INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION

INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
34100 TRIESTE (ITALY) - P.O.B. 586 - MIRAMARE - STRADA COSTIERA 11 - TELEPHONES: 224281/2/3/4/5/6
CABLE: CENTRATOM - TELEX 460392-I

SMR/107 - 6

WORKSHOP ON PATTERN RECOGNITION AND ANALYSIS OF SEISMICITY

(5 - 16 December 1983)

PATTERN RECOGNITION APPLIED TO EARTHQUAKE EPICENTERS IN CALIFORNIA

I.M. GELFAND
S.A. GUBERMAN
V.I. KEILIS-BOROK
L. KNOPOFF
F. PRESS
E.Y. RANZMAN
I.M. ROTWAIN
A.M. SADOVSKY

Physics of the Earth and Planetary Interiors, 11 (1976) 227–283
© Elsevier Scientific Publishing Company, Amsterdam – Printed in The Netherlands

PATTERN RECOGNITION APPLIED TO EARTHQUAKE EPICENTERS IN CALIFORNIA

I.M. GELFAND [1], Sh.A. GUBERMAN [1], V.I. KEILIS-BOROK [2], L. KNOPOFF [3], F. PRESS [4], E.Ya. RANZMAN [5], I.M. ROTWAIN [6] and A.M. SADOVSKY [2]

[1] *Institute of Applied Mathematics, Academy of Sciences, Moscow (U.S.S.R.)*
[2] *Institute of Physics of the Earth, Academy of Sciences, Moscow (U.S.S.R.)*
[3] *Institute of Geophysics, University of California, Los Angeles, Calif. (U.S.A.)*
[4] *Department of Earth and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Mass. (U.S.A.)*
[5] *Institute of Geography, Academy of Sciences, Moscow (U.S.S.R.)*
[6] *Laboratory of Mathematical Methods in Biology, Moscow University, Moscow (U.S.S.R.)*

(Received October 3, 1975; accepted for publication November 21, 1975)

"This is the only way it clicks ...
It doesn't make sense otherwise ...
But this is just a theory, isn't it? ...
Call it any name you like. It's good enough for me."

Dashiel Hammett

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

A pattern recognition procedure is explained which uses geological data and the earthquake history of a region, in this case California, and learns how to separate earthquake epicenters from other places. Sites of future earthquake epicenters are predicted as well as places where epicenters will not occur. The problem is formulated in several ways and control experiments are devised and applied in order to test the stability of the procedures and engender confidence in the results. Some of the combinations of geological features which the computer recognized as significant discriminants are discussed.

1. Introduction

Pattern recognition is not a new subject but its application to geological and geophysical problems has been minimal. This is surprising because the methods of pattern recognition, though paradoxically simple, can extract more results from a body of data, by a more intensive analysis, than many alternative procedures.

Geology and geophysics are essentially experimental disciplines in which large amounts of data are accumulated and are in need of winnowing, codification, correlation, and interpretation. More so than in many other fields, hypotheses flow from data analysis, and prediction is important in a practical sense as well as a means of hypothesis testing. Pattern recognition offers a powerful tool for achieving these results. Yet care must be taken in its application, for pattern recognition itself is an unexplored field lying somewhere between logic and statistics. Its procedures are not yet formalized and the user must be on guard against self-deception.

In this paper we examine the possibility that geological patterns can be recognized which distinguish places where epicenters of large earthquakes have occurred in the past and can occur in the future, from other places which have not been and will not be the sites of epicenters. The primary data for recognition come from California, though insights and procedures stem from earlier applications of these methods to central Asia and Anatolia. If we are successful we will have predicted the sites of future epicenters (but not the times) and we will have learned something about the geological environment of epicenters.

2. Method

2.1. The problem

We place the problem of recognition in the following form: a set of objects is given, with each object described by the answers to a questionnaire. Each object belongs to one and only one of several (usually two) classes. Our goal is to find which class each object belongs to. To solve this problem we first need to go through a "learning phase" using examples of objects of each class as "learning material".

Let us make this specific for our particular problem. Objects are points on a map, representative of a local geographic region. The problem is to find those objects where the epicenters of strong earthquakes may occur in the future. Note that we mean the epicenter or place where the earthquake is initiated and not the entire length of faulting. We shall call these objects D, "dangerous". The rest of the objects we shall call N, "non-dangerous". By strong we mean earthquakes with magnitude M not lower than some threshold M_0; in the text below, "earthquake" and "epicenter" refer to strong earthquakes unless indicated otherwise.

A specific difficulty arises because the learning material consisting of known epicenters is mixed. Because of errors in locating epicenters, objects called D in the learning phase may really be N and vice versa. Also some objects which are classified N in the learning phase are really D because they will be the sites of future epicenters. It is these objects we wish to identify by pattern recognition in order to predict future epicenters. This leads us to the following problem: the objects are divided a priori into classes I and II – correspondingly close to and far from known epicenters; the goal is to recognize objects D in each class.

2.2. The algorithm

The recognition algorithm consists of three stages: learning, voting and control experiments.

Learning: certain critical questions or combinations of questions, called the distinctive features of objects D and N, are determined by analyzing the learning material, as will be described in the next two sections.

