aligned}$$

Therefore the MDL criterion seeks the best fitting state $\hat{M}_i$ of the best model M_i satisfying

$$\begin{aligned} &L(D|\hat{M}_i, i) + L(\hat{M}_i|i) + L(i) \\ &\qquad < L(D|\tilde{M}_i, i) + L(\tilde{M}_i|i) + L(i) \qquad (11) \end{aligned}$$

$$L(D|\hat{M}_i,i)+L(\hat{M}_i|i)+L(i)$$
$$<L(D|\hat{M}_j,j)+L(\hat{M}_j|j)+L(j) \quad \forall j \neq i \qquad (12)$$

where $\hat{M}_i$ denotes the optimal estimate of the parameters of M_i, $\tilde{M}_i$ denotes any other non optimal estimate of M_i. In case there is only one model type, the MDL criterion is reduced to seek the optimal estimate of the model parameters as expressed in (11).

D. Probability and Description Length

The complexity measure from information theory

According to the information theory initiated by Shannon [15], a discrete information source can be modeled as a Markov process. Consider a variable which may be instantiated to a set of values $\{x_1, x_2, \ldots, x_i, \ldots\}$. If a value x_i is emitted from this source with a probability $P(x_i)$, the information amount of this x_i is:

$$I(x=x_i)=I(x_i)=-lbP(x_i) \qquad (13)$$

where lb denotes the base-2 logarithm. An optimal description language $\mathcal{L}$ for this information source is supposed to minimize the description length of the variable x. In such an optimal language, the description length $L(x)$. equals the information amount $I(x)$:

$$L(x)=-lbP(x) \qquad (14)$$

This is the very basic relation between the probability of a variable and its description length. With this relation, we may compare the MDL criterion with the Bayes decision rule and MAP criterion.

The Bayes rule and MAP criterion

The Bayes strategy for selecting the best model out of the model space $\mathcal{M}$ by considering the statistical properties of alternative models. The selected model is optimal in the sense of yielding the lowest probability. The Bayes decision rule is that the best model M_i. out of the whole model space should satisfy

$$P(D|M_i)P(M_i)$$
$$>P(D|M_j)P(M_j) \quad \forall j \neq i \qquad (15)$$

where $P(D|M)$ denotes the probability of data D given model M. This form of objective function is often referred as the Maximum Aposterior Probability (MAP) criterion. In case $P(D|M_i)$ or $P(M_i)$ is unknown or not explicitly computable, the objective function (15) may be transformed to an equivalent expression as

$$P(D \mid M_i)P(M_i)$$
$$=P(D,M_i)=P(M_i|D)P(D) \qquad (16)$$

Dropping the term $P(D)$ which is common for the variable i, yields an alternative form of the objective function:

$$P(M_i|D) \qquad (17)$$

MAP versus MDL

By applying the relation (14) between the description length and the probability, the conditional description length of the data D given a model M_i and that of the model M_i are given by

$$L(D|M_i)=-lbP(D|M_i) \qquad (18)$$
$$L(M_i)=-lbP(M_i) \qquad (19)$$

The MDL criterion can be rewritten as choosing the M_i that minimizes

$$-lbP(D|M_i)-lbP(M_i)L(M_i) \qquad (20)$$

This is equivalent to the MAP criterion which maximizes

$$P(D|M_i)P(M_i) \qquad (21)$$

The MDL and MAP criteria may be compared in several aspects:

1. Prior versus Commonsense Knowledge: MAP criterion uses the prior probability $P(M_i)$ of the model M_i while MDL uses its description length $L(M_i)$. If the description language is optimal, the two terms should be equivalent. This means the prior probability is implicitly specified by the given description language. In case the prior probability is unknown, $L(M_i)$ in fact represents our commonsense knowledge instead of prior knowledge about the model M_i.

2. Uniformity and Flexibility: MAP uses the joint probability of the data D and the model M_i, which is a real value, as the quantity for optimization, while MDL uses the total number of bits which is an integer. With the relation (18) and (19), we can see MDL is more flexible. If $P(M_i)$ is unknown, we can calculate the description length of the model

parameters. $P(D|M_i)$ usually refers to the probability of the residuals of the data D from the estimated model M_i. There are quite a few commonly accepted probability functions for residuals e.g. uniform or Gaussian noise, and outliers. Therefore $L(D|M_i)$ can be computed from the assumed probability functions $P(D|M_i)$ and also some direct description of the outliers.

3. Mathematical Truth versus Discrete Reality: Notice the probability by definition represents a mathematical idealism, while the description length in units of bits is a practical measure. Consider every value represented and manipulated by computer is in fact discretized, we must take into account the limitation of finite resolution for any parameter.

4. Computational Complexity: It appears that the MAP involves only some conventional computations when the probability functions are known, and MDL computation would be too complex. In fact, what we only need is the description length of the models and residuals; it is not necessary to actually encode them in their optimal description languages. It is however true that an application of the ideal MDL criterion generally requires a global optimization which may involve a large effort of search or relaxation in the model space.

E. Some Priors and Finite Resolution

We now consider the problem of finite resolution for any variable to be encoded and introduce some priors.

Integers

Consider an integer variable k with the domain $[i,j]$, where i, j are integers and $i<j$, which takes a set of discrete integer values $K=\{k_1,k_2,\ldots,k_n\}$. The total description length of this integer set (without any model) is given by

$$L(K)=L(i)+L(j-i)+n lb(j-i) \tag{22}$$

If we know i varies from 0 to $N-1$, then

$$L(i)=lb(N) \tag{23}$$

If the range of an integer i is unknown, we may use a universal prior for the natural numbers proposed by Rissanen (1983a):

$$L(i)=lb^*(i)+lb(c) \tag{24}$$

where

$$lb^*(i)=lb(i)+lb(lb(i))+lb(lb(lb(i))) +\ldots up\ to\ all\ positive\ terms \tag{25}$$

$$c\approx 2.865064 \tag{26}$$

Real values

A real value variables x with resolution ε can be encoded as a transformation to integer

$$L(x|\varepsilon)=L([\frac{x}{\varepsilon}]) \tag{27}$$

where [] denotes the roundoff of a real to its closest integer. The resolution ε is the minimum difference between any two values of the real variable x If ε is unknown, it can be estimated through a code length optimization process. Notice that the smaller ε is, the greater the precision is and the better the model fits to the data, but the longer the code length will be.

