COENOSES 6 (1): 15-20, 1991
© 1990 by Edizioni LINT Trieste

THE MINIMUM DESCRIPTION LENGTH PRINCIPLE AS APPLIED TO MULTIVARIATE CLUSTERING ANALYSIS WITH ISODATA[1]

Qiong Gao, Department of Ecology, Institute of Botany, Academia Sinica Beijing, P.R. China 100044

Keywords: Clustering, Number of Classes, Minimum Description Lenght Principle, Fuzzy Sets, Vegetation

Abstract. The ISODATA clustering analysis based on the theory of fuzzy sets makes use of the concept of membership degree, and thus more objectively and reasonably handles the relation between individuals and their representative classes. However the number of classes into which all individuals are classified has to be predetermined. This requires certain insight into the structure and distribution of the data set to which the analysis is to be applied, otherwise arbitrariness and subjectivity are inevitabe to be involved in the results of the analysis. It is natural and reasonable to think that the number of classes should be determined according to the behavior of the observed data set. In the light of the Minimum Description Lenght Principle, developed in the field of statistical modeling, an optimal number of classes may be identified. The method has been implemented in the general purpose computer software, with sample analyses included.

Introduction

One of the important objectives of quantitative vegetation ecology research is to relate the distribution of plant species in the community to the environmental factors, and hence to discover the laws or regulations of plant distribution. The observed data set for an area or plant community often can be represented by a matrix **X** with element X_{ij} being the value of the j'th attribute of the i'th individual in a total of n individuals each with m attribute. The attribute value, for example, can be the percent of plant cover, the abundance or the importance value of a plant species in a sample. Multivariate clustering is to classify the n individuals according to their attributes into a number of representative classes, thus to compress the data for the purpose of futher interpretation (Gauch 1982, Digby and Kempton 1987, Chang 1985).

If the predetermined number of classes is C, the basic idea of ISODATA (Interactive Self-Organizing Data Analysis Technique A) is to construct a partition matrix $\mathbf{R}=R_{ij} \in \{0, 1\}$ for i = 1, 2, ..., C and j = 1, 2, ..., n, and a cluster center matrix $\mathbf{V}=V_{ik}$ for i = 1, 2, ..., C and k = 1, 2, ..., m, to minimize the following residual function J,

$$J=\sum_{i=1}^{C}\sum_{j=1}^{n}R_{ij}\sum_{k=1}^{m}(X_{jk}-V_{ik})^2 \quad (1)$$

where R_{ij} is 1 if individual j belongs to class i, and 0 otherwise. The partition matrix R_{ij} has the following properties (Bezdek 1987): 1) The sum over the second index j is not less than 1, implying that each class has at least one individual; 2) The sum over the first index i is 1, i.e., only one element in each column is 1, and the rest of the elements are 0's. This is because each element can only belong to one class.

The above clustering method is suitable when the observed data are indeed clustered around some center points, i.e., the data points are more densely scattered around the centers in the multidimensional spaces, reflecting the discontinuity of distribution of the individuals in the attribute space. However, the situation in reality is usually complicated and the sample data set, or even the whole population, sometimes has both continuous and discontinuous characteristics in distribution. In a certain local region it may appear to be continuous, but discontinuous globally in the whole area of interest. The opposite case, i.e., discontinuous locally but continuous globally, is also not rare. For instance, if local factors such as soil composition vary only in a moderate range in a large area, the vegetation may look continuously distributed. In contrast, on a steep mountain slope with considerable altitude range, the discontinuity is often more evident. To handle the interlace of continuity and discontinuity, the classic theory of sets, base on which ISODATA was developed, often can only give poor clustering results.

The modified ISODATA clustering (Dunn 1974, Bezdek 1976) introduced the "fuzzy" concept into the relation between the individuals and the respective classes. Specifically, the elements of partition matrix R_{ij} is generalized into real domain [0, 1], rather than

[1] This work was supported by The National Laboratory of Quantitative Vegetation Ecology, Academia Sinica under Grant No. 89003

the two integer set {0, 1}. The extension of R_{ij} implies that an individual can belong to all classes in the set simultaneously to different extent, with values 0 and 1 being the special cases meaning "absolutely not belong to" and "absolutely belong to" respectively. Similar to the original partition matrix, **R** has the following properties: 1) The sum of all elements in a row is larger than 0; 2) The sum of all elements in each column is 1. Therefore **R** describes the membership degree of a individual with respect to the respective classes.

