united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

H4.SMR/1519-9

"Seventh Workshop on Non-Linear Dynamics and Earthquake Prediction"

29 September - 11 October 2003

Pattern Recognition Methods and their Applications to Geophysical Problems

V. Keilis-Borok & A. Soloviev

International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Sciences
Moscow, Russia

www.mitp.ru

I. INTRODUCTION

Pattern recognition is a useful tool for the analysis of behaviour of nonlinear complex systems in absence of fundamental equations describing them. Using this methodology creates possibility for a so-called "technical" analysis that involves a heuristic search for relationships between available system information and its features inaccessible for direct measurements.

Let a set of objects, phenomena or processes, which are connected with the same or similar systems, is considered. Certain information (for example, results of measurements) is available about each element of the set, and there is some feature, possessed only by a part of the elements. If possessing this feature by an element does not present evidently in the information available, then a problem arises to distinguish elements that possess this feature. This problem could be solved by constructing a model on the basis of mechanical, physical, chemical or other scientific laws, which could explain the relationship between the available information and the feature under consideration. But in many cases the complexity of the system makes the construction of such model difficult or practically impossible and it is natural to apply pattern recognition methods.

1.1 Examples of Problems to Apply Pattern Recognition Methods

Recognition of earthquake-prone areas (e.g., *Gelfand et al.*, 1976). A seismic region is considered. The problem is to determine in the region the areas where strong (with magnitude $M \geq M_0$ where M_0 is a threshold specified) earthquakes are possible. The objects are the selected geomorphological structures (intersections of lineaments, morphostructural nodes, etc.) of the region. The possibility for a strong earthquake to occur near the object is the feature under consideration. The available information is the topographical, geological, geomorphological and geophysical data measured for the objects.

The problem as the pattern recognition one is to divide the selected structures into two classes:

- structures where earthquakes with $M \geq M_0$ may occur;
- structures where only earthquakes with $M < M_0$ may occur.

Intermediate-term prediction of earthquakes (e.g., *Keilis-Borok and Rotwain*, 1990). A seismic region is considered. The problem is to determine for any time t will a strong (with magnitude $M \geq M_0$ where M_0 is a threshold specified) earthquake occur in the region within the period $(t, t + \tau)$. Here τ is a given constant. The objects are moments of time. The occurrence of a strong earthquake in time period τ after the moment is the feature under consideration. The available information is the values of functions on seismic flow calculated for the moment t.

The problem as the pattern recognition one is to divide the moments of time into two classes:

- moments, for which there is (or will be) a strong earthquake in the region within the period $(t, t + \tau)$;
- moments, for which there are not (or will not be) strong earthquakes in the region within the period $(t, t + \tau)$.

Recognition of strata filled with oil. The strata encountered by a borehole are considered. The problem is to determine what do the strata contain: oil or water. The objects are the strata. The filling of the strata with oil is the feature under consideration. The geological and geophysical data measured for the strata are the available information.

The problem as the pattern recognition one is to divide the strata into two classes:

- strata, which contain oil;
- strata, which contain water.

Medical diagnostics. A specific disease is considered. The problem is to diagnose the disease by using results of medical tests. The objects are examined people. The disease is the feature under consideration. The available information is the data obtained through medical tests.

The problem as the pattern recognition one is to divide examined people into two classes:

- people who have the disease;
- people who do not have it.

1.2 General Formulation of the Pattern Recognition Problem

One may give the general abstract formulation of the problem of pattern recognition as follows.

The set $W = \{ \mathbf{w}^i \}$ is considered, where objects $\mathbf{w}^i = (w_1^i, w_2^i, \ldots, w_m^i)$, $i = 1, 2, \ldots$ are vectors with real (integer, binary) components. Below these components will be called functions.

The problem is to divide the set W into two or more subsets, which differ in certain feature or according to clustering themselves.

There are two kinds of pattern recognition problems and methods:

- classification without learning;
- classification with learning.

1.3 Classification without Learning (Cluster Analysis)

The set W is divided into groups (clusters, see Fig. 1) on the basis of some measure in the m-dimensional space $w_1, w_2, \ldots, w_m$.

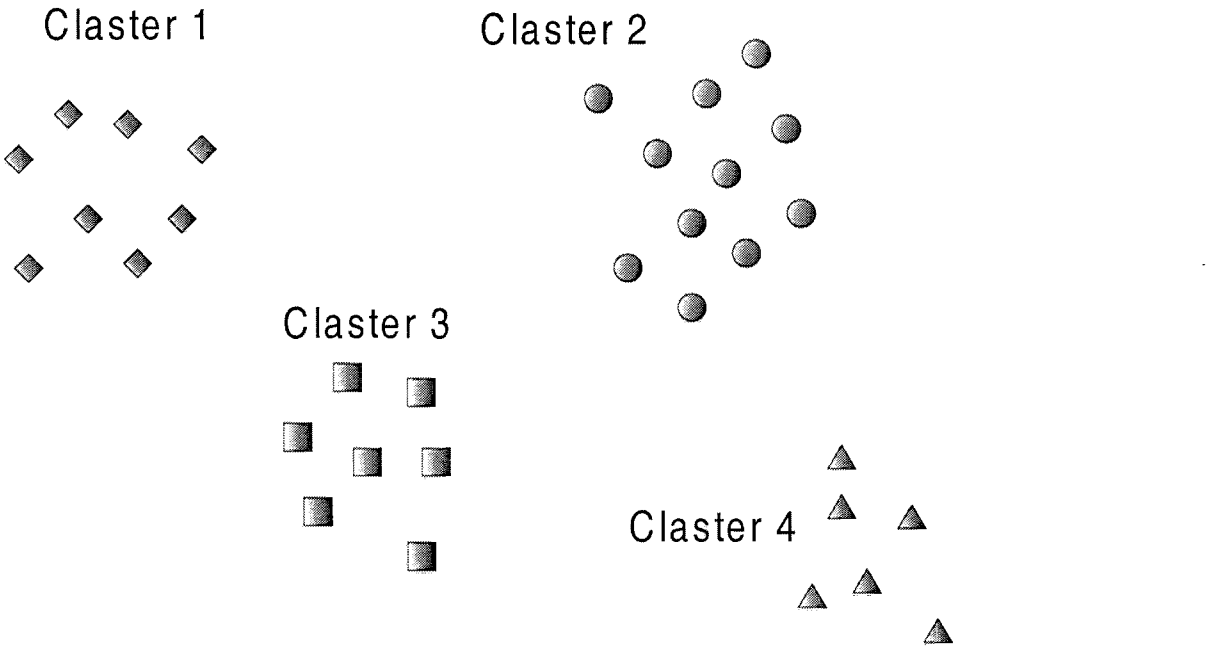

FIGURE 1 Clustering of objects in two-dimensional space

Denote $\rho(\mathbf{w}, \mathbf{v})$ a distance between two m-dimensional vectors $\mathbf{w} = (w_1, w_2,..., w_m)$ and $\mathbf{v} = (v_1, v_2,..., v_m)$.

To define classification and to estimate at the same time its quality the special function is introduced. The best classification gives the extremum of this function.

Examples of the functions. Let W is a finite set. The following two functions can be used.

$$J_1 = \frac{(K-1)\sum_{k=1}^{K} \rho_k}{2\sum_{k=1}^{K-1}\sum_{j=k+1}^{K} \rho_{kj}} \Rightarrow \min$$

$$J_2 = \frac{1}{K}\left(\sum_{k=1}^{K} \rho_k - \frac{2}{K-1}\sum_{k=1}^{K-1}\sum_{j=k+1}^{K} \rho_{kj}\right) \Rightarrow \min$$

Here K is the number of groups,

$$\rho_k = \frac{2}{m_k(m_k-1)}\sum_{i=1}^{m_k-1}\sum_{s=i+1}^{m_k} \rho(\mathbf{w}^i, \mathbf{w}^s),$$

$$\rho_{kj} = \frac{1}{m_k m_j}\sum_{i=1}^{m_k}\sum_{s=1}^{m_j} \rho(\mathbf{w}^i, \mathbf{v}^s),$$

m_k, m_j are the numbers of objects in the group numbered k and in the group numbered j respectively; $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^{m_k}$ are the objects of the group numbered k; $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^{m_j}$ are the objects of the group numbered j.

After the groups are determined the next problem can be formulated: to find common feature of objects, which belong to the same group.

1.4 Classification with Learning

If it is a priori known about some objects to what groups (classes) they belong, then this information can be used to determine classification for other objects.

As a rule the set W is divided into two classes, say D and N.

The a priori examples of objects of each class are given. They are called the training set W_0:

$W_0 \subset W$,

$W_0 = D_0 \cup N_0$.

Here D_0 is the training set (the a priori examples) of objects belonging to class D, N_0 is the training set of objects belonging to class N.

The training set W_0 is used to determine a priori unknown distribution of objects of the set W_0 between the classes D and N.

The result of the pattern recognition is twofold:

- the rule of recognition; it allows to recognize which class an object belongs to knowing the vector $\mathbf{w}^i$ describing this object;
- the actual division of objects into separate classes according to this rule (Fig. 2):

$W = D \cup N$

or if there are objects with undefined classification then

$W = (D \cup N) \cup U.$

Analysis of the obtained rule of recognition may give information for understanding the connection between the feature, which differs the classes D and N, on one hand and description of objects (components of vectors $\mathbf{w}^i$) on another.

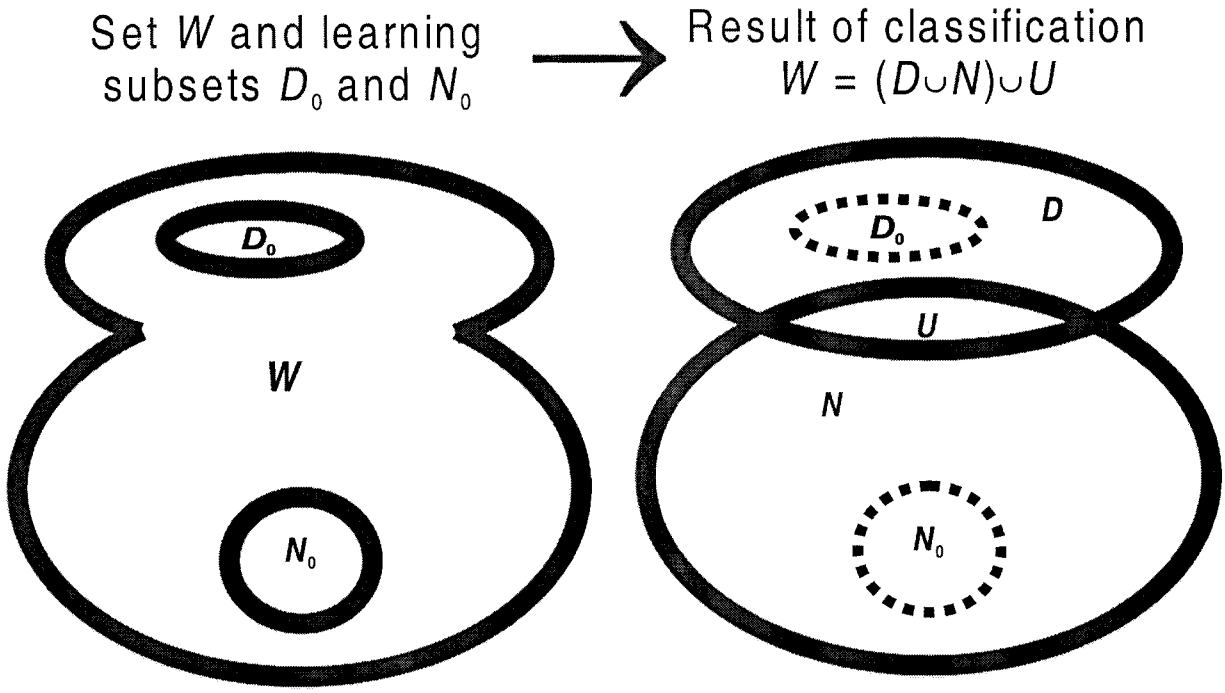

FIGURE 2 Classification with learning

II. EXAMPLES OF ALGORITHMS

Some algorithms used to solve problems of classification with learning are described below.