Variables of known probability function

Two widely used probability functions are equal and normal distributions. Their description lengths without or with a resolution ε: $L(x)$, $L(x|\varepsilon)$ are given below:

- Equal Distribution:

$$x\sim E[a,b] \tag{28}$$

$$P_E(x)=\begin{cases}\frac{1}{b-a} & : \quad a\le x\le b\\ 0 & : \quad x<a\ or\ x>b\end{cases} \tag{29}$$

$$L_E(x)=lb(b-a) \tag{30}$$

$$L_E(x|\varepsilon)=lb\left[\frac{b-a}{\varepsilon}\right] \tag{31}$$

- Normal Distribution:

$$x\sim N(\mu,\sigma) \tag{32}$$

$$P_N(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty<(x,\mu)<\infty,\quad \sigma>0 \tag{33}$$

$$L_N(x)=\frac{1}{2}lb(2\pi)+lb(\sigma)$$

$$+\frac{1}{2\ln 2}\left(\frac{x-\mu}{\sigma}\right)^2 \qquad (34)$$

$$L_N(x|\varepsilon)=\frac{1}{2}lb(2\pi) + lb\left[\frac{\sigma}{\varepsilon}\right]+\frac{1}{2\ln 2}\left(\frac{x-\mu}{\sigma}\right)^2 \qquad (35)$$

where P_E, L_E denote the probability and description length under the equal distribution, and P_N, L_N under the normal distribution respectively.

III. MDL-BASED INTERPRETATION OF REMOTELY SENSED IMAGES

Images are observations. This concept is a substantial generalization of the classical surveying and measurement notions where observations are referred only to collection of isolated and point-wise values, e.g. an angle, a distance, a height, a temperature, a force, etc. An image is an observed mapping of a surface. The amount of data in an image observation is at least 2-orders higher than pointwise observations. Neighboring pixel values are correlated in general, which is a reflection of the constraints over the physical properties of the imaged surfaces. Interpretation of an image aims at recovering the physical properties of the surfaces by exploitation of various constraints. Mathematical formulation of any constraints leads to models for image interpretation. Models can be either for pure image phenomena, for pure scene phenomena, or for image formation process. Models of diversified information sources should be combined in the sole interpretation process. At this point, the MDL principle plays the role of the overall criterion for optimal model selection and simultaneous estimation of model parameters.

To demonstrate the approach, two examples will be described in the following. The first one is a global formulation of the problem of interpreting remotely sensed images for landuse mapping. The second one is generalization of extracted edges onto more significantly structured edges.

A. Image Interpretation for Landuse Mapping

This is still an ongoing effort [8,10]. For landuse mapping purpose, remotely sensed images are interpreted with three levels of modeling: structure S, geometry G, and image intensity I (or say, radiometry. If we only consider single channel of spectrum, radiometry is reduced to image intensity).

The structure S refers to the topology of landuse parcels, e.g. containment of a small parcel in a larger one, and neighborhood of two parcels, etc. If we limit our attention only to agriculture fields, we may discover some constraints over the topology of agricultural landuse parcels. In fact, we have observed that the macro structure of polygon maps representing the boundaries of landuse parcels is the result of a spatial process in which larger parcels are sequentially or parallelly partitioned into smaller ones during reallotment. The most significant characteristic of this landuse structure is its fractal-like recursiveness of the polygon partitioning. We have developed a Stochastic Polygon Map Grammar (SPMG) as a generic model of this structure. In this grammar, there is one generic structural primitive: quadrilateral, and two types of spatial relations: spatial containment and neighborhood. A large quadrilateral polygon may be partitioned into a sequence of smaller neighboring quadrilateral polygons with a certain probability which is a function of many factors relating the overall statistics, local geometric shapes and size, etc. For a given polygon split tree over several generations, the joint probability of this tree is the multiplication of all the probabilities associated with each splitting action. Therefore, the probability $P(S)$ of a given polygon structure S is computable. Fig. 1 shows a polygon split tree. The input image in Fig. 2(a) is generated by using this grammar.

The geometry G refers to the shape, size, orientation, and position of each polygon. The geometry of a polygon is completely determined by its boundary which is an ordered list of edges. Each edge is supposed starting from a knotting point, called vertex, which is an intersection of at least three edges. Therefore, the geometry of the whole polygon map is completely determined by the full enumeration of vertices and edges. The total description length $L(G|S)$of the geometry G, given the structure S, of a polygon map is computable via summing up the encoding lengths of vertices and edges. Each vertex is determined by its x- and y-coordinates. Each edge is determined by its starting and ending vertex numbers, and internal knotting points. The encoding length of these values can be computed by using the formulas given in the section II.E.