Voting: the numbers n_D and n_N of distinctive features characteristic of D and N are determined for each object. Recognition is based on their difference $\Delta = n_D - n_N$. We recognize an object as D if $\Delta \geqslant \bar{\Delta}$, where $\bar{\Delta}$ is some threshold. Making $\bar{\Delta}$ smaller will reduce the number of D objects which will be misidentified as N. At the same time more "false alarms" will occur, i.e. more N will be misidentified as D.

Control experiment: the stability and reliability of recognition are tested by numerical (logical) experiments, in order to convince ourselves that the results are significant since recognition cannot be proven either mathematically or statistically. This is a decisive stage which takes up most of the effort, after the questionnaire is formulated and answered.

Learning is based on the algorithm CORA-3. This algorithm in slightly different form has been used in several other problems (Bongard et al., 1966). For our particular problem it can be formulated as follows. The objects, used in learning, are divided into classes I and II, correspondingly close to and far from known epicenters. As stated earlier, each class may contain objects D and N:

class I is $\{D_{\mathrm{I}}, N_{\mathrm{I}}\}$

class II is $\{D_{\mathrm{II}}, N_{\mathrm{II}}\}$

Here D_i and N_i are objects D and N, respectively, in the ith class; i = I and II. In other words, D_{I} and N_{I} are objects D and N, which lie close to at least one known epicenter. D_{II} and N_{II} are the objects D and N, which lie far from all known epicenters. Each object is described in binary code, which contains the answers to the questionnaire. All binary digits and the combinations of two and three digits are considered as traits of an object.

We repeat this definition more formally.

The object A is described in binary code:

$$A = A_1 A_2 \ldots A_L \qquad (1)$$

where the component A_i is the answer to the ith question in the questionnaire, expressed as 1 for yes and 0 for no. *The trait* τ is an array of six integers:

1

$$\tau = p, q, r, P, Q, R \qquad (2)$$

Here $p = 1, 2, \ldots, L$; $q = p, p+1, \ldots, L$; $r = q, q+1, \ldots, L$; L is the same as in object (1); $P = 0$ or 1; $Q = 0$ or 1; $R = 0$ or 1. If $q = r$, then $Q = R$; if $p = q = r$, then $P = Q = R$; if $q = p$, then $Q = P$. An object (1) has the trait (2), if $A_p = P$; $A_q = Q$; $A_r = R$. Here A_p is the pth component in (1) etc. In the cases $p = q = r$ and $p \neq q$, $q = r$ the trait corresponds to one and two components of (1), respectively. The cases $p = q$, $q \neq r$ are redundant.

As an example, the object A = 0110 has traits as represented in Table I.

Thus the value of PQR in the trait with pqr equal to 222, for example, is the response to question $i = 2$ in A and has been identified in the pqr array by the value of P; the values of q and r are redundant and have been set equal to 2 to provide a complete syntactical statement. Similarly, the value of PQR in the trait with pqr equal to 233, for example, is the response to both questions $i = 2$ and 3 in A and is identified in pqr by the value of p and q; the value of r is redundant. The value of PQR in trait with pqr equal to 134 is the response to questions $i = 1, 3$ and 4 in A.

Thus, an object (1) has $L + C_{L2} + C_{L3}$ traits, where C_{LM} is the number of combinations of L things taken M at a time ($C_{L1} = L$). Eight different traits are possible for fixed p,q,r, if $p \neq q \neq r$: $(P,Q,R) = (0,0,1)$ or (0,1,0) etc. up to (1,1,1). If $p \neq q$, $q = r$, only four different traits are possible for fixed p,q: $(P,Q) = (0,0)$, (0,1), (1,0) or (1,1). If $p = q = r$, only two different traits are possible for fixed p: $P = 0$ or 1.

Therefore, the total number of different traits which are possible in objects of length L is:

$$2L + 4C_L^2 + 8C_L^3$$

The features of D and N are selected from among the traits. A *feature* of D is a trait which is present relatively frequently in class I and relatively infrequently in class II. A feature of N is defined analogously. The formal rule defining features is given in Table II.

Next we eliminate equivalent and weaker features, whose definitions follow. Consider two features of a class and the objects of this class only. Some set of these objects has the first feature, another has the second. If these sets coincide, the features are *equivalent.* If one set is part of another, the corresponding feature is *weaker* than the other. If $M_n/M_u \geqslant 1 - \epsilon$, these features are called "ϵ-equivalent". Here M_n is the number of objects which belong to both sets; M_u is the total number of objects which belong to both or to any one of the sets. Evidently, for equivalent features, $M_n = M_u$, i.e. $\epsilon = 0$. The definition of ϵ-equivalence is used later.

We eliminate all but one feature from each group of equivalent features, and all features which are weaker than some other one. The features remaining after this elimination are called distinctive features and are used in voting. As mentioned above voting consists of tallying the number of D and N distinctive features at each object of class I, II and III.

Class III consists of objects not used in learning.

TABLE I

Traits for the object A = 0110

p	q	r	P,Q,R
1	1	1	0
2	2	2	1
3	3	3	1
4	4	4	0
1	2	2	0,1
1	3	3	0,1
1	4	4	0,0
2	3	3	1,1
2	4	4	1,0
3	4	4	1,0
1	2	3	0,1,1
1	2	4	0,1,0
1	3	4	0,1,0
2	3	4	1,1,0

TABLE II

Definition of features* in learning by algorithm "CORA-3"

A feature of objects	Should be found in the class	
	I	II
D	in $\geqslant \underline{k}_1$ objects	in $\leqslant \tilde{k}_1$ objects
N	in $\leqslant k_2$ objects	in $\geqslant k_2$ objects

* Each feature is a unit or a combination of two or three units in the binary code of the object.