2.1 Statistical Algorithms

These algorithms are based on the assumption that distribution laws are different for vectors from classes D and N (see Fig. 3). The samples D_0 and N_0 are used to define the parameters of these laws.

The recognition rule includes calculating for each object $\mathbf{w}^i$ an estimation of conditional probabilities $P_D{}^i$ and $P_N{}^i$ that the object belongs to class D and N respectively. Classification of the objects according to these probabilities is performed as follows:

$$\mathbf{w}^i \in D, \text{ if } P_D{}^i - P_N{}^i \geq \varepsilon,$$
$$\mathbf{w}^i \in N, \text{ if } P_D{}^i - P_N{}^i < -\varepsilon,$$
$$\mathbf{w}^i \in U, \text{ if } -\varepsilon \leq P_D{}^i - P_N{}^i < \varepsilon,$$

where $\varepsilon \geq 0$ is a given constant.

Bayes algorithm. This is an example of a statistical algorithm. According to Bayes formula

$$P(\mathbf{w} = \mathbf{w}^i | \mathbf{w} \in D)\, P(\mathbf{w} \in D) = P(\mathbf{w} \in D | \mathbf{w} = \mathbf{w}^i)\, P(\mathbf{w} = \mathbf{w}^i) \qquad (1)$$

It follows from (1) that

$$P_D^i = P\left(\mathbf{w} \in D \middle| \mathbf{w} = \mathbf{w}^i\right) = \frac{P\left(\mathbf{w} = \mathbf{w}^i | \mathbf{w} \in D\right) P(\mathbf{w} \in D)}{P\left(\mathbf{w} = \mathbf{w}^i\right)}.$$

Similarly

$$P_N^i = P\left(\mathbf{w} \in N \middle| \mathbf{w} = \mathbf{w}^i\right) = \frac{P\left(\mathbf{w} = \mathbf{w}^i | \mathbf{w} \in N\right) P(\mathbf{w} \in N)}{P\left(\mathbf{w} = \mathbf{w}^i\right)}.$$

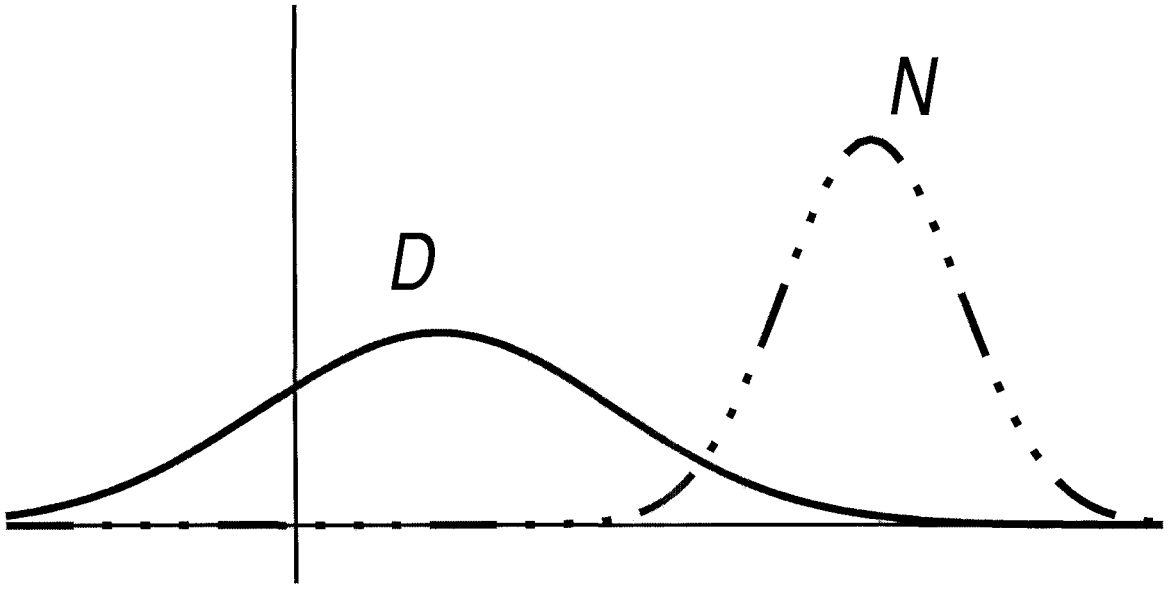

FIGURE 3 Different distribution laws for classes D and N

Estimations of probabilities in the right side of these relations are given by following approximate formulae, in which the samples D_0 and N_0 are used:

$$P(\mathbf{w} = \mathbf{w}^i | \mathbf{w} \in D) \approx P(\mathbf{w} = \mathbf{w}^i | \mathbf{w} \in D_0),$$

$$P(\mathbf{w} = \mathbf{w}^i | \mathbf{w} \in N) \approx P(\mathbf{w} = \mathbf{w}^i | \mathbf{w} \in N_0),$$

$$P(\mathbf{w} = \mathbf{w}^i) \approx P(\mathbf{w} = \mathbf{w}^i | \mathbf{w} \in D_0)\, P(\mathbf{w} \in D) + P(\mathbf{w} = \mathbf{w}^i | \mathbf{w} \in N_0)\, P(\mathbf{w} \in N).$$

Probability $P(\mathbf{w} \in D)$ is a parameter of the algorithm and has to be given, $P(\mathbf{w} \in N) = 1 - P(\mathbf{w} \in D)$.

2.2 Geometrical Algorithms

In these algorithms surfaces in the space w_1, w_2,..., w_m are constructed to separate classes D and N (see Fig. 4).

Algorithm Hyperplane. This is an example of a geometrical algorithm.

The hyperplane $P(\mathbf{w}) = a_0 + a_1 w_1 + a_2 w_2 + \ldots + a_m w_m = 0$ is constructed in the space w_1, w_2,..., w_m to separate the sets D_0 and N_0 by the best way. It means that some function on the hyperplnane has to have extremum value.

The example of the function is

$$J(a_0, a_1, \ldots, a_m) = \sum_{i=1}^{n_1} P(\mathbf{w}^i) - \sum_{i=1}^{n_2} P(\mathbf{v}^i) \Rightarrow \max.$$

Here $\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^{n_1}$ are objects of D_0, $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^{n_2}$ are objects of N_0.

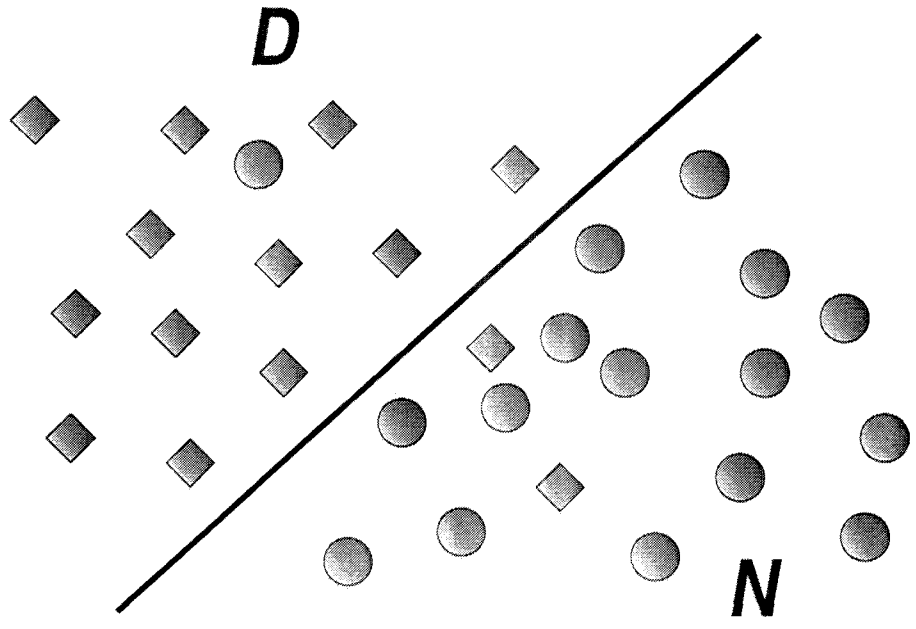

FIGURE 4 Separation of objects from classes D (rhombs) and N (circles) in two-dimensional space by a straight line.

The recognition rule is formulated as follows:

$$\mathbf{w}^i \in D, \text{ if } P(\mathbf{w}^i) \geq \varepsilon,$$
$$\mathbf{w}^i \in N, \text{ if } P(\mathbf{w}^i) < -\varepsilon,$$
$$\mathbf{w}^i \in U, \text{ if } -\varepsilon \leq P(\mathbf{w}^i) < \varepsilon,$$

where $\varepsilon \geq 0$ is a given constant.

2.3 Logical Algorithms

In these algorithms characteristic traits of classes D and N are searched using the sets D_0 and N_0. Traits are Boolean functions on w_1, w_2,..., w_m. The object $\mathbf{w}^i$ has the trait, if the value of the corresponding function, calculated for it, is ***true***, and does not have the trait, if it is ***false***. A trait is a characteristic trait of the class D, if the objects of the set D_0 have this trait more often than the objects of the set N_0. A trait is a characteristic trait of the class N, if the objects of the set N_0 have this trait more often than objects of the set D_0.

Using the searched characteristic traits the recognition rule is formulated as follows:

$$\mathbf{w}^i \in D, \text{ if } n_D{}^i - n_N{}^i \geq \Delta + \varepsilon,$$
$$\mathbf{w}^i \in N, \text{ if } n_D{}^i - n_N{}^i < \Delta - \varepsilon,$$
$$\mathbf{w}^i \in U, \text{ if } \Delta - \varepsilon \leq n_D{}^i - n_N{}^i < \Delta + \varepsilon.$$

Here $n_D{}^i$ and $n_N{}^i$ are the numbers of characteristic traits of classes D and N, which the object $\mathbf{w}^i$ has, Δ and $\varepsilon \geq 0$ are given constants.

Logical algorithms are useful to apply in cases then the numbers of objects in sets D_0 and N_0 are small.

As a rule logical algorithms are applied to vectors with binary components. An example of logical algorithm is the algorithm CORA-3. It is applied to geophysical problems, in particular to the problems of recognition of earthquake-prone areas and intermediate-term prediction of earthquakes. The detailed description of this algorithm can be found in *Gelfand et al.* (1976) and is given below.

III. PRELIMINARY DATA PROCESSING

As it was mentioned above some pattern recognition algorithms (e.g., CORA-3) do classify the vectors with binary components. Therefore, if the set W initially consists of vectors with real components (functions) then prior to an algorithm application, the coding of objects in the form of vectors with binary components has to be carried out. For this purpose, the characteristics are discretized, i.e. ranges of their values are represented as the union of disjoint parts. Then each of these parts is given accordingly by the value of a component of a binary vector or by the combination of values of its several components.

After discretization the data become robust. For example, if a range of some function is divided into three parts then only three gradations for this function ("small", "medium", "large") are used after the discretization instead of its exact value. Do not regret the loss of information. This makes results of recognition stable to variations of data.

3.1 Discretization

Let us consider some component (function) w_j of vectors (objects), which form the set W. Let the range of values of the function is limited with the numbers x_0^j and x_f^j ($x_0^j < x_f^j$). The procedure of discretization for the function w_j consists of dividing the range into k_j intervals by thresholds of discretization (Fig. 5):

$$x_1^j, x_2^j, \ldots, x_{k_j-1}^j \quad \left(x_0^j < x_1^j < x_2^j < \ldots < x_{k_j-1}^j < x_f^j\right)$$

Assume that the value w_j^i of the function numbered j of the object numbered i belongs to the interval numbered s, if $x_{s-1}^j < w_j^i \le x_s^j$, where $x_{k_j+1}^j = x_f^j$. After discretization we replace the exact value of the function by the interval, which contains this value.