The partition matrix **R** is usually obtained using an iterative method with asymptotic modification of both **R** and **V** until they become stabilized.

The modified ISODATA analysis is expected to perform better when both continuous and discontinuous distributions are involved in the sample data, and thus give more objective clustering result. One difficulty in using ISODATA clustering is that the analyst is asked to give the number of classes into which all the individuals are classified before the analysis starts. While there should be an optimal or most appropriate number of classes in nature determined by the structure of the observed data set, the problem is that the analyst often does not have much idea about how large it should be because the clustering is often the first stage of data processing and at that moment he or she cannot have much idea about the data. Therefore, the determination of this parameter until now was often more or less a subjective and arbitrary aspect of the analysis, resulting in some subjectivity in the clustering results. The fuzzy entropy of the clustering analysis takes the following form (Dunn 1974):

$$H_f = -\frac{1}{C} \sum_{i=1}^{C} \sum_{j=1}^{n} R_{ij} \log R_{ij} \qquad (2)$$

where H_f is the fuzzy entropy. Small H_f indicates a better result of clustering. However this quantity obviously cannot serve as a standard for selection of C, because it will approach zero when C increases to n since R_{ij} will approach either 1 or 0 when each individual belongs to a different class, and the clustering result is thus obviously trivial.

The objective of this work is to apply MDLP (Minimum Description Length Principle, Rissanen 1978, 1983) to the clustering problem to determine the optimal number of classes from the data set.

The minimum description leght principle

MDLP has its roots in the well known Bayesian Inference which, in turn, is purely based the Bayes formula:

$$P\ (H_i|D) = \frac{P\ (D|H_i)\ P\ (H_i)}{P\ (D)} \qquad (3)$$

where D denotes the data set and H_i the ith hypothesis or theory in a given exclusive and complete set of hypotheses or theories about the population from wich D is extracted; P (D) is the probability of observing data set D; P $(H_i|D)$ is the final or a posteriori probability, and is actually the conditional probability that H_i is the true theory for the population when observing data set D. Similarly P $(D|H_i$ is the conditional probability of observing data set D, given H_i is true. Finally P (H_i) is the so called prior probability which denotes the chances for H_i to be the true theory among a given set of theories. The theory of Bayesian inference states that for a given set of candidate hypotheses or theories describing the mechanism of the population generating the data, if there exists an H_i rendering P $(H_i|D)$ a maximum compared to P $(H_j|D)$ for any j not equal to i, then H_i is the best approximation of the true hypothesis or theory about the mechanism behind the observed data set.

The dilemma to apply Bayesian inference is the determination of the prior probability P (H_i). This quantity often can only be determined subjectively by querring, and as a result the inference loses its reliability. The maximum likelihood method completely ignores the prior probability and directly maximizes P $(D|H_i)$ in order to get around the problem. A number of approaches have been proposed to approximate the prior probability (Solomonoff 1964, Jeffreys 1961). Rissanan (1978, 1983) suggested that the prior

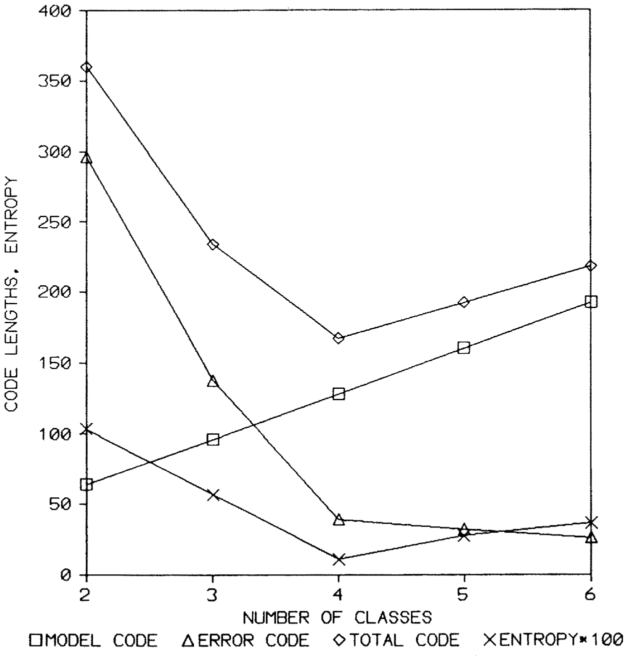

Fig. 1. Optimization for 4-Centre Data Set.