3

We could divide all objects into classes I and II only, according to whether or not the point lies near an epicenter. However, we separate some objects into class III if we are uncertain about their assignment or if we use them for control experiments. Objects in class III are considered in voting, however.

The algorithm assumes that the number of epicenters is sufficient for the following two hypotheses to hold:

Hypothesis 1: most objects, which are far from known epicenters, are N.

Hypothesis 2: most objects near known epicenters are D.

The second hypothesis will be replaced by a weaker one, when we use another algorithm for learning called "CLUSTERS", which we describe next.

Algorithm "CLUSTERS" is fully described for the first time in this paper; a brief description was given in Gelfand et al. (1974a). The objects of class I are subdivided into clusters. Each cluster corresponds to a certain epicenter and includes those objects which are close to this epicenter. An object may be included in several clusters. Class II is specified as before. Thus:

Class I is $\{C[D_1^C, N_1^C]\}$

Class II is $\{D_{II}, N_{II}\}$

Here C is a cluster of objects near some epicenter; D_1^C and N_1^C are the objects D and N in the cluster C. The problem is to recognize D_I and D_{II}.

As in CORA-3 all traits of all objects in the learning materials are considered in the learning phase. We assume that a cluster has some trait if at least one object in the cluster has this trait. (In another version this condition is more flexible in that at least N/K objects in the cluster must have this trait. Here N is the number of the objects in the cluster, and K is some constant common for all clusters; naturally, if $N/K < 1$ it is replaced by 1.) With these definitions the features are selected by the rules specified in Table III.

As before equivalent and weaker features are then eliminated. Although they can be determined exactly as in CORA-3, we preferred to use the clusters rather than objects in testing for equivalent or weaker D features. After this elimination we are left with the distinctive features of D and N, which are used in voting. The voting is carried out separately for each

TABLE III

Definition of features* in learning by algorithm "CLUSTERS"

A feature of objects	Should be found in the clusters C **	single objects in clusters (class I)	single objects outside the clusters (class II)
D	in $\geq k_1$ clusters	in $\geq k_3$ objects	$\leq \bar{k}_1$ objects
N	–	in $\leq \bar{k}_2$ objects	$\geq k_2$ objects

* Each feature is a unit or a combination of two or three units in the binary code of the object.

** A cluster has some trait if at least one object in the cluster has this trait.

object, whether or not it has been assigned to a cluster.

This voting enables us to recognize the objects D near each known epicenter (D_I) and far from all of them (D_{II}). In particular we can divide each cluster into D_I and N_I. The essential advantage, compared to CORA-3 is that objects D_I need not dominate the clusters. They may even be in the minority in many clusters, or in every cluster if $k_3 = 0$.

Formally this algorithm differs from CORA-3 only by the introduction of an additional threshold (compare Tables II and III). It may seem therefore, that the algorithm constrains the learning materials more severely. In fact, the opposite occurs, since k_3 in Table III can be much smaller than k_1 in Table II. Frequently we even take $k_3 = 0$. In effect we have replaced hypothesis 2 by a much broader one:

Hypothesis 3: most of the known epicenters have at least one D object nearby.

Hypotheses 1 and 3 seem to be close to minimally necessary ones. Without them our problem is hopeless.

2.3. Control experiments

The main control experiment will now be described. Let us neglect the data on a specified part of the list of earthquakes. These earthquakes can be the most recent ones or can be selected otherwise. Objects which are near the epicenters of these and only these earthquakes will be reassigned in the learning stage to class II. We apply the same algorithms of learning and

voting and check whether these objects will be recognized as D_{II}. If the earthquakes we have neglected are the most recent ones we refer to this experiment as experiment EH ("earthquake history"), since it is equivalent to learning on the basis of an early part of an earthquake catalog and testing our ability to predict the subsequent seismic history, i.e. recognize the earthquakes in the later part of the catalog.

We must also take care that the number of objects identified as dangerous does not grow unduly large "in the future". To check this possible instability, we continue the EH experiment into the future, reassigning into group I all objects recognized as D_{II}. This is referred to as "experiment EF" ("earthquake future"). This procedure is evidently iterative and can be carried out on both experiments EH and EF.

As a further check we have compared our results against those obtained using variations of the data, the questionnaire and its answers, the criteria for the selection and classification of objects, the catalog of epicenters, and the values of numerical parameters, including k_i and $\bar{k}_i$. We have also used tests involving random variation and even random generation of data, the use of different algorithms of recognition, and the application of the characteristic features identified in one region to recognition in another, but seismically similar region.