Usually we divide the range of function values into two intervals ("small" and "large" values) or into three intervals ("small", "medium" and "large" values).

Thresholds of discretization can be introduce manually on the basis of various considerations for the nature of the given function.

The other way to determine the thresholds is to compute them so as to make the numbers of objects with the function values within each interval (x_{s-1}^j, x_s^j), $s = 1, 2, \ldots, k_j$, being roughly equal to each other. In this case one should specify the number of intervals k_j only. Then the thresholds of discretization may be calculated by using a special algorithm. All objects together or only objects of D_0 and N_0 may be considered. This type of discretization is called here and below as ***objective*** or ***automatic***.

Our purpose is to find such intervals where values of the function w_j for objects from one class occur more often than for objects from another class.

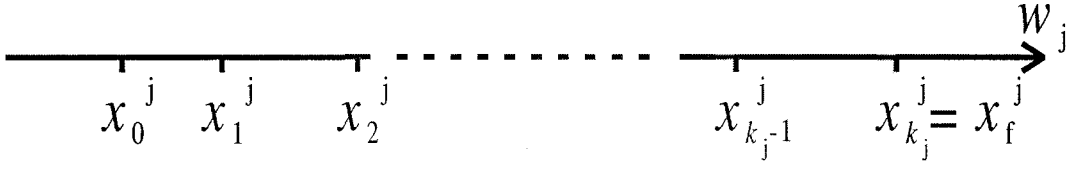

FIGURE 5 Discretization of the function w_j.

How informative is the function w_j in a given discretization can be characterized as follows.

Let us compute for each interval (x_{s-1}^j, x_s^j) the numbers P_s^D and P_s^N ($s = 1, 2, ..., k_j$), which give for the sets D_0 and N_0 respectively the percent of objects, for which the value of the function w_j falls within the interval numbered s.

Let us denote $P_{\max} = \max\limits_{1 \le s \le k_j} \left| P_s^D - P_s^N \right|$.

In other words P_s^D and P_s^N are empirical histograms of the function w_j for the sets D_0 and N_0, and $P_{\max}$ is the maximal difference of these histograms.

The larger is $P_{\max}$, the more informative is the function w_j.

Functions for which $P_{\max} < 20\%$ are usually excluded.

Another criterion of the quality of a discretization is ***monotonous dependence*** of P_s^D and P_s^N on s. Let $k_j = 3$. Let us denote:

$$M_D = \frac{\left|P_2^D - P_1^D\right| + \left|P_3^D - P_2^D\right|}{\left|P_3^D - P_1^D\right|},$$

$$M_N = \frac{\left|P_2^N - P_1^N\right| + \left|P_3^N - P_2^N\right|}{\left|P^N{}_3 - P_1^N\right|}.$$

If P_s^D changes monotonously with s, $M_D = 1$; the larger is M_D, more jerky is P_s^D. This is clear from Fig. 6. Similar statements are true for M_N, P_s^N.

The smaller are M_D and M_N, the better is the discretization of the function w_j. Functions with both M_D, $M_N \ge 3$ are usually excluded.

Samples D_0 and N_0 are often marginally small, so that their observed difference may be random. Therefore the relation between functions P_s^D and P_s^N after discretization should be not absurd according to the problem under consideration, though they may be unexpected indeed.

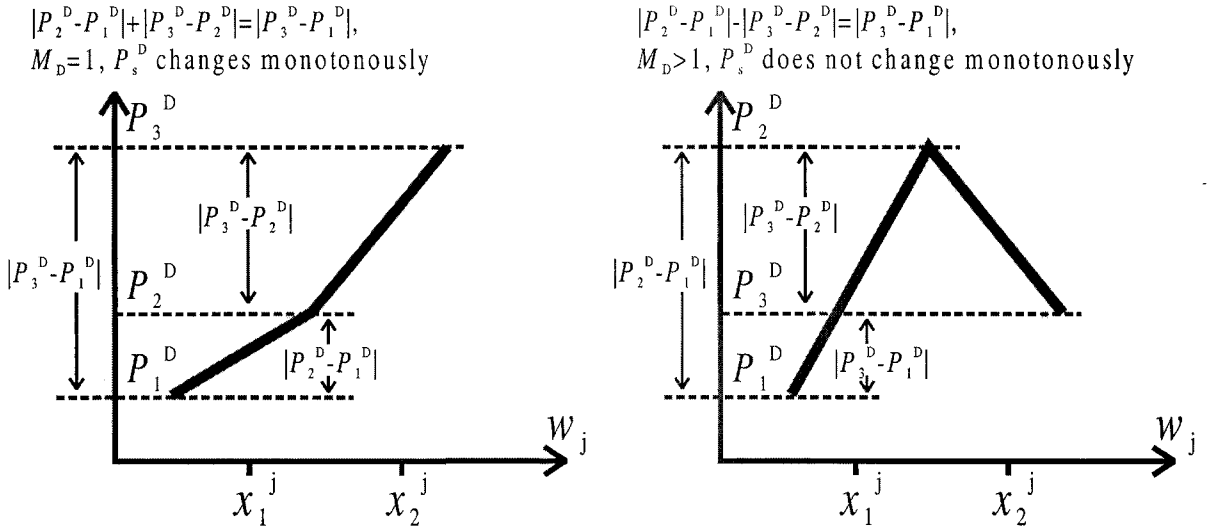

FIGURE 6 Monotonous and non-monotonous changing of P_s^D

3.2 Coding

With discretization thresholds determined, vectors $\mathbf{w}^i$ are coded into binary vectors. Only the functions selected at the stage of discretization are considered for coding. At the stage of coding l_j components of binary vectors are determined for the function w_j. Number l_j depends on the number of thresholds as well as on the type of coding procedure applied to the function w_j.

The following two types of coding are used.

1. ***I*** ("impulse") type. In this case $l_j = k_j$, i.e. the number of binary vector components allocated for the coding of the function w_j is equal to the number of intervals into which the range of its values is divided after discretization.

Let us denote as $\omega_1, \omega_2, ..., \omega_{lj}$ the values of binary vector components, which code the function w_j. If the value w_j^i of the function w_j for the object numbered i falls within the s-th interval of its discretization, i.e. $x_{s-1}^j < w_j^i \le x_s^j$, then we set

$\omega_1 = \omega_2 = ... = \omega_{s-1} = 0, \omega_s = 1, \omega_{s+1} = 0 = ... = \omega_{lj} = 0.$

2. ***S*** ("stair") type. In this case $l_j = k_j - 1$, i.e. the number of binary vector components, allocated for the coding of a function, is equal to the number of the thresholds of discretization. If the value w_j^i for the object numbered i falls within the s-th interval of its discretization, then we set

$\omega_1 = \omega_2 = ... = \omega_{s-1} = 0, \omega_s = \omega_{s+1} = ... = \omega_{lj} = 1.$

The case when the codes of the function w_j are constructed for $k_j = 3$ is considered below.

If the value w_j^i belongs to the first interval ($x_0^j < w_j^i \le x_1^j$) ***I*** type coding has the form: 100. ***S*** type coding for the same value w_j^i has the form: 11.

For the second interval ($x_1^j < w_j^i \le x_2^j$) the codes are 010 (***I*** type) and 01 (***S*** type).

For the third interval ($x_2^j < w_j^i \le x_3^j$) they are 001 and 00 respectively.

Discretization and coding procedures transform the set of vectors $W = \{ \mathbf{w}^i \}$, $i = 1, 2, ..., n$, which correspond to all objects, into a set of vectors with l binary components. Here $l = \sum' l_j$, where summation is implemented only over the functions left after discretization.

Thus, discretization and coding transform the initial problem in the form of the classification within the finite set of l-dimensional vectors with binary components. These vectors will be also called objects of recognition.

IV. ALGORITHM CORA-3

Algorithm CORA-3 operates in two stages:

- selection of characteristic traits (*learning*);
- *voting*.

4.1 Learning

In the learning stage, the algorithm determines characteristic traits for classes D and N using vectors from sets D_0 and N_0.

Traits. A matrix **A**,

$$\mathbf{A} = \begin{Vmatrix} i_1 & i_2 & i_3 \\ \delta_1 & \delta_2 & \delta_3 \end{Vmatrix},$$

denotes a trait, where i_1, i_2, i_3, $1 \le i_1 \le i_2 \le i_3 \le l$ are the numbers of binary vector components and δ_1, δ_2, δ_3 are their binary values. We say that a binary vector (an object) $\omega^i = (\omega_1{}^i, \omega_2{}^i, ..., \omega_l{}^i)$ has the trait **A** if $\omega^i_{i_1} = \delta_1$, $\omega^i_{i_2} = \delta_2$, $\omega^i_{i_3} = \delta_3$.

Characteristic traits. Let $W' \subseteq W$. Denote the number of vectors $\omega^i \in W'$ that have trait **A** by $K(W', \mathbf{A})$.

The algorithm has four free parameters $k_1, \overline{k}_1, k_2, \overline{k}_2$, which are nonnegative integers used to define characteristic traits of the two classes.

Trait **A** is a characteristic trait of class D if

$K(D_0, \mathbf{A}) \ge k_1$ and $K(N_0, \mathbf{A}) \le \overline{k}_1$.

Trait A is a characteristic trait of class N if

$K(N_0, \mathbf{A}) \ge k_2$ and $K(D_0, \mathbf{A}) \le \overline{k}_2$.

Parameters k_1 and k_2 are called selection thresholds for characteristic traits of classes D and N respectively. Parameters $\overline{k}_1$ and $\overline{k}_2$ are called contradiction thresholds for characteristic traits of classes D and N.

Equivalent, weaker, and stronger traits. The number of characteristic traits may be rather large. Some of them occur on the same vectors from training sets. The algorithm distinguishes such cases and does not include all characteristic traits in the final list.

Specifically, denote by $\Omega(\mathbf{A})$ a subset of set W such that $\omega^i \in \Omega(\mathbf{A})$ has trait **A**. Let, $\mathbf{A}_1$ and $\mathbf{A}_2$ be two characteristic traits of class D. Trait $\mathbf{A}_1$ is *weaker* than trait $\mathbf{A}_2$ (or $\mathbf{A}_2$ is *stronger* than $\mathbf{A}_1$), if

$\Omega(\mathbf{A}_1) \cap D_0 \subset \Omega(\mathbf{A}_2) \cap D_0$ and $(\Omega(\mathbf{A}_2) \cap D_0) \backslash (\Omega(\mathbf{A}_1) \cap D_0) \ne \varnothing$.

This condition means that all vectors from D_0 that have $\mathbf{A}_1$ also possess $\mathbf{A}_2$; at the same time there is at least one vector from D_0, which has trait $\mathbf{A}_2$, and does not have $\mathbf{A}_1$.

A similar definition is valid for characteristic traits of class N. Let $\mathbf{A}_1$ and $\mathbf{A}_2$ be two characteristic traits of class N. Then the $\mathbf{A}_1$ is weaker than trait $\mathbf{A}_2$ (or $\mathbf{A}_2$ is stronger than $\mathbf{A}_1$) if

$\Omega(\mathbf{A}_1) \cap N_0 \subset \Omega(\mathbf{A}_2) \cap N_0$ and $(\Omega(\mathbf{A}_2) \cap N_0) \backslash (\Omega(\mathbf{A}_1) \cap N_0) \ne \varnothing$.