probability of a particular hypothesis or theory is related to the length of the binary codes which represents the theory in a universal computing machine, the so-called Turing mechine. Roughly speaking, a more complicated theory requires lengthier codes to describe and has less chance to occur in the real world. Specifically, taking the negative logarithm of Equation (3), we have

$$-\log_2 [P(H_i|D)] = \log_2 [P(D|H_i)] - \\ -\log_2 [P(H_i)] + A_1 \qquad (4)$$

where A_1 is a constant. The first term on the right is called *Data* or *Error Code Length* and the second is the *Theory* or *Model Code Length*. The left hand side is the *Total Code Length* or *Total Description Length*. Therefore Rissanen successfully related all the probabilities in the Bayesian formula to their corresponding descriptive code lengths in the machine. Beside the theoretical bases relating the description lengths to their corresponding probabilities, here is an informal argument: Suppose there are N independent, exclusive events each with an equal probability P = 1/N, the code needed to distinguish the N events from each other, or to represent N unique events in machine is about $\log_2$ (N) bits long, i.e., $-\log_2$ (P). On the other hand, if a model needs M bits to represent in the machine, its prior probability is roughly 2^{-M} (Jeffreys 1964).

The problem of maximizing the posteriori probability becomes minimizing the total description length; the minimum description length principle states that for a given candidate set of theories or models to describe the mechanism generating the data set, the one with the minimum total description length (or minimum code representation in the machine) will be the best theory or model. To understand this argument, suppose we are to select the best model to describe a given data set from a series of models which have different complexities. If a model is too simple, it may fail to capture the essence of the data, resulting in too much error. This situation reflected in Equation (4) is a small model code length with a large error code length. On the other hand, if a model or theory is too complicated and trying to include everything in the data, corresponding to the situation of a large model code length and a small error or data code length in Equation (4), the model is also not necessarily superior to the simpler models because the data set is just a sample from the population with random errors and the complicated model may be too sensitive to random errors and irregularities when used to predict the future data from the same population, since it has included some of the random error from the observed data set. Between these two extremes, MDLP says that the model of minimum description length will be the best approximation to the mechanism generating the data. An alternative way to elucidate MDLP is as follows: if there are two models of the same complexity, MDLP gives preference to the one with smaller error code length, i.e., smaller error in describing the data. On the other hand, the principle will select a simpler model if the two models of different complexities give the same error with respect to the observed data set.

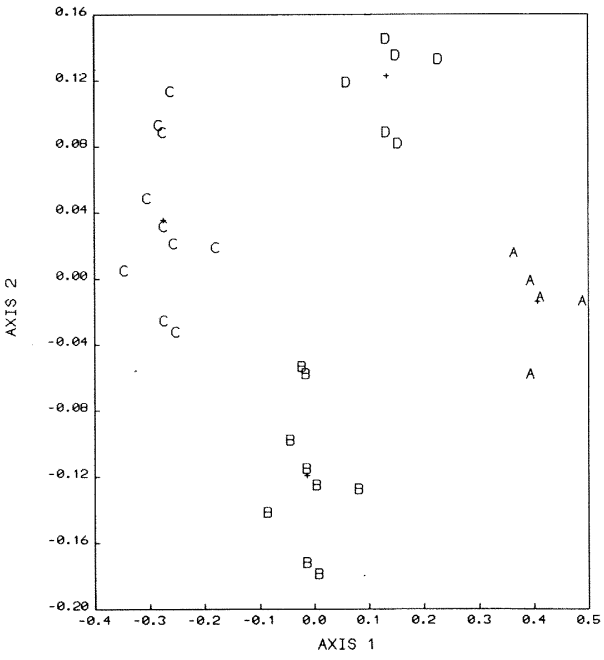

Fig. 2. Classification for 4-Centre Data Set.

The MDLP model in ISODATA clustering analysis

Let us suppose there are n points in the m dimensional space normally distributed around C center points in the same space. Each axis in the space is an attribute variable and the data points are the individuals of the sample data. Furthermore, the distribution of all attributes are independent of each other. If the data point vector in the space is $\mathbf{X}_i = \{X_{ik}\}$ (i = 1, 2, ..., n, k = 1, 2, ..., m), the center point vector is $\mathbf{V}_j = \{V_{jk}\}$ j = 1, 2, ..., C; k = 1, 2, ..., m), and the components of these two quantities can only be represented in machine with a finite number of binary digits, then the probability for X_i to occur with V_j as its center can be expressed as