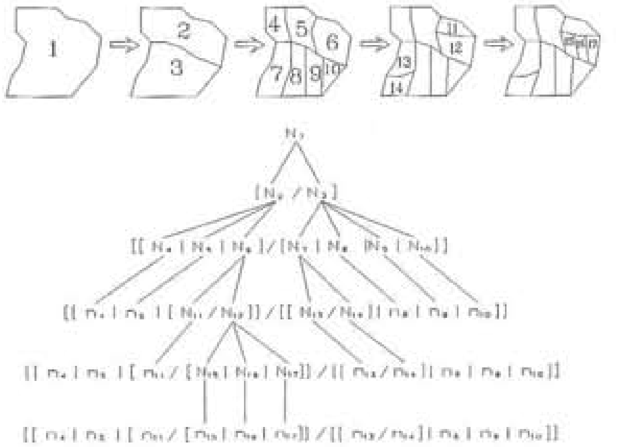


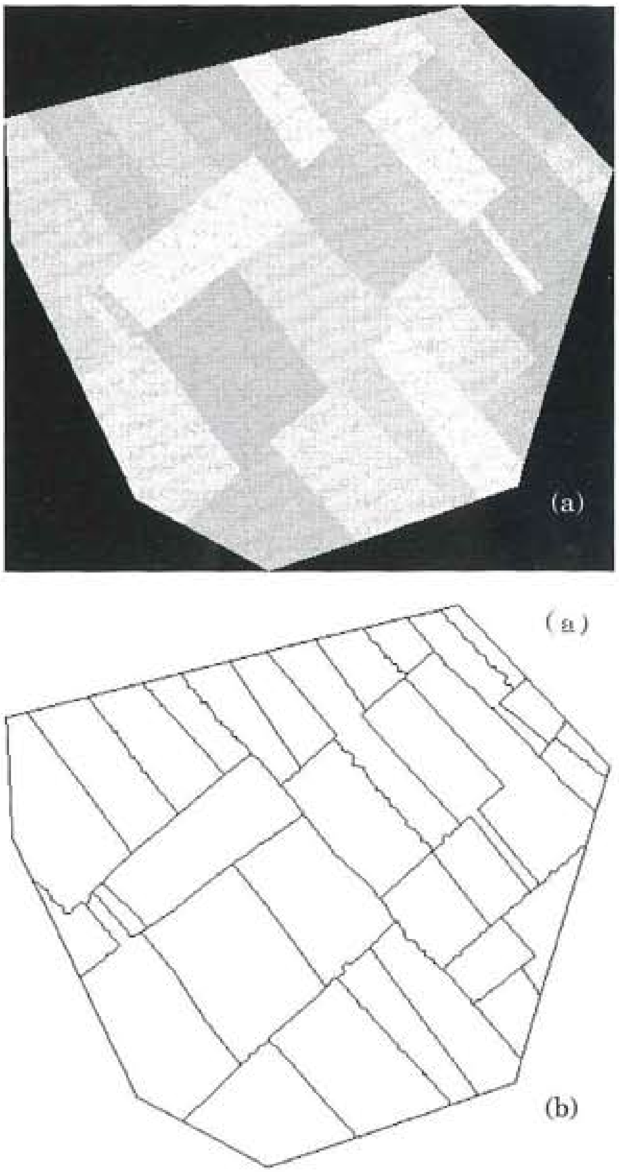
Figure 1. Recursive Polygon Partitioning and its tree representation

The image intensity I refers to any model of pixel values within an uniform image region which corresponds to one parcel. The simplest model is a horizontal plane for the image intensity surface. A more general model is a polynomial surface. Once the ideal image intensity surface is given, the real image data D can be considered as a corruption of the ideal surface by image noise N and outliers O. The ideal image intensity I can be encoded as an integer or as a set of parameters in the polynomial surface model, so $L(I \mid G,S)$ is computable. The image noise is generally assumed to be either Gaussian or white, so the probability $P(N \mid I,G,S)$ is computable. The outliers refers to those pixels whose values are far from the assumed probability distribution. Each outlier can be encoded as a triplet: the x- and y-coordinates and the intensity. Thus $L(O \mid N,I,G,S)$ is computable. Therefore, the total description length of the whole image given these models is

$$\begin{aligned} L(D,I,G,S) &= L(D \mid I,G,S) \\ &\quad + L(I \mid G,S) + L(G \mid S) + L(S) \\ &= L(O \mid N,I,G,S) + L(N \mid I,G,S) \\ &\quad + L(I \mid G,S) + L(G \mid S) + L(S) \qquad (36) \end{aligned}$$

where $L(G \mid S)$, $L(I \mid G,S)$ and $L(O \mid N,I,G,S)$ are directly computable from the ways at which they are encoded, no actual encodings are needed; and $L(S)$ and $L(D \mid I,G,S)$ are computable from their probabilities:

$$L(S) = -lbP(S) \qquad (37)$$



$$L(N \mid I,G,S) = -lbP(N \mid I,G,S) \qquad (38)$$

Figure 2. (a). An input image; and (b). Its segmentation.

Solving the interpretation problem is via a complex process of minimizing the global quantity $L(D,I,G,S)$ For detailed information on this approach, see [9]. Detailed formulation of $L(D,I,G)$ is given in [7]. Fig. 2 shows an example of image segmentation by using this approach.

B. Line Generalization

In image analysis, line generalization refers to generalize raw edges extracted from an image either by a local edge detector or through an image segmentor. The raw edges contain too many details that are not useful or meaningful for further interpretation. Generalization is to remove those details that may be considered as positional noise.

An edge in a digital image can be represented as a polyline which is characterized by a starting point, a sequence of internal points, and an ending point. The generalization of such a polyline is to remove some internal points, so the remaining points bear significant information on the shape of the edge. Let S be an original polyline which is a series of points:

$$S=[p_1\ p_2 \cdots p_n] \tag{39}$$

Suppose S is generalized to a new polyline G with fewer internal points

$$G=[p_{i1}\ p_{i2} \cdots p_{im}] \qquad m<n, \quad i_j \in [1,n], \quad j=1,2,\ldots,m \tag{40}$$

As G consists of a number of line segments, say, $[p_{i_1}\ p_{i_2}],[p_{i_2}\ p_{i_3}],\ldots,[p_{i_{m-1}}\ p_{i_m}]$, so the generalization G can be decomposed as

$$G=\Diamond_{k=1}^{m-1}G_k \tag{41}$$

$$G_k=[p_{i_k}\ p_{i_{k+1}}] \tag{42}$$

We consider all other points in S that are not in G as positional noise points which also form a series N

$$N=\Diamond_{k=1}^{m-1}N_k \tag{43}$$

$$N_k=[p_{i_k+1}\ p_{i_k+2} \cdots p_{i_{k+1}-1}] \tag{44}$$

where we use $\Diamond$ to denote the concatenation of two or more series. With this decomposition, N_k can be considered as the noise to G_k. Because G_k is supposed to be an ideal line segment and N_k is a rasterized series, so the positional noise N_k takes the form of shift of each internal point in the direction perpendicular to G_k, because the shift along G_k is negligible to the shape of G_k.