3. Strike-slip earthquakes of California

3.1. The problem

Here we consider strong ($M \geq 6.5$) strike-slip earthquakes in the region shown in Fig. 1a. The catalog of earthquakes is given in Table IV. We assume that the epicenters are associated with major strike-slip faults (referred to as faults in what follows). The system of faults is depicted in Fig. 1a. The *Tectonic Map of the United States* (USGS, 1962) is the prin-

TABLE IV

Epicenters of strike-slip earthquakes with $M \geq 6$

N	Year	Day, month	Latitude (°N)	Longitude (°W)	M	Closest point (see Fig. 1)
1	1836		37.5	121.9	>7	*3*
2	1836		37.7	122.1	>7	*4*
3	1836		37.8	122.3	>7	*5*
4	1857		34.7	118.8	>7	*11*
5	1906	18, IV	38.25	122.95	8.25	*2*
6	1911	1, VII	37.25	121.75	6.6	*6*
7	1918	21, IV	33.75	117	6.8	*13*
8	1922	10, III	35.75	120.25	6.5	*10*
9	1923	22, I	40.5	124.5	7.2	*23*
10	1923	23, VII	34	117.25	6.25	*12*
11	1926	22, X	36.75	122	6.1	*7*
12	1933	11, III	33.6	118	6.25	*22*
13	1934	8, VI	36	120.5	6.0	*9*
14	1934	30, XII	32.25	115.5	6.5	*20*
15	1934	31, XII	32	114.75	7.0	*21*
16	1937	25, III	33.5	116.5	6.0	*15*
17	1940	19, V	32.7	115.5	6.7	*19*
18	1942	21, X	33	116	6.5	*18*
19	1948	4, XII	33.9	116.4	6.5	*14*
20	1951	8, X	40.25	124.5	6.0	*1*
21	1952	22, XI	35.8	121.2	6.0	*8*
22	1954	19, III	33.28	116.18	6.2	*17*
23	1968	9, IV	33.4	116.2	6.9	*16*

cipal source of fault information, with a few additions on our part. The problem is to find where on these faults earthquake epicenters may fall.

3.2. The objects

Objects of recognition are points on faults. We consider points of the following three kinds:

(1) Projection of each epicenter onto the closest fault. Epicenters of all strike-slip earthquakes with $M \geqslant 6.0$ (Table IV) are used. We call these points centers. They have ordinal numbers from *1* to *23* in Fig. 1b.

(2) Points on faults at distances of 25 km from centers. They have ordinal numbers from *24* to *61* in Fig. 1b. To locate these points we drew circles of 25-km radius around each center. The intersections of the circles with the faults are these points of the second kind. We call these points associates of corresponding centers. For example, point *60* is an associate of point *22* in Fig. 1b. If it occurs that two centers lie on the same fault with distance in the range $25 < r \leqslant 50$ km from each other, then the two associates between them lie closer than 25 km. In this case the two associates are replaced by one point in the middle. Points *38* and *48* are examples of such points.

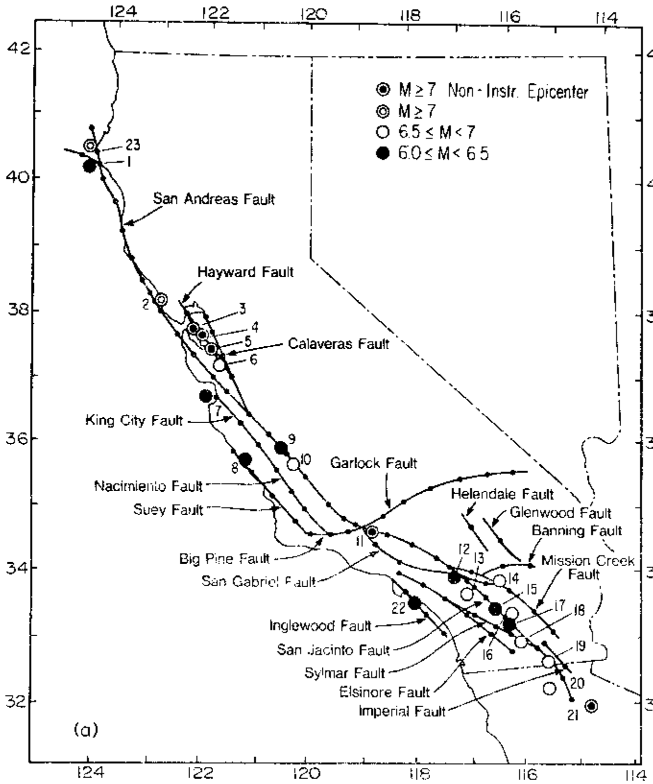


Fig. 1. (a) Major strike-slip faults of California.