$$P(\mathbf{X}_i|H_c, \mathbf{V}_j) = \frac{\delta V}{B} \exp\left[-\frac{1}{2} \sum_{k=1}^{m} \left[\frac{X_{ik} - V_{jk}}{\sigma_k} \right]^2 \right] \quad (5)$$

where H_c denotes the hypothesis of C clustering centers; P ($\mathbf{x}_i|H_c$, $\mathbf{V}_j$) is the conditional probability for $\mathbf{X}_i$ to occur given H_c and $\mathbf{V}_j$; B is the normalizing costant; δV is the volume of a small hypercube in the m dimensional space and is related to the precision of the representation of $\mathbf{X}_i$ and $\mathbf{V}_j$; and δ_k is the standard deviation of the distribution of the k'th attribute variable.

Following Equation (4), we have the total error code length

$$EL = \log_2(e) \sum_{i=1}^{n} \sum_{j=1}^{C} \delta_{ij} \sum_{k=1}^{m} \frac{1}{2} \left[\frac{X_{ik}-V_{jk}}{\sigma_k} \right]^2 + A_2 \quad (6)$$

where δ_{ij} is 1 if the ith individual is classified to the jth class, otherwise it will be zero; under the concept of fuzzy sets, $\delta_{ij} = R_{ij}$; and A_2 is another constant.

It is more reasonable to use t-distribution instead of normal distribution since δ_k's are unknown. Thus the error code length will be:

$$EL = -\sum_{i=1}^{C} \sum_{j=1}^{n} \delta_{ij} \sum_{k=1}^{m} \quad (7)$$

$$\log_2 \left[t_{n-C} \left[\frac{X_{jk}-V_{ik})}{S_k(C)} \right] \right] + A1$$

where $S_k(C)$ is

$$S_k(C) = \left[\frac{1}{n-c} \sum_{i=1}^{C} \sum_{j=1}^{n} \delta_{ij} (X_{jk}-V_{ik})^2 \right]^{1/2} \quad (8)$$

The assumption of distribution independence among attribute variables needs justification. In general, the attribute variables are very likely to be related to each other. As a result, covariances will be involved in the probability density. This can be avoided by tâking a coordinate transformation such as the principal component analysis and using the principal coordinates as the new attribute values for clustering. One additional benefit of doing so is that the dimensionality of spaces can be considerably reduced because most of the data variation can be explained by the first few axes, making the rest of the components negligible (Digby and Kempton 1987).

The model of fuzzy ISODATA clustering is a set of clustering centers $\mathbf{V}_j$, j = 1, 2, ..., C. If each element V_{jk} of the clustering center matrix $\mathbf{V}$ is represented by N_b bits in the machine, then the total model code length will be

$$ML = N_b \, C \, m \quad (9)$$

where N_b determines the precision of the clustering centers. Larger N_b gives higher precision for V_{jk} and hence more complex model and longer model code; ML is the total model code length and is proportional to the number of classes C. As the number of classes increases the model gets more complex and more codes are needed to represent it. The sum of ML and EL is the total code length DL.

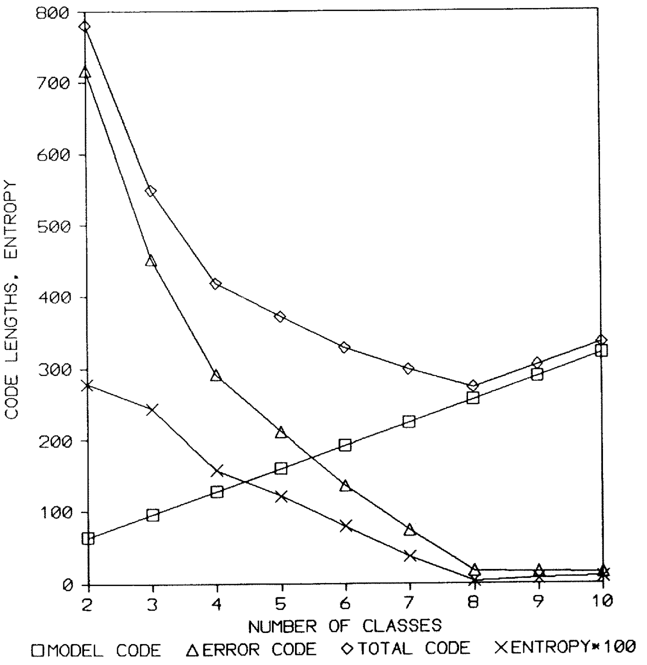

Fig. 3. Optimization for 8-Centre Data Set.