The total description length of S given a generalization G as a model is

$$L(S)=L(G)+L(S-G)=L(G)+L(N) \tag{45}$$

This is the objective function used to seek the best generalization $\hat{G}$:

$$L(\hat{G})+L(S-\hat{G})=\min_G(L(G)+L(S-G)) \tag{46}$$

We first transform the coordinates of points of N_k into a local reference system taking G_k. as the x-axis direction. Then, ideally the y-coordinate for each internal point of N_k should be zero if there is no noise. The positional noise takes the form of non-zero y-coordinates. Generally the positional noise for two neighboring points are correlated. It is obvious that the difference between two neighboring y_{j-1} and y_j is generally smaller than $|y_{j-1}|$ and $|y_j|$, so it is cheaper to only encode such differences. This kind of encoding may lead to a Random Markov Chain model for the positional noise along L_i:

$$y_{j+1}=ay_j+\varepsilon_k \quad (j=i_{k,}\ i_k+1,\ldots,i_{k+1}-1) \tag{47}$$

where a is a parameter of correlation, ε_k is a variable (error term) which follows a normal distribution:

$$\varepsilon \sim N(0,\sigma_k) \tag{48}$$

where σ_k is to be estimated from the given data N_k,

With $a=0$ the noise is uncorrelated, which refers to roughness. It is equal to say that the y_j themselves follow a normal distribution. With $a=1$ the noise is correlated, which refers to smoothness. In general, estimated $\hat{a}$ is between $(0,1)$. Because we use a fixed description length of this parameter, so a needs not to be encoded. Therefore, according to formula (34), the description length of N_k is then

$$\begin{aligned} L(N_k) &= L(\hat{\sigma}_k)+\sum_{j=i_k}^{i_{k+1}} l(p_j) \\ &= L(\hat{\sigma}_k)+L(y_{i_k}) \\ &\quad +\frac{i_{k+1}-i_k}{2\ln 2}(\ln(2\pi)+2\ln\hat{\sigma}_k+1) \end{aligned} \tag{49}$$

The total description length of the positional noise N is

$$L(N)=L(S|G)=\sum_{k=1}^{m-1}L(N_k) \tag{50}$$

An approach to minimize the objective function (46) is to use a recursive mechanism similar to that of Douglas and Peuker [2] but without any control parameter.

For generality, let us suppose we are considering a portion $S_{i,j}$ of an original linear pattern $S: S_{i,j} = [p_i \;\; p_{i+1} \;\; \dots p_j]$, where $j - i \geq 2$. Alternative generalizations may be hypothesized with some most significant points of $S_{i,j}$ (see Fig. 3): the starting point p_i, the ending point p_j., the two points with either the largest positive or negative new y coordinate p_{s_1} and p_{s_2}, $s_1 < s_2$. Let p_s denote one of p_{s_1} and p_{s_2} with the largest absolute y value, i.e. the farthest from the line defined by p_i and p_j. The following three hypotheses are the most significant:

Hypothesis 0:

$$G_{i,j}^{(0)} = [p_i \;\; p_j] \tag{51}$$

Alternative Hypothesis 1:

$$G_{i,j}^{(1)} = [p_i \;\; p_s \;\; p_j] \tag{52}$$

Alternative Hypothesis 2:

$$G_{i,j}^{(2)} = [p_i \;\; p_{s_1} \;\; p_{s_2} \;\; p_j] \tag{53}$$

If one of p_{s_1} and p_{s_2} does not exist, $G_{i,j}^{(2)}$ is reduced to $G_{i,j}^{(1)}$. In general, we assume three hypotheses are there. For each hypothesized generalization, a total description length can be computed:

$$L^{(k)}(S_{i,j}) = L(G_{i,j}^{(k)}) + L(N_{i,j}^{(k)}) \quad k = 0, 1, 2 \tag{54}$$

where $N_{i,j}^{(k)} = S_{i,j} - G_{i,j}^{(k)}$. $L(N_{i,j}^{(k)})$ is calculated with formula (50).

$L(G_{i,j}^{(k)})$ can be easily formulated as its contents are known discrete values. As p_i and p_j are common to all three hypotheses, they need not to be considered for comparison. Let $w_{i,j}$ and $h_{i,j}$ be the width and height of the bounding box of all the points in $S_{i,j}$, assume each of points p_{s_1} and p_{s_2} to be a random point within this bounding box, then

$$L(G_{i,j}^{(0)}) = 0 \tag{55}$$

$$L(G_{i,j}^{(1)}) = lbw_{i,j} + lbh_{i,j} \tag{56}$$

$$L(G_{i,j}^{(1)}) = 2(lbw_{i,j} + lbh_{i,j}) \tag{57}$$

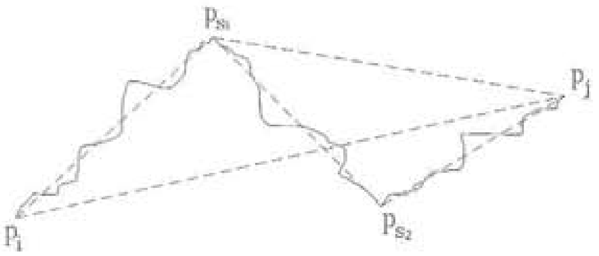


Figure 3. Alternative hypotheses of line generalization

The decision is now to select one from the three hypotheses with smallest description length $L(S_{i,j})$. If the hypothesis 0 is selected, then stop for this local $S_{i,j}$, otherwise, $S_{i,j}$ is split into two or three new subseries. Starting from $i = 1$ and $j = n$, the original series S is first split into two or three subseries. Each new subseries can be tested again for further splitting. The test with a selection of the hypothesis 0 is a hard criterion to stop a local split. This recursive splitting of $S_{i,j}$ is a gradient descendent approach to reach the minimization of the total description length of the original series S together with the final series G as the most probable model. Fig. 4 shows an example produced by using this mechanism. The input is taken from Fig. 2(b).