Two characteristic traits $\mathbf{A}_1$ and $\mathbf{A}_2$ of class D are called *equivalent* if they are found on the same vectors of set D_0, i.e.,

$\Omega(\mathbf{A}_1) \cap D_0 = \Omega(\mathbf{A}_2) \cap D_0$.

Similarly, characteristics traits $\mathbf{A}_1$ and $\mathbf{A}_2$ of class N are called equivalent if

$\Omega(\mathbf{A}_1) \cap N_0 = \Omega(\mathbf{A}_2) \cap N_0$.

The algorithm excludes from the lists of characteristic traits those that are weaker or equivalent to a selected trait.

Thus, the learning stage results in the final list of q_D and q_N characteristic traits of classes D and N. respectively. Any member of this list does not have weaker or equivalent members.

4.2 Voting and Classification

In the second stage the algorithm performs voting and classification using the final list of characteristic traits. For each vector $\omega^i \in W$, it calculates the number n_D^i of characteristic traits of class D, which the vector possesses, the number n_N^i of those of class N, and the difference $\Delta_i = n_D^i - n_N^i$ called voting.

The classification is defined as follows.

Class D (set D) is formed from the vectors ω^i, for which $\Delta_i \geq \Delta$. The vectors, for which $\Delta_i < \Delta$, are included in class N (set N).

Here Δ is a parameter of the algorithm as well as k_1, $\bar{k}_1$, k_2, and $\bar{k}_2$.

This recognition rule corresponds to $\varepsilon = 0$ in the description of logical algorithms given above.

4.3 Algorithm CLUSTERS

Algorithm CLUSTERS is the modification of algorithm CORA-3 (*Gelfand et al.*, 1976). It is applied in the case when set D_0 consists of S subsets (subclasses):

$D_0 = D_0^1 \cup D_0^2 \cup \ldots \cup D_0^S$,

and it is known a priori that at least one element of each subclass belongs to class D but some elements of set D_0 may belong to class N.

The learning stage of algorithm CLUSTERS differs from that of CORA-3 in the following.

First, by definition, a subclass has a trait if it contains at least one vector with this trait. Trait $\mathbf{A}$ is a characteristic trait of class D, if

$K^S(D_0, \mathbf{A}) \geq k_1$ and $K(N_0, \mathbf{A}) \leq \bar{k}_1$.

Here $K^S(D_0, \mathbf{A})$ is the number of subclasses that have the trait $\mathbf{A}$.

Second, the definition of the weaker and equivalent traits for characteristic traits of class D is different. A characteristic trait $\mathbf{A}_1$ of class D is weaker than a characteristic trait $\mathbf{A}_2$ of the same class if any subclass that has trait $\mathbf{A}_1$ also has $\mathbf{A}_2$ and there is at least one subclass, which has trait $\mathbf{A}_2$ but does not have trait $\mathbf{A}_1$. Traits $\mathbf{A}_1$ and $\mathbf{A}_2$ are equivalent if they are found in the same subclasses.

Algorithm CLUSTERS forms the sets of characteristic traits of classes D and N like CORA-3.

The stage of voting and classification is the same as in algorithm CORA-3.

V. ALGORITHM HAMMING

Another algorithm applied to geophysical problems is algorithm HAMMING (*Gvishiani and Kossobokov*, 1981). There are also other possible applications of this algorithm (e.g., *Keilis-Borok and Lichtman*, 1981).

This algorithm operates also in two stages: learning and voting.

5.1 Learning

In the first stage (learning), the algorithm computes for each component ω_k (k = 1, 2, ..., l) of binary vectors the following values:

$q_D(k|0)$ - the number of objects of the set D_0, which have $\omega_k = 0$,

$q_D(k|1)$ - the number of objects of the set D_0, which have $\omega_k = 1$,

$q_N(k|0)$ - the number of objects of the set N_0, which have $\omega_k = 0$,

$q_N(k|1)$ - the number of objects of the set N_0, which have $\omega_k = 1$.

Then the relative number of vectors, for which this component equals to 1, is determined for the set D_0:

$$\alpha_D(k|1) = \frac{q_D(k|1)}{q_D(k|0) + q_D(k|1)}$$

and for the set N_0:

$$\alpha_N(k|1) = \frac{q_N(k|1)}{q_N(k|0) + q_N(k|1)} .$$

A binary vector $\mathbf{K} = (\kappa_1, \kappa_2, ..., \kappa_l)$ called the ***kernel of class D***, is determined as follows:

$$\kappa_k = \begin{cases} 1, \text{ if } \alpha_D(k|1) \geq \alpha_N(k|1), \\ 0, \text{ if } \alpha_D(k|1) < \alpha_N(k|1). \end{cases}$$

The calculation of the kernel **K**, whose components are more typical of set D_0 than of N_0 completes the first stage.

NOTE: It may be more reliable to eliminate the components, for which $|\alpha_D(k|1) - \alpha_N(k|1)| < \varepsilon$, where ε is a small positive constant.

5.2 Voting and Classification

In the second stage, the algorithm computes Hamming's distances

$$\rho_i = \sum_{k=1}^{l} \left| \omega_k^i - \kappa_k \right|$$

from each vector $\boldsymbol{\omega}^i \in W$ to the kernel of class D.

The classification is defined as follows.

Class D (set D) is formed from vectors ω^i, for which $\rho_i \le R$.

The vectors, for which $\rho_i > R$, are included in class N (set N).

Here R is a parameter of the algorithm.

Algorithm HAMMING-1 is generalization of HAMMING. It operates with the generalized Hamming's distance

$$\rho_i = \sum_{k=1}^{l} |\omega_k^i - \kappa_k| \xi_k .$$

Weights $\xi_k > 0$ ($k = 1, 2, ..., l$) are parameters of the algorithm. They can be assigned arbitrarily or computed from objective considerations that reduce the danger of self-deception; for example, by formula:

$$\xi_k = \frac{|\alpha_D(k|1) - \alpha_N(k|1)|}{\max_k |\alpha_D(k|1) - \alpha_N(k|1)|}$$

where maximum is taken over all components used in the given run of the algorithm.

VI. EVALUATION OF THE CLASSIFICATION RELIABILITY

Reliability of results of recognition is evaluated by several methods including control tests, statistical analysis of the established classification and other techniques. These tests are necessary to be sure in the obtained results. It is especially important in the case of small samples D_0 and N_0. The tests illustrate - how reliable are the results of the pattern recognition. However they do not provide a proof in the strict statistical sense if the training material is small.

The following simplest tests are useful.

1. To save the part of objects from W_0 for recognition only, not using it in learning.
2. To check the conditions: $D_0 \subset D$, $N_0 \subset N$.

NOTE: Sometimes these conditions are not valid because sets D_0 and N_0 are not "clear" enough. For example, in the case of recognition of earthquake-prone areas objects of D_0 are structures where epicenters of earthquakes with $M \geq M_0$ are known and objects of N_0 are structures where epicenters of such earthquakes are not known. Objects of N_0 may belong to the class D, because in some areas earthquakes with $M \geq M_0$ may be possible, though yet unknown. Objects of D_0 may belong to the class N due to the errors in the catalog (in epicenters and/or magnitude).

The examples of some other tests are listed below. These tests include some variation of the objects, used components of vectors, numerical parameters etc. The test is positive if the results of recognition are stable to these variations. Since the danger of self-deception is not completely eliminated by these tests the design and implementation of new tests should be pursued.

6.1 Using a Result of Classification as a Training Set (RTS test)

This test is an attempt to repeat the established classification $W = D \cup N$, using the resultant sets D and N as the new training sets instead of D_0 and N_0. We usually consider this test as successful if not more than 5% of the total number of objects are classified in the test differently comparing with their initial classification. The "physical" idea of the test is rather obvious and natural: if our classification is correct then such changing of training material should not change the result of classification.

Note that algorithm CORA-3 allows easy repetition of initial classification if one takes $\bar{k}_1 = \bar{k}_2 = 0$ and sufficiently small k_1 and k_2. Therefore, it is advisable to perform this test with nonzero thresholds $\bar{k}_1$ and $\bar{k}_2$. For example, $\bar{k}_1 = \bar{k}_2 = 1$, or $\bar{k}_1 = \bar{k}_2 = 2$, or the same as in the initial classification. In the case of $\bar{k}_1 = \bar{k}_2 = 0$ the substantial information is carried with maximum values of k_1 and k_2, under which the initial classification can be repeated.

In the case of any algorithm used to obtain the initial classification, it's advisable to repeat it in making the test by using HAMMING algorithm. We consider success of RTS test as the necessary condition for the classification obtained to pretend to be the problem solution. In this sense RTS test is obligatory to check the reliability of the classification.

6.2 Stability Testing (ST tests)

These tests generalize RTS test. Their goal is to obtain the initial classification $W = D \cup N$, using the various subsets $D_0' \subseteq D$, $N_0' \subseteq N$ as D_0 and N_0 training sets. The test is considered successful if the initial classification is rather stable while we change training material. Usually we accept the result if not more than 10% of the total number of objects change their classification in the result of the test. The choice of D_0' and N_0' used as training sets in ST test

can be different. For instance, in the case of recognition of earthquake-prone areas the region at hand can be divided into two parts, and subsets D_0' and N_0' then formed from objects of the sets D and N objects with preimages belong to one part. The other way of selecting D_0' and N_0' can be based on voting results in the initial classification. If algorithm HAMMING (or HAMMING-1) is used, the objects $\mathbf{w}^i \in D$ close to the kernel $\mathbf{K}$ can be assigned to D_0', and those far from it are assigned to N_0'. When algorithm CORA-3 (or CLUSTERS) is used, the objects $\mathbf{w}^i \in D$ with larger values of Δ_i can be assigned to D_0', whereas those with small Δ_i form N_0'.

Successful results of different ST tests are appealing indirect arguments favoring the validity of an established classification. At the same time, a success in a single test with an arbitrary choice of D_0' and N_0' is by no means a proof of reliability.

6.3 Sliding Control (SC test)

This test is designed for establishing classifications on the basis of the training sets $(D_0 \backslash \mathbf{w}^i)$ and $(N_0 \backslash \mathbf{w}^{i+n_1})$, $i = 1, 2, ..., \max(n_1,n_2)$. The idea of SC test is very clear. We just want to check weather classification of the objects belonging to the training set is stable while they are excluded from the training set. The first variant discards the objects $\mathbf{w}^1 \in D_0$ and $\mathbf{w}^{1+n_1} \in N_0$, the second variant resets them but discards the objects $\mathbf{w}^2 \in D_0$ and $\mathbf{w}^{2+n_1} \in N_0$, etc. If one of sets D_0 or N_0 (with a smaller number of objects) has already all its objects discarded once, we proceed only with the other set. In case of algorithm CLUSTERS the whole subclasses are excluded in turn from the set D_0.

Formal criteria of success of the test is small value of ratio $\frac{m_D}{|D_0|}$ or $\frac{m_D + m_N}{|D_0| + |N_0|}$. Here m_D and m_N show how many objects of D_0 and N_0 respectively change classification after they were excluded from learning. We usually consider SC test as successful if not above 20% of objects in each of D_0 and N_0 sets change their classification while neglecting.

This test is very similar to the well-known "jackknife" procedure, under which each variant discards only one object, first from D_0, and then from N_0. On the other hand SC is preferable because it needs executing less variants of classification.

6.4 Voting by Equivalent Traits (VET test)

This test is applied only if classification is obtained by CORA-3 (or CLUSTERS) algorithm. In both cases the result of classification depends on the choice of traits picked up from equivalence groups. The VET test aims at evaluating the classification stability under such a choice.