As C increases, the error code length EL generally decreases because the distance between the data points and their respective class centers decreases as the number of classes increases. However, the model code length ML linearly increases with C. MDLP provides a balance between these two terms (A_2 is dropped in the process minimizing DL). Between the two extremes of very large and very small C, there must be a value of C, denoted as C_{min} which renders the total description length DL a minimum DL_{min}. According to MDLP, C_{min} is the optimized number of classes.

If Equation (6) is used for EL instead of (7). The application of MDLP model will need the determination of the variances σ_k^2 for all attribute variables. Theoretically these quantities must be estimated from the random deviation of the individuals from their clustering centers. To do so, we compute S_k (C) for C = 1, 2, ... In general S_k contains both the deterministic and random variation. At C = 1, this quantity contains the total variation in the data. When the number of classes increases from 1, S_k will rapidly decrease as the deterministic part of the variation is gradually transferred to the variation between classes. The decreasse of S_k will diminish when most of the deterministic variation is eliminated from S_k (C), and the value of S_k (C) at that point can be regarded as the estimate of σ_k^2. In particular for the current work, if

$$|S_k(i+1) - S_k(i)| \leq 0.1 |S_k(2) - S_k(1)| \quad (10)$$

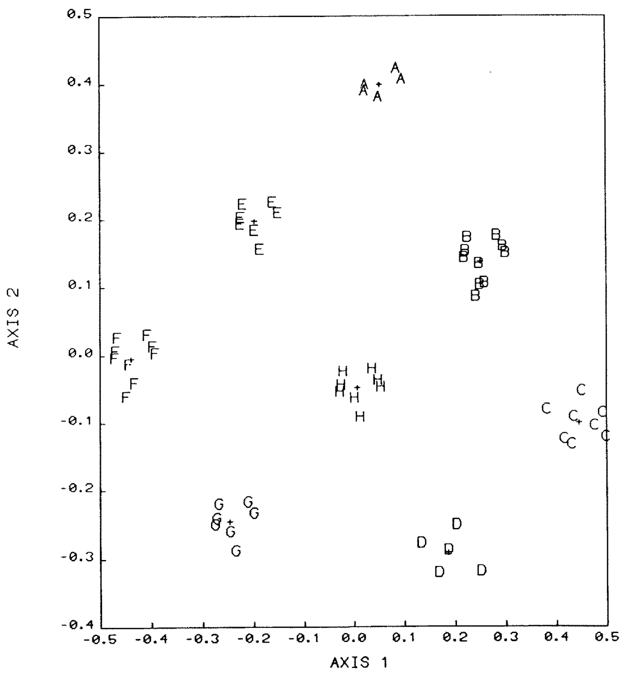

Fig. 4. Classification fo 8-Centre Data Set.

is satisfied, then we take $\sigma_k = S_k$ (i+1), where S_k (i) denotes the value of S_k (C) when C=i.

The implementation of the model

The author has implemented the above MDLP model in a small system using TURBO C under MS DOS environment. The specific algorithm of computation is as follows:

1. Reading in the data matrix **X** and normalizing all the variables according to the following equation

$$Y_{ik}=\frac{X_{ik}-X_{min\ k}}{X_{max\ k}-X_{min\ k}} \qquad (11)$$

where Y_{ik} is the element of normalized data matrix **Y**; $X_{max\ k}$ and $X_{min\ k}$ are respectively the maximum and minimum values of variable k in the original data matrix.

2. Applying principal component analysis to the transformed data matrix **Y**. As stated above, the application of PCA is to make the new attributes linearly independent of each other and to reduce the amount of computation in the successive clustering stage.

Only the first t principal components (t is given by user) will be used as attributes in the subsequent clustering analysis and the rest axes are discarded.

To do the clustering analysis after the principal component analysis, we need to replace X_{ik} by the principal component Z_{ik} and m by t in Equations (1), (5), (6), (7), (8) and (9).

3. Increasing C starting from 1, and for each value of C do an ISODATA clustering, i.e. find R_{ij} and V_{ik} for all i, j and k, and compute S_k (C). The following two equations are used to compute iteratively R_{ij} and V_{ik} (Bezedk, 1980).

$$R_{ij}=\left[\sum_{p=1}^{C}\left[\frac{\sum_{k=1}^{t}(Z_{jk}-V_{ik})^2}{\sum_{k=1}^{t}(Z_{jk}-V_{pk})^2}\right]^{\frac{1}{\mu-1}}\right]^{-1} \qquad (12)$$

$$V_{ik}=\frac{\sum_{j=1}^{n}(R_{ij})^{\mu}Z_{jk}}{\sum_{j=1}^{n}(R_{ij})^{\mu}} \qquad (13)$$

for each C, the iteration procedure to obtain **R** and **V** starts with random initialization of **V** and proceeds with iteration using equation (12) and (13) until both **R** and **V** become stablized within a given tolerance.