In comparison with the well-known recursive splitting algorithm of Douglas and Peuker[2], this algorithm is objective without requiring any subjective control parameters. This is a major advantage.

IV. MDL-BASED ANALYSIS OF SPATIAL DATA IN GIS

Spatial data in GIS refer to those quantitative values that are spatially indexed, and describe some physical or cultural aspects of the earth surfaces. Spatial data may come from different information sources such as photogrammetric production, map digitization, field survey, etc. Redundancy, noise, and gross errors may exist in a spatial data set, which should be removed or identified through data modeling. In these cases, the MDL criterion is usable and useful for modeling or compression of raw spatial data. In the following, two apparently distinct problems will be studied: (1) to search for the best model of a digital terrain, (2) to seek the best way of indexing spatial data. It must be pointed out that these two problems are not new, however our formulation of each problem under the MDL criterion opens a new view and potential solution which may be pursued in the future.

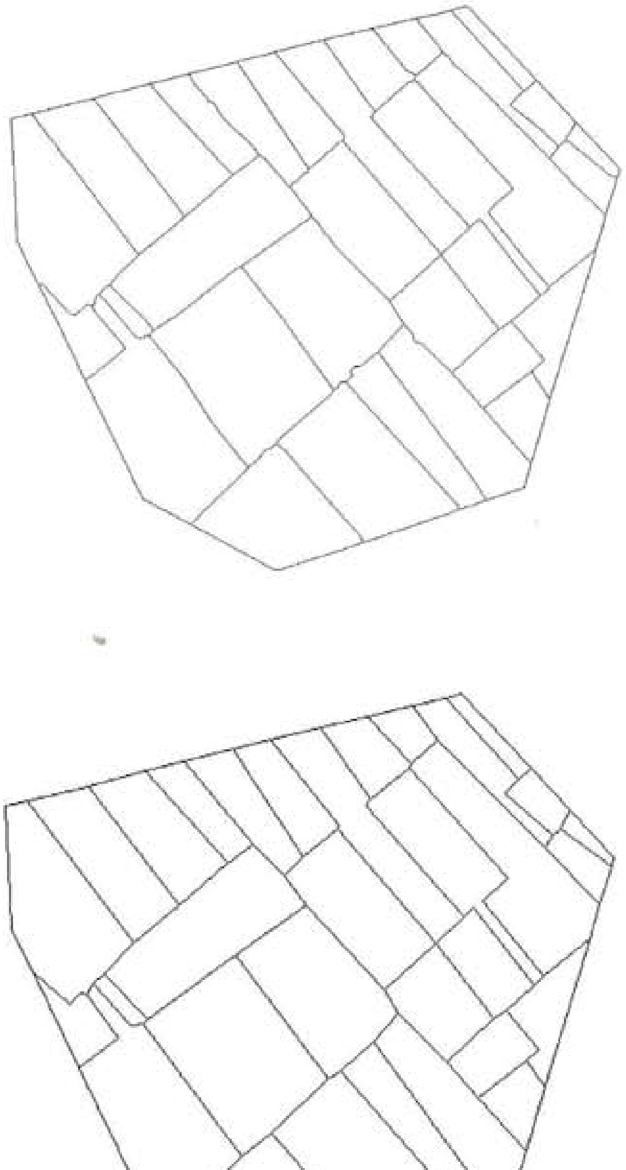


Figure 4. Generalization of edges in Fig.2(b) under correlated (top) or uncorrelated noise model (bottom)

A. Digital Terrain Modelling

A raw data set D of a digital terrain in general takes the form of a collection of three-dimensional points:

$$D = \{p_i = (x_i, y_i, z_i) \mid i = 1, 2, \ldots, N; \quad 0 \le x_i \le a; 0 \le y_i \le b; 0 \le z_i \le c\} \quad (58)$$

where a, b, and c are three positive real constants. It is generally assumed that the data points are distributed randomly, though denser point groups may corresponds to finely detailed landform variations. Let ε_x, ε_y, and ε_z be the resolution of the coordinates x, y, and z, the length of direct encoding the data set D, according to formula (30) and (31) is

$$L(D) = \sum_{i=1}^{N} (L(x_i) + L(y_i) + L(z_i)) = \sum_{i=1}^{N} \left(lb\left[\frac{a}{\varepsilon_x}\right] + lb\left[\frac{b}{\varepsilon_y}\right] + lb\left[\frac{c}{\varepsilon_z}\right] \right) \quad (59)$$

Digital terrain modeling refers to a complete description of a digital surface by using the raw point set D. A straightforward and the stablest surface model is the Triangulated Irregular Network (TIN) of this terrain which is generated through distance transform (Voronoi graph transform) from the points of D.

The primitive model M_0

Given a point set D a TIN all whose nodes coincide with all the points in D is unique to D, and the transformation is a deterministic procedure. We denote this unique TIN by $M_0(D)$ as the most primitive model of D. We naturally assume this procedure is provided by our DTM description language, so it needs not to be encoded. Thus the description length of D including $M_0(D)$ should be

$$L(D, M_0) = L(D \mid M_0) + L(M_0) = L(D) \quad (60)$$

There are many alternative ways to describe such a digital terrain based on the raw data set D. In the following, we describe two major approaches of digital terrain modeling: the generalized TIN and the regular lattice.