Let object $\mathbf{w}^i$ possesses u^i_{Dj} traits, which are equivalent to j-th trait of class D, and u^i_{Nj} traits, which are equivalent to j-th trait of class N. We define on the bases of numbers u^i_{Dj} and u^i_{Nj} the numbers of "votes" in favor of classes D and N respectively as follows.

$$u^i_D = \sum_{j=1}^{p_D} \frac{u^i_{Dj}}{p^j_D}, \quad u^i_N = \sum_{j=1}^{p_N} \frac{u^i_{Nj}}{p^j_N}.$$

Here p^j_D is the total number of traits equivalent to j-th trait of class D, p^j_N - the number of traits equivalent to j-th trait of class N. In calculation of both numbers p^j_D and p^j_N j-th trait itself is obviously included. In the test the set D is formed from the objects, which satisfy the condition $u^i_D - u^i_N \geq \Delta$ and the rest of objects forms the set N.

The results of the VET test are claimed successful if it is possible to find Δ such that the total change in classification is less than 5% of the total number of recognition objects. We consider a success of the VET test as a necessary precondition for claiming the validity of the resultant classification obtained with CORA-3 or CLUSTERS.

6.5 Randomization of Data

These tests (*Gvishiani and Kossobokov*, 1981) are used to estimate the probability of an erroneous classification and its nonrandomness in the absence of a control sample.

The sequence of intermixed problems is considered in these tests. An intermixed problem is formulated on the basis of initial one by a random choice of n_1 objects from given n objects of the set W and also by a random choice n_2 objects from the rest of n - n_1 objects of the set W. These two new random training sets we symbolize as D_0' and N_0'. Coding of the objects in the form of binary vectors remains the same for an intermixed problem as it is in the real one. In other words it means that we preserve the relationship between the characteristics, which organic one to the set W as a whole. The total number $C_n^{n_1} C_{n-n_1}^{n_2} = n!/[n_1! n_2! (n - n_1 - n_2)!]$ of intermixed problems may be defined.

A pattern recognition algorithm is applied to each intermixed problem, and the classification $W = D \cup N$ based upon the training sets D_0' and N_0' is obtained in the given intermixed problem. The condition that $|D|$ is not greater than the number of objects in the set D obtained in the initial classification is imposed on the classification in the intermixed problem.

Assume that F of intermixed problems have been formed and f_1 among them succeeded to include $D_0' \subseteq D$. Then f_1/F ratio may be used as the measure of the result to be non-random. If the values of f_1/F are small it obviously means that, it is complicated to obtain a random result of the same quality as the real one. In this sense the small values of f_1/F speak for the fact that the real result obtained is non-random. On the other hand it cannot of course be used as a necessary condition to proceed with the classification.

Gvishiani and Kossobokov (1981) showed that under some natural additional requirements classifications in intermixed problems offer to define the upper estimate of classification error probability for the original problem. This upper estimate is calculated by the formula

$$\overline{p} = |\overline{\mathbf{N}}|/n - \overline{\nu_D}/n_1 .$$

Here $|\overline{N}|$ is the average number of objects allocated to class N in the intermixed problems, $\overline{\nu_D}$ - the average number of objects from sets D_0' allocated to N in the intermixed problems.

Naturally, a small value of $\overline{p}$ is the argument favoring the validity of classification obtained for the original problem. If the estimation results in a large value ($\overline{p} > 0.5$), it is advisable to return to the original problem. Such a situation may indicate, for instance, an insufficient size of D_0. On the other hand, one should remember that $\overline{p}$ gives only the upper estimation of the error probability, though its value is usually much less.

6.6 Result Replication Tests

These tests are the attempts to replicate the obtained result by altering the solution procedure starting with some intermediate stage. The application of another pattern recognition algorithm is used in the simplest example of such experiment. For example, classification was established by performing CORA-3 algorithm, then, using that same coding of objects, an attempt is made to repeat the classification by applying HAMMING algorithm. This test is

usually considered as satisfactory one if not more than 20% of objects change their classification.

When application of a simpler algorithm results in repeating almost entirely the initial classification, its validation rises, of course. On the other hand replication of the classification by another algorithm cannot be considered, of course, as the necessary condition for the result to be valid.

The set of used components of binary vectors may be changed. In particular this may include elimination of each used component in turn.

An attempt may be also made to repeat the classification altering discretization thresholds for the functions describing the objects. Corresponding changes in coding of the objects should be also made. New functions may be included in the description of the objects. Then by replication of all subsequent stages of the problem consideration, a new classification is established and its comparison with the initial is made.

VII. APPLICATION OF PATTERN RECOGNITION METHODS TO GEOPHYSICAL PROBLEMS

Application of the pattern recognition methods to the problems of earthquake-prone areas determination and intermediate-term earthquake prediction is considered below.

7.1 Recognition of Earthquake-prone Areas

The problem under consideration is to determine in the region the areas where strong (with magnitude $M \geq M_0$ where M_0 is a threshold specified) earthquakes are possible. The basic assumption is that strong earthquakes associate with morphostructural nodes, specific structures that are formed about intersections of fault zones. This gives possibility to apply the pattern recognition approach.

The nodes are considered as objects of recognition. They are identified by means of the morphostructural zoning and described by characteristics determined on the basis of the topographical, geological, geomorphological and geophysical data. When these characteristics are measured, the objects are represented by vectors with components, which are values of the characteristics.

The problem as the pattern recognition one is to divide the vectors into two classes: vectors D (Dangerous) and vectors N, which represent correspondingly the nodes where earthquakes with $M \geq M_0$ may occur and the nodes where only earthquakes with $M < M_0$ may occur. Application of the pattern recognition algorithms requires a training set of vectors, for which we know *a priori* the class they belong to. The training set is formed on the basis of the data on seismicity observed in the region. It consists of vectors D_0 and N_0 representing correspondingly the nodes where strong earthquakes occurred and the nodes, which are far from the known epicenters of such earthquakes.

Formulation of the Problem and the Main Stages of Its Investigation

Consider a selected magnitude cutoff M_0 that defines large earthquakes in the region under study. Roughly speaking, the problem of determining earthquake-prone areas aims at separating places of potential earthquakes into two parts, D where earthquakes with magnitude $M \geq M_0$ can happen and N where earthquakes with magnitude $M \geq M_0$ are impossible.

The first question arising in a strict formulation of the pattern recognition problem is how to select the region and the magnitude cutoff M_0. The experience accumulated in *Gelfand et al.* (1972, 1973, 1974a, 1974b, 1976), *Zhidkov et al.* (1975), *Gvishiani et al.* (1978, 1987), *Caputo et al.* (1980), *Zhidkov and Kossobokov* (1980), *Gvishiani and Kossobokov* (1981), *Kossobokov* (1983), *Gvishiani and Soloviev* (1984), *Cisternas et al.* (1985), and *Gorshkov et al.* (1987) suggests the following heuristic criteria.

- The number of large earthquakes with $M \geq M_0$ in the region should be at least 10-20.
- The circles centered at epicenters of reported earthquakes with $M \geq M_0$ that have radii about the size of their source should not cover all of the region (otherwise, the problem has a trivial solution where the whole region is D).
- The region has to be tectonically uniform in sense of the similarity of possible causes of earthquakes with $M \geq M_0$.

These criteria establish certain limitations on the size of the region and the threshold M_0. For instance, M_0 = 5.0-6.0 implies the linear size of a region of the order of hundreds kilometers, whereas for M_0 = 7.0-7.5 this size should be larger than a thousand kilometers. M_0 = 8.0 requires a region tens of thousands kilometers long. These limitations were met in

practice, for example, in Italy, $M_0 = 6.0$ (*Caputo et al.*, 1980); in California, $M_0 = 6.5$ (*Gelfand et al.*, 1976); in South America and Kamchatka, $M_0 = 7.75$ (*Gvishiani and Soloviev*, 1984), and in the whole Circumpacific, $M_0 = 8.0$ (*Gvishiani et al.*, 1978). The experience accumulated in a decade confirmed that pattern recognition methods might reliably distinguish earthquake-prone areas on different scales of lithospheric block hierarchy and in different seismic and tectonic environments (*Gelfand et al.*, 1972, 1973, 1974a, 1974b, 1976; *Zhidkov et al.*, 1975; *Gvishiani et al.*, 1978, 1987; *Caputo et al.*, 1980; *Zhidkov and Kossobokov*, 1980; *Gvishiani and Kossobokov*, 1981; *Kossobokov*, 1983; *Gvishiani and Soloviev*, 1984; *Cisternas et al.*, 1985; *Gorshkov et al.*, 1987).

When selecting the region and threshold magnitude M_0, it is necessary to define the objects of recognition.

Gelfand et al. (1972) were the first who applied pattern recognition methods to determine earthquake-prone areas in the Pamirs and Tien Shan. Since then, several important improvements in such a determination have been developed, including a broader choice of natural objects for recognition. In general, one may consider three types of objects in a study of earthquake-prone areas: planar areas, segments of linear structures, and points.

Gelfand et al. (1972) used planar morphostructural nodes of the Pamirs and Tien Shan as candidates for earthquake-prone areas. At that time, even a formal definition of this structure that permits reproducible identification did not exist and was subject of further analysis by gemorphologists and mathematicians (*Alekseevskaya et al.*, 1977). However, because most fractional areas are characterized by multidirectional intensive tectonic movements, nodes essentially attract epicenters of large earthquakes. The fact that most earthquakes with $M \geq M_0$ in a region originate within nodes is a necessary precondition for using them as objects of recognition. *Ranzman* (1979) formulated the geomorphological basis that favors this precondition. *Gvishiani and Soloviev* (1981) suggested a statistical method for testing it in practice, even when the boundaries of nodes are not defined precisely.

In planar nodes, pattern recognition algorithms classify morphostructural node in the region either as a D node, which is prone to earthquakes with $M \geq M_0$, or as a N node, where strong earthquakes are not possible. Such a classification determines the area D as the union of all D nodes and the area N as the union of all N nodes. The remaining territories of the region complementary to the nodes are not assumed to be dangerous (they are rejected with a certain level of confidence by preconditioning strong earthquake – node association).

This natural choice of objects entails a difficult problem outlining the boundaries of morphostructural nodes. When the difficulty is overwhelming, one may try substituting the nodes with intersections of morphostructural lineaments as done by *Gelfand et al.* (1974b). Tracing lineaments and their intersections is much easier task for a geomorphologist that essentially delivers similar (though less complete) information on the most fractured places of multidirectional intensive tectonic movements. That is why intersections of morphostructural lineaments were commonly used for determining of earthquake-prone areas (*Gelfand et al.*, 1974b, 1976; *Zhidkov et al.*, 1975; *Caputo et al.*, 1980; *Zhidkov and Kossobokov*, 1980; *Gvishiani and Soloviev*, 1984; *Cisternas et al.*, 1985; *Gorshkov et al.*, 1987; *Gvishiani et al.*, 1987). The necessary precondition of using nodes as recognition objects is transformed to a hypothesis that epicenters of strong earthquakes originate near intersections of morphostructural lineaments (*Gelfand et al.*, 1974b). This hypothesis is likely to be confirmed in a region if the following two conditions are valid: (1) the distance from all accurately determined epicenters of earthquakes with $M \geq M_0$ to the nearest intersection does not exceed a predefined distance r; (2) the area covered by circles of radius r centered in all intersections is a small part of the total area of the region. A statistical justification of the hypothesis can be obtained by using the algorithm developed by *Gvishiani and Soloviev* (1981).

Pattern recognition algorithms assign the vectors that describe intersections of lineaments to two classes: class D of intersections having vicinities prone to earthquakes with

$M \geq M_0$ (D intersections) and class N. The classification of vectors determines the preimage of area D as the union of all vicinities of D intersections. The area N is the complement of area D in the union all vicinities of intersections. It is assumed that the remaining territories of the region complementary to all vicinities of intersections are not dangerous.

Usually, earthquakes are associated with segments of faults that they rapture. Therefore linear objects of recognition, like segments of active faults or fault zones, may seem most natural to many seismologists (*Gelfand et al.* (1976) give an excellent demonstration of how the problem is viewed differently). Pattern recognition algorithms divide linked linear objects into two classes: D segments capable of originating earthquakes with $M \geq M_0$ and N segments that are not.