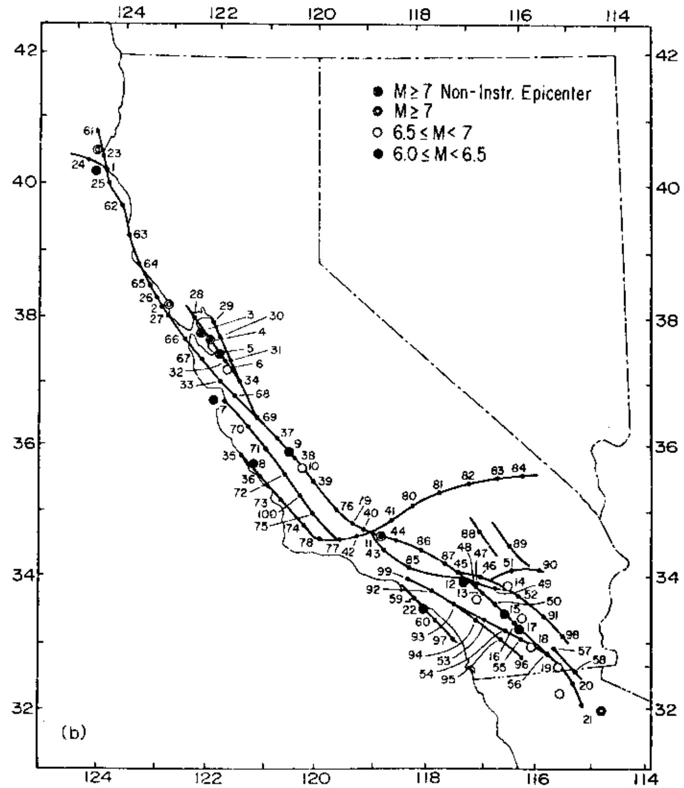


Fig. 1. (b) Identification numbers of objects of recognition (points).

If two centers lie on the same fault with distance $r \leqslant 25$ km, then we eliminate the associate of each center which lies in the direction of the other center on the same fault.

(3) We eliminate from consideration all parts of faults between each center and its associates, i.e. parts of the faults within the above mentioned circles of radius 25 km. On the remaining parts of each fault we place the points at intervals of 50 km, beginning with the northern end of the fault. These are points of the third kind.

If it occurs that a point of the third kind is at the distance $r \leqslant 100$ km from some associate point on the same fault, the next point of the third kind is placed at a distance $r/2$ km (instead of 50 km) from the preceding point. Examples are points *28* and *59*.

The class definitions for objects, in the learning stage, are as follows: Class I contains the centers and their associates (points of the first and second kind), corresponding to earthquakes with $M \geqslant 6.5$. Each of these centers together with its associates forms a cluster. Two clusters were eliminated in the learning stage because the corresponding epicenters occur too far from the nearest fault. Points *16* and *20* are the centers of these clusters. The remaining 13 clusters were used for learning. Class II contains points of the third kind (which are at a distance $r > 25$ km from all centers).

With this classification of objects we proceed through a learning stage in which each cluster (formed of a center and two associate points on the same fault) is in class I and the remaining objects are in class II. We are uncertain about the classification of objects in clusters associated with earthquakes for which $6.0 \leqslant M < 6.5$, so we place them in class III. Some associates were recognized as D. These and only these associates were used with centers to form clusters in the final variant of learning, leading to the results described in later sections. Voting was carried out for all 100 points shown in Fig. 1b.

3.3. The data

We considered the parameters listed in Table V for use in characterizing the objects. Some comments on their choice are in order. The diversity of data that might be used is limited in this case by the small number of known epicenters. Increasing the assortment of data would lead to more distinctive features and in the subsequent voting the objects would divide more readily into two classes. However, this would also increase the probability that the division is random. Thus, we cannot include all data which might be relevant. Our choice depends strongly on physical and geological intuition and experience. Parameters 1–6 and 19–24 (Table V) are considered because they characterize the intensity and degree of contrast of neotectonic movements represented in topography. Parameters 8–11 and 18 are considered because they characterize fracturing of the crust. The rest of the parameters reflect various hypotheses on conditions favorable for the occurrence of strong earthquakes.

Each parameter will be assigned a threshold in the form of an inequality, and Table V becomes a questionnaire applied to each object. In coding the answers to the questionnaire which will be used to describe each object, "1" will signify that the corresponding inequality is satisfied, "0" indicates that it is not satisfied at the object.

Two questions now arise. Which of the parameters in Table V should be used for recognition and what discrete thresholds should we assign to them? Two extreme situations should be avoided. Too narrow a range for the parameters is barred by the limited learning material because too few objects will fall in each interval. On the other hand, with too few parameters and/or overly wide ranges, much needed information might be eliminated and recognition will fail because few if any distinctive features will be found.

The parameters and their assigned thresholds were selected in the following way: For each parameter we determined a one-dimensional distribution of its occurrence among the objects as a function of the value of the parameter. These are shown as histograms in Fig. 2. Dashed lines refer to objects of class I and solid lines to objects of class II; they are more likely D and N, respectively. Those parameters for which the distributions turned out to be sufficiently different were selected for further analysis. They are listed in Table VI, together with the thresholds assigned to discriminate between the two groups. Some comments on the choice of thresholds are in order. For some parameters, such as No. 12, we could have chosen the thresholds in such a way that classes I and II would separate better. Such parameters would then play a dominant role in recognition. However, the sharp division of histograms may be random, since the learning material is rather small. Experience tells us that in such cases many errors in recognition of D_{II} objects can occur. We have found that it is safer to choose the thresholds in such a way that objects of both groups are slightly mixed with no group forming too large a majority in any interval. Then more parameters will be used in distinctive features and recognition as a rule becomes more reliable.

3.4. Results

Distinctive D and N features found by algorithm CLUSTERS are given in Table VIIA. The table lists eight D features and eleven N features, each a combination of two or three parameters. Table VIIB lists distinctive D and N features for an EH experiment in which learning was based on a portion of the earthquake catalogs extending through the year 1922.