The generalized TIN model M_1

It is generally true that high redundancy exists in the primitive TIN model $M_0(D)$ for a given D. If we keep those points D_1 of D which are characteristic in describing the surface, and then generate a TIN from the points in D_1, this TIN is, in fact, a generalization of $M_0(D)$. With a proper interpolation scheme, each of all other points which are in D but not in D_1 can be interpolated from a triangular facet containing it. The total description length of D is then

$$\begin{aligned} L(D, M_1) &= L(D \mid M_1) + L(M_1) \\ &= \sum_{p_i \in (D-D_1)} L(p_i \mid M_1) + L(D_1) \\ &= \sum_{p_i \in (D-D_1)} (L(x_i \mid M_1) + L(y_i \mid M_1) \\ &\qquad + L(z_i \mid M_1)) + L(D_1) \\ &\qquad = \sum_{p_i \in (D-D_1)} (L(x_i) + L(y_i) \\ &\qquad + L(z_i \mid M_1)) + L(D_1) \qquad (61) \end{aligned}$$

Compare formula (61) with (60) and (59), the difference

$$\begin{aligned} L(D, M_0) - L(D, M_1) &= L(D) - L(D, M_1) \\ &= \sum_{p_i \in (D-D_1)} (L(z_i) - L(z_i \mid M_1)) \qquad (62) \end{aligned}$$

where $L(z_i \mid M_1)$ is the number of bits to encode the z-coordinate of the $i-th$ point given the generalized TIN model M_1. As x_i and y_i are encoded ordinarily as in M_0, there is a deterministic procedure to determine in which triangle $T_j = (p_{j1}, p_{j2}, p_{j3})$ a known position (x_i, y_i) is located. And, this procedure is already provided in our description language. However, we need one bit to indicate that p_i is not in the set D_1. If we use a triangular plane model for interpolation, let $z_i^{(1)}$ denote the simple interpolated z-coordinate for the position (x_i, y_i), the residual will be $z_i - z_i^{(1)}$, therefore

$$L(z_i \mid M_1) = 1 + L(z_i - z_i^{(1)}) \qquad (63)$$

Therefore, the difference (62) in encoding length with model M_1 and M_0 is computable. The necessary condition to accept the alternative model M_1 is

$$L(D) - L(D, M_1) > 0 \qquad (64)$$

In general, there may be many different choices of M_1 from D, the best choice $\hat{M}_1$ should satisfy

$$\begin{aligned} L(D) - L(D, \hat{M}_1) &= \max_{M_1}\left(L(D) - L(D, M_1)\right) \\ &= \max_{M_1}\left(\sum_{p_i \in (D-D_1)} (L(z_i) - L(z_i \mid M_1)) \right) \qquad (65) \end{aligned}$$

Regular lattice model M_2

Suppose we divide the rectangular domain in the plane $(x \in [0,a], y \in [0,b])$ by a regular lattice with spacing s_x and s_y in $x-$ and $y-$dimension, so the $x-$ and $y-$coordinates can be replaced by i and j integer indices,

$$i = 0, 1, \ldots, n; \; n = \left[\frac{a}{s_x} \right] \qquad (66)$$

$$j = 0, 1, \ldots, m; \; m = \left[\frac{a}{s_y} \right] \qquad (67)$$

In this case, the model M_2 is a $(n+1) \times (m+1)$ matrix

$$M_2 = (z_{ij}) \qquad (68)$$

The description length of M_2 is

$$L(M_2) = L(s_x) + L(s_y) + \sum_{i,j=0}^{n,m} L(z_{ij}) \qquad (69)$$

Given M_2, it can be determined by a fixed procedure to which rectangular cell $R = ((i,j), (i,j+1), (i+1,j), (i+1,j+1))$ each point p_k of D is located. Therefore, its supposed height $z_k^{(2)}$ can be interpolated via certain well chosen and fixed interpolation scheme. The z-residual of the kth point is: $z_k - z_k^{(2)}$ The total description length of the data D including the model M_2 can be formulated, in the way similar to (61), as

$$\begin{aligned} L(D, M_2) &= L(D \mid M_2) + L(M_2) \\ &= \sum_{p_k \in D} (L(x_k) + L(y_k) + L(z_i \mid M_2)) + L(M_2) \\ &= \sum_{p_k \in D} (L(x_k) + L(y_k) + L(z_k^{(2)})) + L(M_2) \qquad (70) \end{aligned}$$

The difference in encoding length with M_2 and M_0 is

$$\begin{aligned} L(D, M_0) - L(D, M_1) &= L(D) - L(D, M_1) \\ &= \sum_{p_k \in D} (L(z_k) - L(z_k^{(2)})) - L(M_2) \qquad (71) \end{aligned}$$

The necessary condition to accept M_2 is

$L(D)-L(D,M_1)>0$, which requires

$$\sum_{p_k\in D}(L(z_k)-L(z_k^{(2)}))>L(M_2) \qquad (72)$$

or in detail,

$$\sum_{p_k\in D}(L(z_k)-L(z_k^{(2)}))$$

$$>L(s_x)+L(s_y)+\sum_{i,j=0}^{n,m}L(z_{ij}) \qquad (73)$$

The best spacings of grid $\hat{s}_x$ and $\hat{s}_y$ which lead to the best grid model $\hat{M}_2$, can be determined via an optimization procedure:

$$L(D)-L(D,\hat{M}_2)=\max_{M_2}(L(D)-L(D,M_2))$$

$$=\max_{s_x,s_y}(\sum_{p_k\in D}(L(z_k)-L(z_k^{(2)}))$$

$$-(L(s_x)+L(s_y)+\sum_{i,j=0}^{n,m}L(z_{ij}))) \qquad (74)$$

There may be many other ways of digital terrain modeling, e.g. a proper combination of TIN and regular grid. Each of these alternative model can be treated in the similar manner of reasoning. In case the interpolation function has local parameters associated with each triangle or polygon in order to achieve a high fidelity, the encoding length of these parameters should be taken into account in the description length of the model. In case different models are compared under the criterion of interpolation precision which is evaluated by using additional points of known height, these additional points should be considered as a part of the data set D.

It is worth mentioning that the wavelet transform may be used as a promising approach for digital terrain modeling. The optimal way of using an optimal wavelet package for compressing digital terrain data can also be determined by using the MDL criterion.