As C increases, the change of S_k is traced until (10) is satisfied and σ_k is determined for each k.

4. Computing the total code length according to Equation (6) or (7) and Equation (9) and the entropy according to (2) for each value of C. The iteration stops when a particular model code length is larger than the smallest total length before that particular C. The minimum code length and the corresponding number of classes are then identified from the iteration result.

5. Printing out the clustering results for the optimum number of classes.

Analyses of computer generated sample data sets

As a demonstration of the above implemented MDLP for ISODATA clustering, the author used the computer's random number generator to make up two artificial data sets, each with two attribute variables. The procedure was as follows: 1) choose a number of centers on the plane (4 centers for the first data set and 8 for the second); 2) use the random number generator to generate 31 random number of with normal distribution in certain ranges for the first data set and 59 random numbers of the same property for the second; 3) Add these random numbers to the centers to produce the data sets. Therefore the two data sets were truely made up to satisfy the basic assumptions of the theory. If MDLP is successful to recover true mechanisms generating the data sets, it should be able to identify the correct number of classes for the two data sets respectively. The two data sets were analyzed with the MDLP algorithm using $\mu=1.7$ and $N_b=32$ (4 bytes for each real number in the real computer).

The results of the analyses are in Figure 1, 2, 3 and 4, which show that the MDLP worked perfectly well as expected. Figure 1 and 3 show that the optimization process in terms of MDLP using Equation (6) to compute error code length for the twod ata sets.

As described in the theory above, the model code lengths increased linearly with C. The error code lengths, on the other hand, decreased with C. At $C=4$ for the first data set (Figure 1) and $C=8$ for the second data set (Figure 3), the total code lengths reached their respective minimum, indicating that the $C=4$ and $C=8$ are the best approximations to the true mechanisms behind the two data sets. The classifications using $C=4$ and $C=8$ for the two data sets are shown in Figures 2 and 4 respectively. The classification was done by converting the largest R_{ij} into 1 and the rest into 0's in the jth column of $\mathbf{R}$ for $j=1, 2, ..., n$, i.e., individual j was regarded to belong to the class to which it has the largest membership degree.

REFERENCES

BEZDEK, J.C. 1987. *Some non-standard clustering algorithms.* In P. Legendre and L. Legendre (eds): Developments in numerical ecology, Springer, New York, pp. 225-287.

BEZDEK, J.C. 1980. *A convergence theorem for the fuzzy ISODATA clustering algorithms*, IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. PAM I-2, no. 1, pp. 1-8.

BEZDEK, J.C. 1976. *A physical interpretation of fuzzy ISODATA*, IEEE Transactions SMC, vol. 6, no. 5, pp. 387-389.

CHANG, H.S. 1985. *The Multivariate Analysis of Vegetation and Environmental Factors in Ngari, Tibet. Ph. D. Thesis*, Cornell University.

DIGBY, P.G.N. and R.A. KEMPTON. 1987. *Multivariate Analysis of Ecological Communities.* Chapman and Hall, London.

DUNN, J.C. 1974. *A fuzzy relative of ISODATA process and its use in delecting compact well-separated clusters*, J. of Cyber, vol. 3. pp. 32-57.

GAUCH, H.G. Jr. 1982. *Multivariate Analysis in Community Ecology.* Cambridge University Press.

JEFFREYS, H. 1961. *Theory of Probability.* 3rd Edition. Clarendon Press, Oxford.

RISSANEN, J. 1978. *Modelling by Shortest Data Description.* Automatica. 14: 465-471.

RISSANEN, J . 1983 . *A Universal Prior for Integers and Estimation by Minimum Description Length.* The Annals of Statistics. 11: 416-431.

RISSANEN, J. 1985. Minimum Description Length Principle. In: S. Kotz and N.L. Johnson, eds., *Encyclopedia of Statistical Sciences*, 5, pp. 523-527. Wiley, New York.

SOLOMONOFF, R. 1964. *A Formal Theory of Inductive Inference.* Information and Control. Vol. 7, pp. 1-22 and pp. 223-254.

Manuscript received: November 1990