Results of voting are given in Table VIII and in Table IX under the column for 1948. We note that recognition is formally successful in two ways: many distinctive features of D and N are found and voting divides the objects into two rather distinctive groups,

TABLE V

Characterization of objects: List of parameters considered

N	Parameter name
1	maximal elevation, h_{max} (m)
2*	minimal elevation, h_{min} (m)
3*	elevation difference, $\Delta h = h_{max} - h_{min}$ (m)
4*	distance between points with minimum and maximum elevation, l_1
5*	"gradient", $\Delta h/l_1$
6*	relative area of soft sediments, q (%)
7*	type of rocks (igneous I, metamorphic M, sediments S)
8	distance to closest fault, r_1 (km)
9	distance to closest intersection of faults, R_1 (km)
10	distance to closest end of fault, R_2 (km)
11	distance to closest end or intersection of faults, r_2 (km)
12	distance to geothermal zone, r_3 (km)
13	distance to a region of large precipitation, R_3 (km)
14	distance to closest large water reservoir, r_6 (km)
15	distance to closest occurrence of Franciscan rocks along the fault, R_4 (km)
16	distance to the closest spreading center, r_4 (km) (40.5° N, 126.6° W or 33.2° N, 115.6° W)
17	distance to the reference point (intersection of San Andreas and Big Pine faults), r_5 (Km)
18*	the number of unnamed faults on the *Tectonic Map of U.S.* (USGS, 1962), N_1
19*	the types of relief (surface morphology) featured across the fault
20*	the number of changes of the types of relief along the fault, n_2
21*	the types of relief featured along the fault
22**	maximal elevation, H_{max} (m)
23**	minimal elevation, H_{min} (m)
24**	elevation difference, $\Delta H = H_{max} - H_{min}$ (m)
25**	distance between points with minimum and maximum elevation, l_2
26**	"gradient", $\Delta H/l_2$
27**	the number of unnamed faults on the *Tectonic Map of U.S.* (USGS, 1962), n_1
28**	the types of relief (surface morphology), featured across the fault
29**	the types of relief featured along the fault
30**	the number of contacts between rocks of different age on the *Geological Map of North America* (USGS, 1965), n_3
31***	the number of parallel faults, n_4
32**	the number of faults, n_5
33***	the number of ends and intersections of faults, n_6
34***	the number of intersections, N_2
35	the angle between the fault and the dominant structural trend in the region

* Measured inside the circle $r = 12.5$ km.
** Measured inside the circle $r = 25$ km.
*** Measured inside the circle $r = 50$ km.

with only a small number of "neutral" results. The latter may be seen from the fact that out of 100 objects, 84 have the distinctive features of only D or only N, and only two points in all clusters have $\Delta < 0$. Experience with recognition based on real as well as random data (see below) indicates to us that these results are encouraging, though nothing more.

The threshold of recognition was chosen as follows:

$$\bar{\Delta} = \Delta^* - \tau, \qquad \Delta^* = \min_c [\max \Delta_c]$$

Here $\max \Delta_c$ is the maximal value of Δ for points in the cth cluster; $\min_c$ signifies the smallest value among all clusters. We subtract τ to decrease the probability of missing a D point; although this increases the probability of false alarms, this is preferable in our problem.

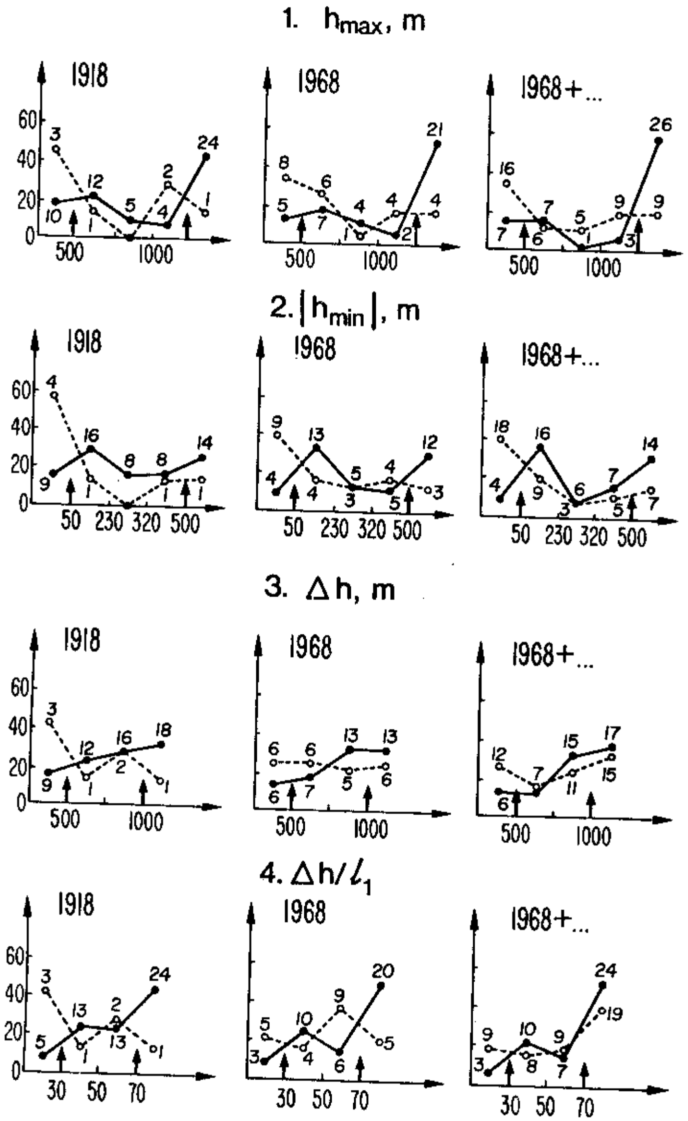


Fig. 2. For caption see p. 245.