B. Spatial Indexing

Spatial indexing refers to encode the spatial data in a certain way in order to facilitate or accelerate the retrieval of information based on location, especially for large databases like GIS's. We only consider point-like spatial data. Suppose there is a set of data points

$$S=\{p_1,p_2,\ldots,p_n\} \qquad (75)$$

Each point p_i is determined originally by two coordinates (x_i,y_i) in a global reference system. These points may be distributed in groups. Among groups, there could be hierarchical relations. To reflect the hierarchical groupings in the point set S, generally three alternative indexing schemes may be used: R-Tree, sphere-tree, and cell-tree.

R-tree (Fig. 5) is a tree of minimum-bounding rectangles. Each rectangle r_j is determined by four parameters (x_j,y_j,w_j,h_j,f_j), where x_j,y_j are the coordinates of the top-left corner of this rectangle in the reference system of its super-rectangle, w_j,h_j are its width and height, f_j is the pointer to its super-rectangle. At the leaf level, each rectangle r_j is minimally bounding a group of points. The position of each point (x_i,y_i) may be considered as a random in the minimum-bounding rectangle r_j:

$$x_i=x_j+u_{ij}, \qquad o\le u_{ij}\le w_j \qquad (76)$$

$$y_i=y_j+v_{ij}, \qquad o\le v_{ij}\le h_j \qquad (77)$$

It is obvious that local coordinates (u_{ij},v_{ij}) of a point p_i in the rectangle r_j is shorter than its original coordinates (x_i,y_i). The total encoding length $L(S,T_r)$ of the point set S in a R-tree T_r will be

$$L(S,T_r)=\sum_{j=1}^{m}((L(x_j)+L(y_j)+L(w_j)$$

$$+L(h_j)+L(f_j))+\sum_{p_i\in r_j}(L(u_{ij})+L(v_{ij}))) \qquad (78)$$

where $p_i\in r_j$ only refers to the lowest-level rectangles r_j's.

To encode x_j,y_j,w_j, and h_j, these variable may be considered as from some distribution within the super-rectangle of r_j. $L(f_j)$ for all rectangles of different levels may be considered as a constant. A R-tree $\hat{T}_r$ of a data set S is optimal among all possible configurations if

$$L(S,\hat{T}_r)=\min_{T_r}L(S,T_r) \qquad (79)$$

A sphere tree T_s (Fig. 6) is a tree of spheres instead of rectangles. In the 2-D space, a sphere is degenerated to a circle, however, without losing generality, let us still call a tree of circles as a sphere tree. Each sphere s_j is determined by its radius r_j, the center coordinates (x_j, y_j) and the pointer f_j to its supersphere. The encoding length of a sphere is obviously shorter than that of a rectangle, but there is no guarantee that the encoding length $L(S, T_s)$ of a data set S in a sphere tree is shorter than that in a R-tree. $L(S, T_s)$ is calculated as

$$L(S,T_s)=\sum_{j=1}^{m}((L(x_j)+L(y_j)+L(r_j)$$
$$+L(f_j))+\sum_{p_i\in s_j}(L(u_{ij})+L(v_{ij}))) \qquad (80)$$

where (u_{ij}, v_{ij}) are the local coordinates of a point p_i in a lowest-level sphere s_j. They can be either Euclidean or spherical coordinates.

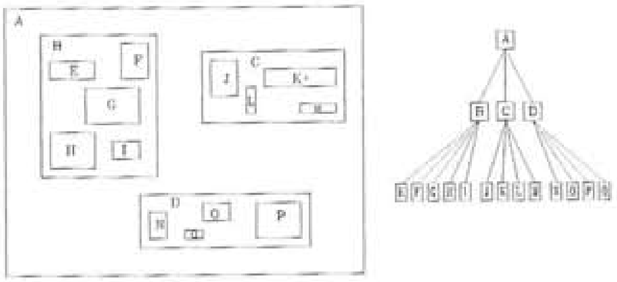


Figure 5. Rectangle tree for indexing data points.

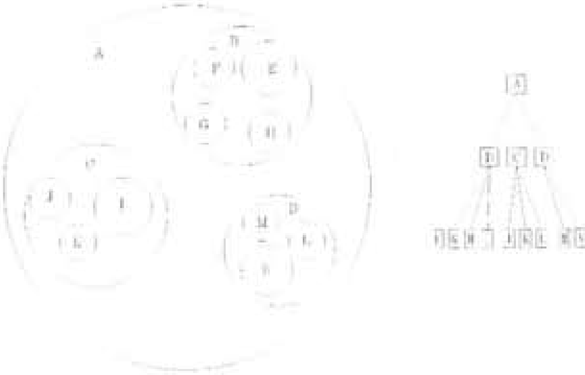


Figure 6. Sphere tree for indexing data points

A cell tree T_c is a tree of polygons (see Fig. 1) instead of rectangles or spheres. It is more complex to encode a polygon obviously. The simplest way to encode a polygon cell c_j is a series of its knotting points $[(x_{j1}, y_{j1}), (x_{j2}, y_{j2}), \ldots, (x_{jn_j}, y_{jn_j})]$. The encoding length of a data set S in a cell tree T_c is given by

$$L(S,T_c)=\sum_{j=1}^{m}(\sum_{k=1}^{n_j}(L(x_{jk})+L(y_{jk}))$$
$$+\sum_{p_i\in c_j}(L(u_{ij})+L(v_{ij}))) \qquad (81)$$

where $p_i \in c_j$ refers only to the lowest-level cells c_j's. Notice that spatial indexing is not only an organization of spatial data for information retrieval, but also an optimal indexing will reveal the hierarchical relations inherent in the data points. In this sense, an optimal spatial indexing T of a data set S of three alternative encodings T_r, T_s, T_c should be

$$L(S,T)=\min_{k=r,s,c} L(S,T_k) \qquad (82)$$

V. A NEW OBJECTIVE FUNCTION FOR UNSUPERVISED CLUSTERING

Unsupervised clustering is a classic topic in pattern recognition with direct application in multispectral classification in remote sensing and in other data analysis tasks of the same nature. Suppose there is a set of data points $S = \{x\}$ in the pattern space. Assume there are totally N_c clusters that have been discovered in the pattern space, and these clusters completely cover the set of data points S. Each i-th cluster includes a set S_i of N_i data points which is a subset of S. The mean vector of the set S_i is m_i:

$$m_i=\frac{1}{N_i}\sum_{x\in s_i}x \qquad (83)$$

In the most general circumstance of unsupervised clustering, the number of clusters and the mean vector for each cluster are unknown. In order to evaluate the performance of any clustering algorithm, one of the most often used criteria is to minimize the sum of the squared intraset distances [17], given by

$$F(N_c, m_1, m_2, \ldots, m_{N_c})$$

$$=\sum_{i=1}^{N_c}\sum_{x\in s_i}\|x-m_i\|^2 \qquad (84)$$

Careful consideration on this criterion reveals the following facts. On the one side, when N_c is fixed, the objective function F to be minimized leads to a reasonable performance which corresponds to the minimum intraset distances after clustering of the data set. However on the other side, when N_c is actually unknown, there may be a fatal collapse which is caused exactly by minimizing the function F: If we take each data point x as a cluster, then $F\equiv 0$, which is the absolute minimum.

At least in order to avoid this fatal collapse in theory, we must set up a constraint to the characteristics of clusters. In this sense, a new objective function L can be constructed based on the MDL criterion. Let M be the model representing all the characteristics of clusters: N_c, and $\{m_i | i=1,2,\ldots,N_c\}$. The new objective function is defined as

$$L(S,M)=L(S\,|\,M)+L(M)$$

$$=\sum_{i=1}^{N_c}\left[\left(\sum_{x\in s_i}L(x-m_i)\right)+L(m_i)\right] \qquad (85)$$

where L denotes the description length. Within the [] brackets, the first term is the description length of the data points within each cluster, which corresponds to the encoding length of the residuals $(S\,|\,M)$ given the model M; the second term is the description length of each cluster center, which corresponds to the encoding length of the model M. To compute the first term, we may assume a statistical distribution within each cluster. To compute the second term, we may use the prior probability of each cluster. If the prior probability is unknown or not computable, we may consider the way of encoding each cluster center as a point in the pattern space, which corresponds to our commonsense knowledge about the cluster centers.

VI. CONCLUSIONS

The MDL principle is a best established criterion for model selection and estimation. It is specially advantageous at its flexibility of comprising multiple sources of information and its uniform measure of the best decision in number of bits. Interpretation of digital images and analysis of spatially indexed data are the two fields where the MDL criterion can be best applied to demonstrate its full usefulness. The two examples in image analysis including a global interpretation of remotely sensed images for landuse mapping, and general line/edge generalization are proved and partially realized applications. The two examples in spatial data analysis are basically novel ideas for the classic problems in GIS. The new formulation of the DTM and spatial indexing problems may lead to new understanding or solutions of these problems. The objective function proposed for unsupervised clustering is novel in general pattern recognition, which is important to image classification in remote sensing and other data interpretation. The approach demonstrated in these examples is also applicable to all other problems in spatial informatics where multiple models or multiple data sources are involved and a somewhat best interpretation of spatial data is required.

ACKNOWLEDGMENT

This work was sponsored in part by the Deutsche Forschungsgemeinschaft. This paper was first drafted in the Institute for Photogrammetry, University of Bonn, Germany. Various discussions with Prof. Dr. Wolfgang Forstner on MDL and its applications in Geoinformatics were helpful in preparing the first draft. Comments from three anonymous referees were very valuable and helpful in finalizing this paper.

REFERENCES

[1] Chaitin G.J., 1977. Algorithmic information theory. *IBM Journal of Research and Development*, 21:350-359.

[2] Douglas D. H. and Peuker T.K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer* 10(2):112-122.

[3] Fua P. and Hanson A. J., 1989. Objective functions for feature discrimination: Theory. *Proc. DARPA*, Image Understanding Workshop.

[4] Georgeff M. P. and Wallace C. S., 1984. A general selection criterion for inductive inference. *Proc. Advance in AI*, Italy, Sept. 1984, T. O'Shea (Ed.), North Holland, Amsterdam.

[5] Kolmogorov A.N., 1964. Three approaches to the quantitative definition of information. *Probability and Information Transmission* 1(1).

[6] Leclerc Y.G., 1989. Constructing simple stable descriptions for image partitioning. *Int. Journal of Computer Vision* 3:73-102.

[7] Pan H.-P., 1994. Two-level global optimization for image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*. 49(2):21-32.

[8] Pan H.-P. and Forstner, W., 1992a. Polygon map grammars: a generic model for understanding landuse maps and images in remote sensing, Forstner and Ruwiedel (Eds.), *Robust Computer Vision*, Wichmann Verlag.

[9] Pan H.-P. and Forstner, W., 1992b. An MDL-principled evolutionary mechanism to automatic architecturing of pattern recognition neural network. *IEEE Proc. 11th Int. Conference of Int. Association of Pattern Recognition (IAPR)*, Vol. II.

[10] Pan H.-P. and Forstner, W., 1994. Segmentation of remotely sensed images by MDL-principled Polygon Map Grammar. *Int. Archives of Photogrammetry and Remote Sensing*, 30(3):648-655.

[11] Rissanen J. (1978): Modelling by shortest data description. *Automatic*, 14:465-471.

[12] Rissanen J., 1983a. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 2(11):211-222.

[13] Rissanen J., 1983b. Minimum Description Length Principle. *Encyclopedia of Statistical Sciences*, Vol. V.

[14] Rissanen J., 1989. *Stochastic complexity in statistical Inquiry*. Series in Computer Science, Vol. 15, World Scientific, Singapore.

[15] Shannon C. E. , 1948. A mathematical theory of communication. *Bell Syst Tech J.*, (3), pp.379-423.

[16] Solomonoff R., 1964. A formal theory of inductive inference I & II. *Infomation and Control,* 7:1-22, 224-254.

[17] Tou J.T. and Gonzalez R.C., 1974. *Pattern Recognition Principles*. Addison-Wesley, Reading, Massachusetts, p.89.