Segments of linear structures were used as objects for recognition of earthquake-prone areas in California (*Gelfand et al.*, 1976), where the basic linear structure was San-Andreas fault, in the whole linear structure of Circumpacific seismic belt (*Gvishiani et al.*, 1978), and in the Western Alps (*Cisternas et al.*, 1985), where the segments of linear structures, forming a neotectonic scheme of the region were considered.

The usage of pattern recognition algorithms with learning necessitates an a priori selection of the training set W_0, which is the union of two subjects that do not overlap: the training set D_0 from class D and the training set N_0 from class N. Such a selection of $W_0 = D_0 \cup N_0$ depends on the types of the objects for recognition. In the case of planar objects, all of those, including known epicenters of earthquake with $M \geq M_0$, form D_0, whereas the subset N_0 consists of all remaining objects from W, $N_0 = W \setminus D_0$, or those of such objects that do not contain known epicenters of earthquakes with $M \geq M_0 - \delta$ (where $\delta > 0$ is usually 0.5 or about this value). It is necessary to emphasize that N_0 is not "pure" training set in the sense that some of its members belong to class D. In the first case, where $N_0 = W \setminus D_0$, the problem consists of distinguishing samples that spoil the purity of N_0. Such a fussy type of learning highlights a specific difficulty in locating possible earthquake-prone areas by pattern recognition techniques.

It is natural to require the condition $D_0 \subseteq D$, where D denotes the vectors classified as belonging to class D. In other words, all places of strong earthquakes that are known should be recognized. When D_0 many vectors a part of it can be excluded from the training set and reserved to verify the reliability of the decision rule obtained.

When recognition objects are points, the training set D_0 is assembled from those that are situated at a distance not exceeded a certain fixed value r from the reported epicenters of earthquakes with $M \geq M_0$. The choice of r must satisfy the condition that the distance from most (practically all) of the well located epicenters of strong earthquakes in the region to the nearest recognition point is less than r. Naturally r scales with M_0. For instance, *Zhidkov and Kossobokov* (1980) used $r = 40$ km for $M_0 = 6.5$ in the eastern part of Central Asia; *Gvishiani and Soloviev* (1984) chose $r = 100$ km for $M_0 = 7.75$ on the Pacific coast of South America. The training set N_0 consists of either all remaining points or those of them that are at a distance r_1 ($r_1 \geq r$) or longer from the epicenters of earthquakes with $M \geq M_0 - \delta$ ($\delta > 0$). In this case the training set N_0 also can contain points that are potentially from class D.

There is a certain difficulty when recognition objects are points; one epicenter can be attributed to several objects if its distance to each of them is r or less. In such case the training set D_0 may have some objects from class N. Algorithm CLUSTERS, which takes into account this specific feature of the training set D_0 is used to overcome this difficulty. In case of ambiguity, the condition that $D_0 \subseteq D$ is changed by another natural one: each epicenter of an earthquake with $M \geq M_0$ has a point D object at a distance r or less.

When recognition objects are linear segments, the training set D_0 assembles those containing a projection of an epicenter of a strong earthquake. The training set N_0 is either $N_0 = W \setminus D_0$ or contains segments from W that are not neighbors of D_0. Another way to form N_0 is

to exclude those segments from $W \setminus D_0$ that contain a projection of an epicenter of an earthquake with $M \geq M_0 - \delta$ (where $\delta > 0$ is a parameter). As a rule, there is a unique projection of an epicenter that does not create ambiguity in selecting D_0: therefore, it is natural to require that $D_0 \subseteq D$.

Pattern recognition algorithms operate with vectors of characteristics representing natural recognition objects. As far as the earthquake-prone areas are considered, it appears natural to use the characteristics describing, either directly or indirectly, the intensity of recent tectonic activity at the locality of each object. The accumulated experience in recognizing earthquake-prone areas has established the following characteristics as tipical:

- a multitude of characteristics describing topography: maximum altitude above see level H_{max} inside the object area, altitude range ΔH; dominating combination of geomorphological structures in the object's vicinities, percentage of the object's area with existing Paleogene Quaternary sediments, etc.;
- characteristics describing the complexity of geomorphological and neotectonic network of structures: number of lineaments forming the object, the highest rank of lineament among those which form the object, etc.;
- characteristics describing gravitational field anomalies.

In case of planar objects the sense of "area" is obvious. When objects are points the area is a circle of the same radius for all objects centered at an object. When objects are linear segments the area is a circle of the same radius for all objects centered at the middle of a segment. Planar objects may have various areas and the area of an object may be used as one of characteristics.

In principle, all available information related directly or indirectly to the level of seismic activity can be used to characterize objects. The only necessary precondition for a characteristic is availability of uniform measurements across the entire region under consideration. After measuring selected characteristics for all the objects, they are converted to vectors $\mathbf{w}^i = \{w_1^i, w_2^i, ..., w_m^i,\}$, $i = 1, 2, ..., n$, where m is the total number of characteristics, n is the total number of objects in W, and w_k^i is the value of the k-th characteristic measured for the i-th object.

The pattern recognition algorithms, which are used to investigate the problem, work in a binary vectors space. Their application requires a transformation of vectors that describe natural recognition objects into binary ones. A specific transformation, so-called coding of characteristics is described above in Section 3.2.

Given the training set of vectors $W_0 = D_0 \cup N_0$, a pattern recognition algorithm determinates a classification $W = D \cup N$ where D and N are sets of vectors of classes D and N, respectively. As pointed above, the resulting classification should satisfy certain conditions, like $D_0 \subseteq D$ for planar objects. To avoid a trivial solution when all places considered belong to D, the following condition is usually introduced:

$$|D| \leq \beta |W|.$$

where $|D|$ and $|W|$ stand for the numbers of objects in sets D and W, respectively; and β, $0 < \beta < 1$, is a real constant, which sets an a priori upper bound for the fraction of D vectors in W. The value and justification of β must result from an expert evaluation of geological, seismological, and other available information on the region.

The quality and reliability of a classification can be verified by control tests. If successful, such test favors the classification that actually divides the region into earthquake-prone areas and areas where earthquakes with $M \geq M_0$ are not likely. Usually, pattern recognition of earthquake-prone areas involves a small sample of natural objects whose size does not allow reserving a control set for verification. Nevertheless, certain verification of the classification can be achieved by a the comprehensive analysis of the result and additional information that was not used initially, of which the most important are data on epicenters of

large earthquakes, e.g., noninstrumental, either historical or paleoseismological. Section VI describes some the typical control tests and other methods for evaluating the reliability of small sample statistics.

Classifications that are not satisfactory and have no meaningful interpretation are usually not reported. To get a satisfactory classification, a researcher can perform several cycles of trial and error through the following stages of recognition:

- definition of the region under study and the magnitude cutoff attributed to earthquake-prone areas;
- choice of the natural recognition objects;
- selection of the training set $W_0 = D_0 \cup N_0$;
- description of objects as vectors;
- discretization and coding of the characteristics;
- classification of vector space $W = D \cup N$ by a pattern recognition algorithm;
- evaluation of the reliability of classification from control tests;
- interpretation of the classification $W = D \cup N$ as a division of the region into earthquake-prone and other areas;
- generalization of geological and geomorphological interpretation of classification and the rules used to obtain it.

After the definition of D and N areas in the region territory it is advisable to do a statistical analysis of the locations of the known epicenters of earthquakes with $M < M_0$ relative to the located areas (as, e.g., in *Kossobokov and Soloviev*, 1983). The result of such comparison can lead, in principle, to the conclusion that the obtained D and N areas are actually earthquake-prone areas for earthquakes with $M < M'_0$ where M'_0 is a smaller magnitude threshold than M_0.

Recognition of earthquake-prone areas for the Western Alps

The problem of recognition of places in the Western Alps where earthquakes with $M \geq 5.0$ may occur (*Cisternas et al.*, 1985) is briefly considered below.

The intersections of the morphostructural lineaments obtained as the result of the morphostructural zoning of the Western Alps are natural objects of pattern recognition. The scheme of the morphostructural zoning of the Western Alps and the objects are shown in Fig. 7. The total number of objects in the set W is 62. The problem is to classify these objects into two classes: objects where earthquakes with $M \geq 5.0$ may occur (class D) and objects where earthquakes with $M \geq 5.0$ are impossible (class N).

Table 1 contains the list of characteristics, which describe the objects. The components of vectors $\mathbf{w}^i$ are the values of these characteristics.

The epicenters of earthquakes with $M \geq 5.0$ or $I \geq 7$ (I is maximum macroseismic intensity) are shown in Fig. 7 by dark circles with years of occurrence. The training set D_0 includes intersections located near epicenters of earthquakes with $M \geq 5.0$, 1900-1980. If an epicenter is at a distance less than 25 km from two intersections, both them are included in D_0. As a result, 14 intersections (3, 12, 13, 14, 20, 30, 31, 35, 40, 41, 42, 44, 51, and 57) constitute D_0. Intersections 1, 5, 6, 8, 53, 55, 56, 58, 60, and 61 hosting historic earthquake with $I \geq 7$ are not included both in D_0 and N_0 training sets as well as intersections 18 and 19. The latter are near the 1905 epicenter represented in D_0 by the nearest intersection 20. The remaining 36 intersections compose the training set N_0.

Table 1 lists the characteristics and the discretization thresholds used for recognition. Except for the combination of topographic forms, their binary coding was ***S*** type (see Section 3.2). The most informative characteristics are: maximum altitude H_{max}, altitude gradient $\Delta H/l$,

the percentage of Quaternary deposits Q, the highest rank of the lineament R_h, the distance to the nearest second rank lineament ρ_2. For all of them $P_{max} > 20\%$.

The value of β, which sets an a priori upper bound for the fraction of D vectors in W, was estimated as 0.6. Therefore classifications with $|D| \leq 0.6\ |W|$ were considered only.