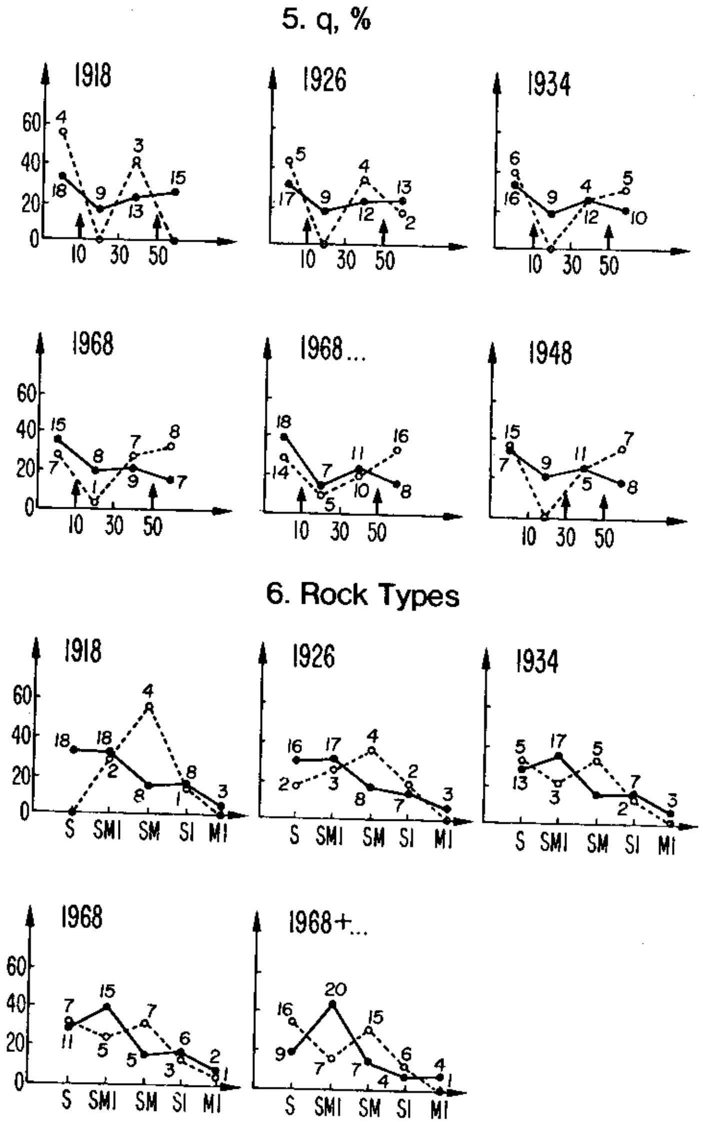
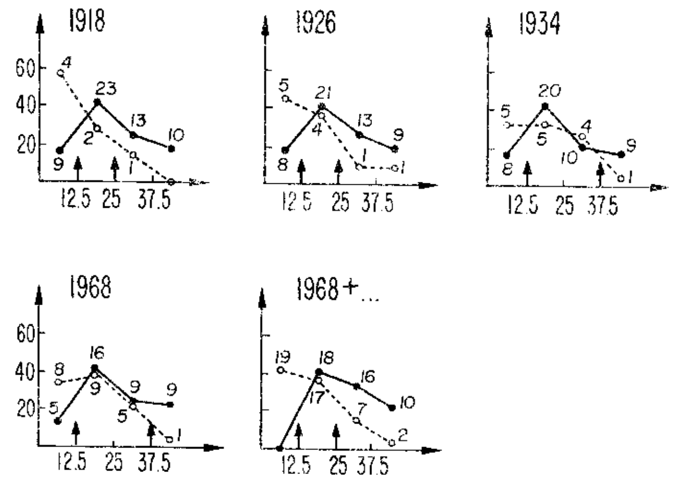


Fig. 2. For caption see p. 245.

7. r_1, km



8. r_2, km

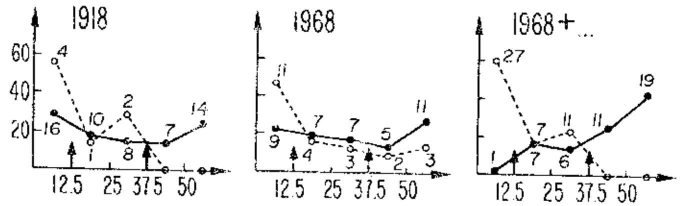
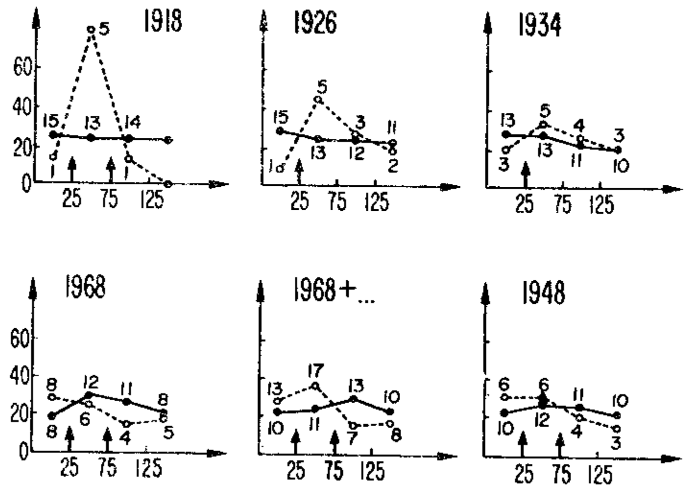


Fig. 2. For caption see p. 245.