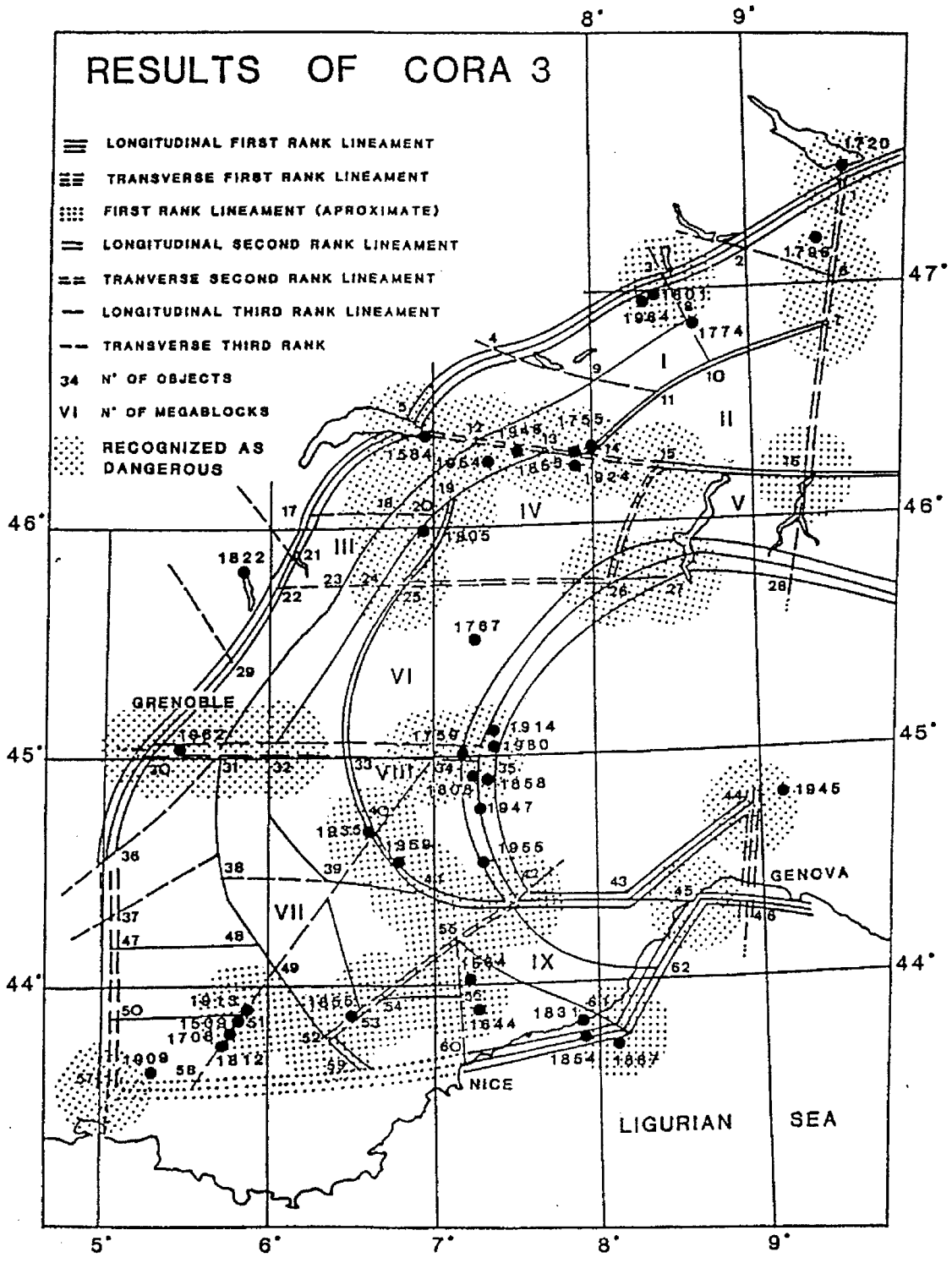

FIGURE 7 The morphostructural scheme of the Western Alps and the result of recognition.

The main case of classifying the 62 binary vectors was obtained through CORA-3 with $k_1 = 3$, $\bar{k}_1 = 2$, $k_2 = 11$, $\bar{k}_2 = 1$, and $\Delta = 0$. The main case resulted in the eleven D traits and eight N traits listed in Table 2. The traits are given in the table as conjunctions of inequalities in the values of the characteristics of the intersections. The classification of the intersections is shown in Fig. 7: 34 intersections are assigned to class D, and the remaining 28 to N. Class D includes all D_0, 11 intersections from N_0, and 9 intersections from outside the training sets.

TABLE 1 Characteristics of intersections in the Western Alps

Functions	Discretization thresholds	
	first	second
Maximum altitude H_{max}, *m*	2686	4807
Minimum altitude H_{min}, *m*	325	-
Altitude at the intersection H_0, *m*	490	900
Distance between points where H_{max} and H_{min} are measured *l*, *km*	32	42
$\Delta H = H_{max} - H_{min}$, *m*	2500	-
Altitude gradient $\Delta H/l$, *m/km*	51	91
Combinations of large topographic forms (yes, no)		
mountain ranges separated by a longitudinal valley (m/m)		
a mountain range and a piedmont plain (m/p)		
a mountain range and piedmont hills (m/pd)		
piedmont hills and piedmont plain (pd/p)		
The percentage of Quaternary deposits *Q*, %	10	-
The highest rank of the lineament R_h	1	2
The number of lineaments forming an intersection n_l	2	-
The number of lineaments in a circle of 25 *km* radius N_l (3 thresholds)	2	3, 4
The distance to the nearest intersection ρ_{int}, *km*	20	31
The distance to the nearest first rank lineament ρ_1, *km*	0	32
The distance to the nearest second rank lineament ρ_2, *km*	0	40
The maximum value of Bouguer anomaly B_{max}, *mGal*	-82	-8
The minimum value of Bouguer anomaly B_{min}, *mGal*	-145	-85
$\Delta B = B_{max} - B_{min}$, *mGal*	45	65
$\bar{B} = (B_{max} + B_{min})/2$, *mGal*	-110	-44
$HB = 0.1\ H_{max}\ [m] + B_{min}\ [mGal]$	153	-
The minimum distance between two Bouguer anomaly isolines spaced at 10 *mGal* $(\nabla B)^{-1}$, km	2	3

TABLE 2 Characteristic traits selected by algorithm CORA-3 for the Western Alps

#	Q, %	n_1	N_1	ρ_1, *km*	ρ_2, *km*	ΔB, *mgl*	$(\nabla B)^{-1}$, km
	D traits						
1				≤32		≤65	≤2
2				>0		≤65	≤2
3				≤32	0	≤65	
4			>3		0	≤65	
5			>4			>45	≤3
6					>0; ≤40	>45	
7		2		>32		>45	
8		2		>32			≤3
9		>2	≤3				≤2
10	>10		>3		≤40		
	N traits						
1						≤45	>2
2					>0	≤45	
3		2				≤45	
4					>40	≤45	
5					>40		>2
6		2			>40		
7		2	≤3		>0		
8		2		0			

7.2 Intermediate-term Prediction of Earthquakes

The pattern recognition methods were used to develop the intermediate-term earthquake prediction algorithm, was initially applied to California-Nevada region and is called algorithm CN (*Keilis-Borok and Rotwain*, 1990).

Objects of Recognition

The objects are moments of time. These moments are described by the functions defined in the lecture "Integrals over Seismic Sequences" (*Kossobokov and Novikova*, 2003). The selection of the moments and the forming of the training sets D_0 and N_0 are described below.

Three types of time periods can be defined for the time from t_0 to T_k, covered by the earthquake catalog of some region:

- periods, which precede strong earthquakes, - periods *D*;
- periods, which follow strong earthquakes, - periods *X*;
- periods, which are not connected with strong earthquakes, - periods *N*.

The formal definition can be formulated as follows.

Let $t_1, t_2, \ldots, t_m$ ($t_0 < t_1 < t_2 < \ldots < t_m < T_k$) be the moments of strong earthquakes of the region under consideration. Here strong earthquakes are the main shocks with magnitude $M \geq M_0$, where M_0 is a given threshold.

Periods *D* are time intervals from $t_i - \Delta t_D$ to t_i ($i = 1, 2, \ldots, m$).

Periods *X* are time intervals from t_i to $t_i + \Delta t_X$ out of periods *D*.

Periods *N* are intervals from t_0 to T_k which remain after exclusion of all periods *D* and *X*.

Here $i = 1, 2, \ldots, m$; Δt_D and Δt_X are given constants.

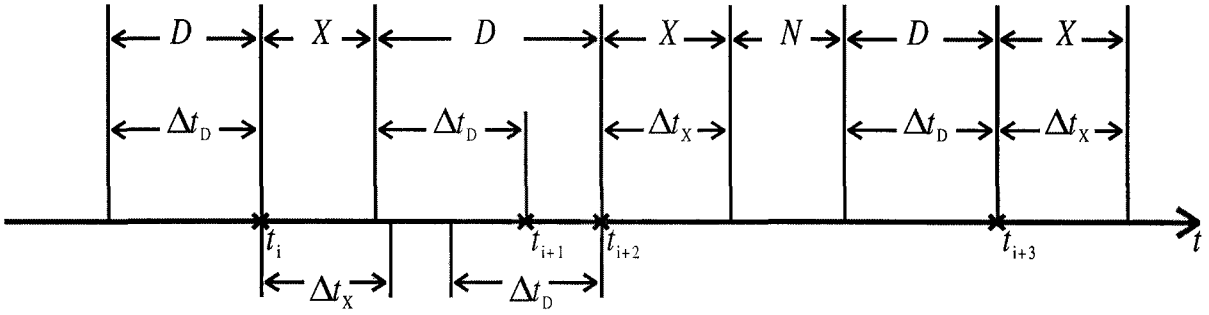

FIGURE 8 Periods D, N, and X.

Example of periods D, X, and N is shown in Fig. 8. The moments t_i, t_{i+1}, t_{i+2}, and t_{i+3} in this figure are the moments of four strong earthquakes.

Moments of time are considered as objects of recognition. For time period from t_0 to T_k three types of moments are defined: D_0, N_0, and X.

Moments D_0 (the set D_0) are the moments before strong earthquakes. For each strong earthquake with origin time t_i the interval from $t_i - \Delta t_D$ to $t_i - \delta t$ is divided into k equal parts of the length $\Delta t_2 = \Delta t_1/k$ where $\Delta t_1 = \Delta t_D - \delta t$. Here $\delta t \geq 0$ and k are specified so to satisfy the relationship $\delta t << \Delta t_2$.

D_0 consists of the moments

$t_i^j = t_i - \Delta t_D + j\Delta t_2$

where $j = 0, 1, 2, ..., k$. The moments D_0, which are earlier than the origin time t_{i-1} of the preceding strong earthquake, are eliminated (see Fig. 9b).

Moments N are selected within periods N with equal steps, unless there is not specific reason to do otherwise.

Moments N_0 (the set N_0) are selected from moments N to be regularly distributed among them. The number of moments N_0 is usually selected about the same as the number of moments D_0.

Moments X are selected from periods X with step Δt_2.

Subclasses

Subclasses are formed of the moments D_0. One subclass includes moments D_0, which precede the same strong earthquake.

Let t_{i-1} and t_i are origin times of two consecutive strong earthquakes. If $t_i - t_{i-1} > \Delta t_D$ then the subclass connected with the strong earthquake numbered i consists of the following moments D_0:

$t_i^j = t_i - \Delta t_D + j\Delta t_2$

where $j = 0, 1, 2, ..., k$. If $t_i - t_{i-1} \leq \Delta t_D$ then only moments t_i^j, which are after t_{i-1} ($t_i^j > t_{i-1}$), are included in the subclass.

Two subclasses are shown in Fig. 9a. The first corresponds to the strong earthquake occurred at time t_{i-1} and consists of three moments D_0: t_{i-1}^0, t_{i-1}^1, and t_{i-1}^2. The second, connected with the strong earthquake, occurred at time t_i, consists also of three moments D_0: t_i^0, t_i^1, and t_i^2. Fig. 9b shows another example of two subclasses. The first, connected with the strong earthquake, occurred at time t_{i-1}, consists also of three moments D_0: t_{i-1}^0, t_{i-1}^1, and t_{i-1}^2, and the second, connected with the strong earthquake, occurred at time t_i, contains only two moments D_0: t_i^1 and t_i^2.

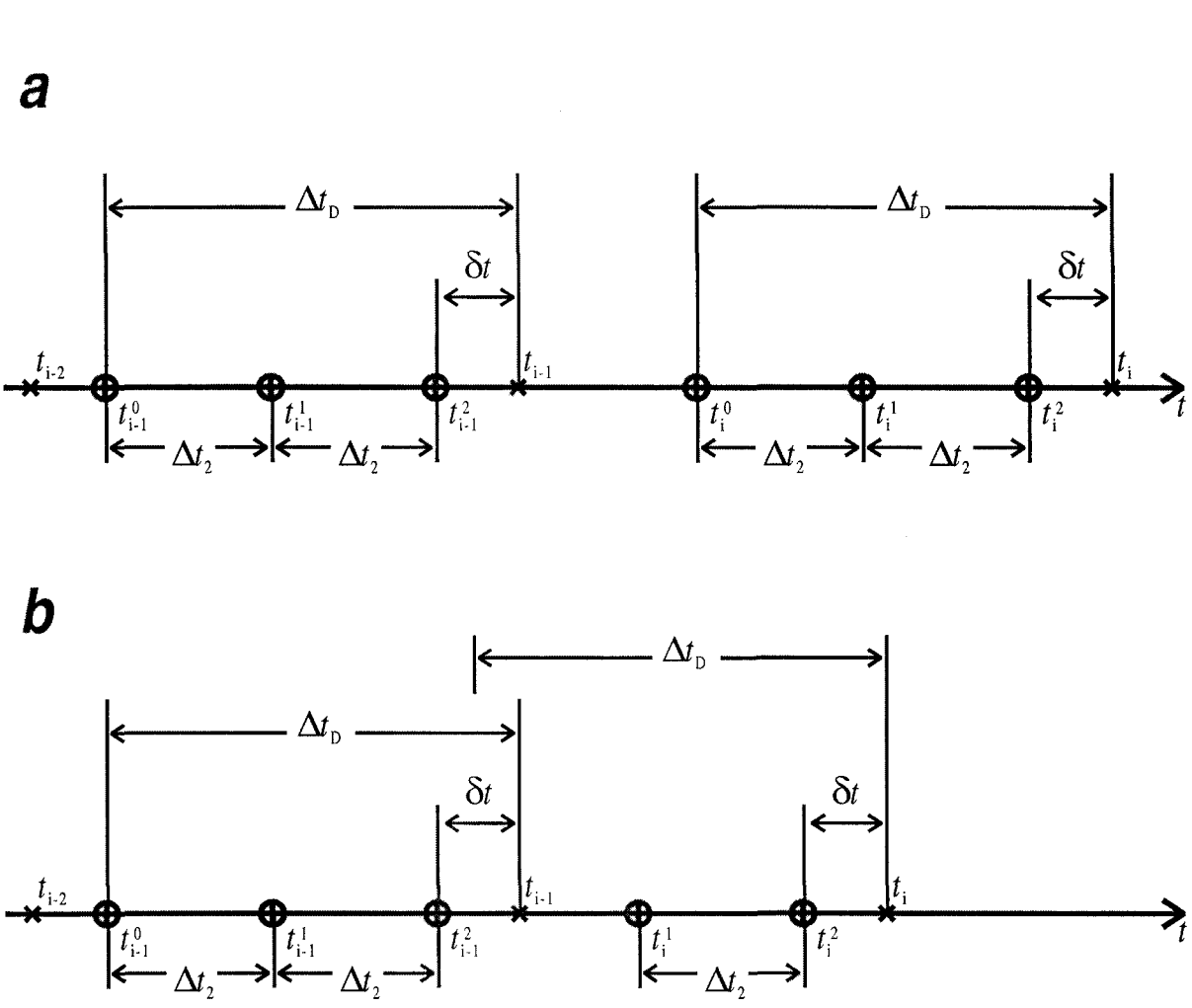

FIGURE 9 Moments D_0 (marked by $\oplus$), $k = 2$.

Algorithm CN

The earthquake catalog of the Southern California for the time period 1938-1984 was used to determine the training set. The threshold magnitude for the strong earthquakes was $M_0 = 6.4$. Table 3 contains the thresholds for discretization of the functions on the earthquake flow, calculated for these moments. The binary coding of all functions was ***S*** type (see Section 3.2).

The algorithm CLUSTERS (Section 4.3) with $k_1 = 7, \bar{k}_1 = 2, k_2 = 10, \bar{k}_2 = 4$ was applied to obtain the characteristic traits of classes *D* and *N*. These traits are listed in Table 4. The time moments are classified by using these traits and $\Delta = 5$. If a moment t is attributed to class *D* then this moment is considered to belong to a period of the time of increased probability (TIP) of a strong earthquake. Formally if t is attributed to class *D* then a TIP is diagnosed from t to $t + \tau$ where τ is a given constant. The value $\tau = 1$ year is used for the Southern California

TABLE 3 Thresholds for discretization of functions on the earthquake flow (Southern California)

Function	Thresholds	
N2	0	-
K	-1	1
G	0.5	0.67
SIGMA	36	71
Smax	7.9	14.2
Zmax	4.1	4.6
N3	3	5
q	0	12
Bmax	12	24

TABLE 4 Characteristic traits of classes *D* and *N* obtained by algorithm CLUSTERS for the moments of the Southern California
(traits of the algorithm CN)

Traits *D*	N2	K	G	SIGMA	Smax	Zmax	N3	q	Bmax
1		0							0
2								0	
3							0	0	0
4						0		0	
5		0					1		0
6		1				0			0
7		0				1			0
8		0	0						0
9					0	0			
10		1		0		0			
11		0 1				0			
12	0	1				0			
13		0			1				
14		0			0				
15		0		0					
16		0	1						

Traits *N*	N2	K	G	SIGMA	Smax	Zmax	N3	q	Bmax
1					1				1
2						1			1
3				1				1	1
4		1						1	1
5							0	1	1
6					1				1
7		1				1			1
8	1					1			1
9				1			0	1	
10					1			1	
11						1	0		
12					1		0		
13		1			1				
14		1		1					
15		1				1			
16		1	1		1				
17		1			1				
18		1		1					

REFERENCES

Alekseevskaya,M.A., A.M.Gabrielov, A.D.Gvishiani, I.M.Gelfand, and E.Ya.Ranzman (1977). Formal morphostructural zoning of mountain territories. *J. Geophys*, **43**, 227-233.

Caputo,M., V.Keilis-Borok, E.Oficerova, E.Ranzman, I.Rotwain, and A.Solovjeff (1980). Pattern recognition of earthquake-prone areas in Italy. *Phys. Earth Planet Int.*, 21: 305-320.

Cisternas,A., P.Godefroy, A.Gvishiani, A.I.Gorshkov, V.Kossobokov, M.Lambert, E.Ranzman, J.Sallantin, H.Soldano, A.Soloviev, and C.Weber (1985). A dual approach to recognition of earthquake prone areas in the western Alps. *Annales Geophysicae*, **3**, 2: 249-270.

Gelfand,I.M., Sh.Guberman, M.L.Izvekova, V.I.Keilis-Borok, and E.Ia.Ranzman (1972). Criteria of high seismicity determined by pattern recognition. In A.R.Ritsema (ed.), *The Upper Mantle. Tectonophysics*, **13** (1-4): 415-422.

Gelfand,I.M., Sh.I.Guberman, M.P.Zhidkov, M.S.Kaletskaya, V.I.Keilis-Borok, and E.Ja.Ranzman (1973). Transfer of high seismicity criteria from the East of Central Asia to Anatolia and adjacent regions. *Doklady Academii Nauk SSSR*, **210**, 2: 327-330 (in Russian).

Gelfand,I.M., Sh.I.Guberman, M.P.Zhidkov, M.S.Kaletskaya, V.I.Keilis-Borok, E.Ja.Ranzman, and I.M.Rotwain (1974a). Identification of sites of possible strong earthquake occurrence. II. Four regions of Asia Minor and South Eastern Europe. In V.I.Keilis-Borok (ed.), *Computer Analysis of Digital Seismic Data*. Moscow, Nauka: 3-40 (Comput. Seismol.; Iss. 7, in Russian).

Gelfand,I.M., Sh.I.Guberman, M.P.Zhidkov, V.I.Keilis-Borok, E.Ja.Ranzman, and I.M.Rotwain (1974b). Identification of sites of possible strong earthquake occurrence. III. The case when the boundaries of disjunctive knots are unknown. In V.I.Keilis-Borok (ed.), *Computer Analysis of Digital Seismic Data*. Moscow, Nauka: 41-65 (Comput. Seismol.; Iss. 7, in Russian).

Gelfand,I.M., Sh.A.Guberman, V.I.Keilis-Borok, L.Knopoff, F.Press, I.Ya.Ranzman, I.M.Rotwain, and A.M.Sadovsky (1976). Pattern recognition applied to earthquake epicenters in California. *Phys. Earth Planet. Inter.*, **11**: 227-283.

Gorshkov,A.I. G.A.Niauri, E.Ya.Ranzman and A.M.Sadovsky (1987). Use of gravimetric data for recognition of places of possible occurance of strong earthquakes in the Great Caucasus. In V. I. Keilis-Borok and A. L. Levshin (eds), *Theory and Analysis of Seismological Information*, Allerton Press Inc, New York: 117-123 (Comput. Seismol., volume 18).

Gvishiani,A.D., A.V.Zelevinsky, V.I.Keilis-Borok, and V.G.Kossobokov (1978). Study of the violent earthquake occurrences in the Pacific Ocean Belt with the help of recognition algorithms. *Izvestiya Acad. Sci. USSR. Physics of the Earth*, 9: 31-42 (in Russian).

Gvishiani,A.D. and V.G.Kossobokov (1981). On foundations of the pattern recognition results applied to earthquake-prone areas. *Izvestiya Acad. Sci. USSR. Physics of the Earth*, 2: 21-36 (in Russian).

Gvishiani,A., and A.Soloviev (1981). Association of the epicenters of strong earthquakes with the intersections of morphostructural lineaments in South America. In V. I. Keilis-Borok and A. L. Levshuin (eds), *Interpretation of Seismic Data - Methods and Algorithms*, Allerton Press Inc., New York: 42+ (Comput. Seismol., volume 13).

Gvishiani,A.D., and A.A.Soloviev (1984). Recognition of places on the Pacific coast of the South America where strong earthquakes may occur. *Earthq. Predict. Res.*, 2: 237-243.

Gvishiani,A., A.Gorshkov, V.Kossobokov, A.Cisternas, H.Philip, and C.Weber (1987). Identification of seismically dangerous zones in the Pyrenees. *Annales Geophysicae*, 5B(6): 681-690.

Keilis-Borok,V.I. and A.J.Lichtman (1981). Pattern recognition applied to presidential elections in the United States, 1860-1980: role of integral social, economic and political traits. *Proceedings of US National Ac. Sci.*, 78, 11: 7230-7234.

Keilis-Borok,V.I. and I.M.Rotwain (1990). Diagnosis of Time of Increased Probability of strong earthquakes in different regions of the world: algorithm CN. *Phys. Earth Planet. Inter.*, **61**: 57-72.

Kossobokov,V.G. (1983). Recognition of the sites of strong earthquakes in East Central Asia and Anatolia by Hamming's method. In V.I.Keilis-Borok and A.L.Levshin (eds), *Mathematical Models of the Structure of the Earth and the Earthquake Prediction*, Allerton Press Inc, New York: 78-82 (Comput. Seismol., volume 14).

Kossobokov,V.G., and A.A.Soloviev (1983). Disposition of epicenters of earthquakes with $M \geq 5.5$ relative to the intersection of morphostructural lineaments in the East Central Asia. In V.I.Keilis-Borok and A.L.Levshin (eds), *Mathematical Models of the Structure of the Earth and the Earthquake Prediction*, Allerton Press Inc, New York: 75-77 (Comput. Seismol., volume 14).

Kossobokov,V.G., and O.Novikova (2003). Integrals over Seismic Sequences. Seventh Workshop on Non-Linear Dynamics and Earthquake Prediction, 29 September – 11 October 2003, Trieste: the Abdus Salam ICTP.

Ranzman,E.Ia. (1979). *Places of Earthquakes and Morphostructures of Mountain Countries*. Nauka, Moscow (in Russian).

Zhidkov,M.P., I.M.Rotwain, A.M.Sadovsky (1975). Recognition of places where strong earthquakes may occur. IV. High-seismic intersections of lineaments of the Armenian upland, the Balkans and the Aegean Sea basin. In V.I.Keilis-Borok (ed.), *Interpretation of Seismology and Neotectonics Data*, Moscow, Nauka: 53-70 (Comput. Seismol.; Iss. 8, in Russian).

Zhidkov,M.P., and V.G.Kossobokov (1980). Identification of the sites of possible strong earthquakes - VIII: Intersections of lineaments in the east of Central Asia. In V.I.Keilis-Borok (ed.), *Earthquake Prediction and the Structure of the Earth*, Allerton Press Inc, New York: 31-44 (Comput. Seismol., volume 11).