12

9. r_3, km



10. r_4, km

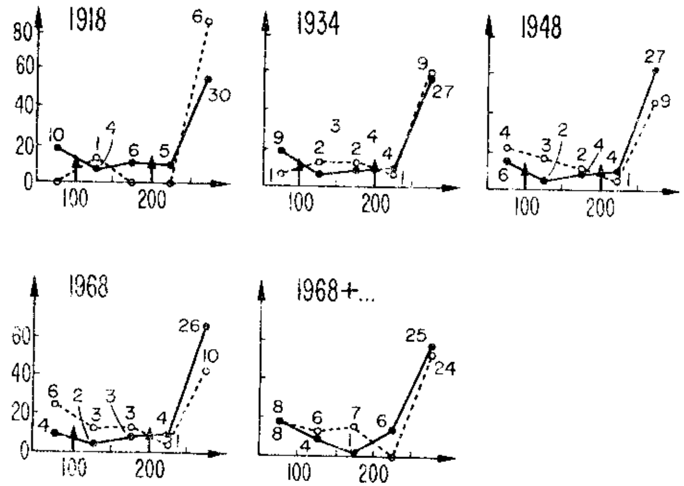
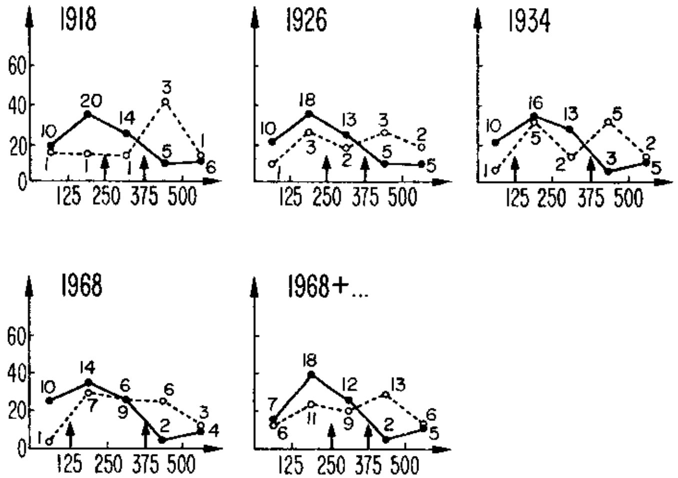


Fig. 2. For caption see p. 245.

13

11. r_5, km



12. r_6, km

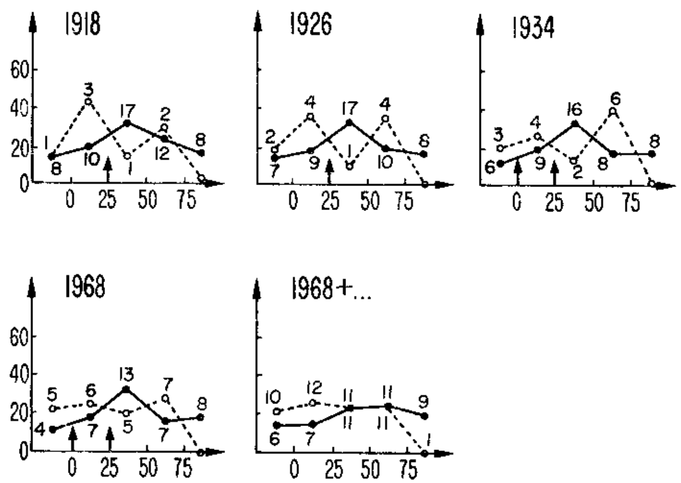
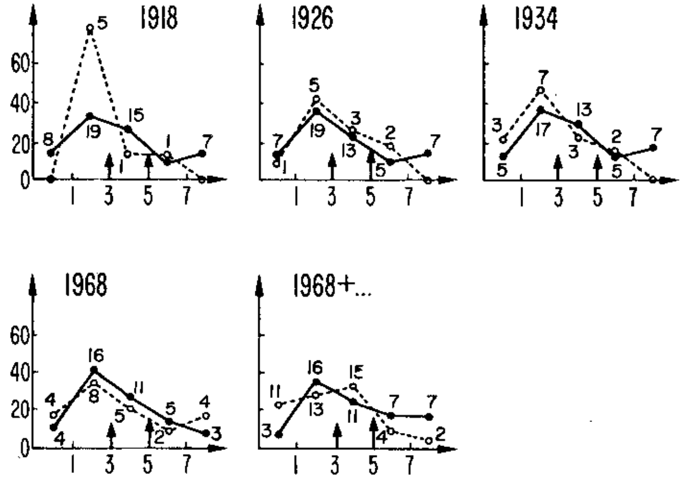


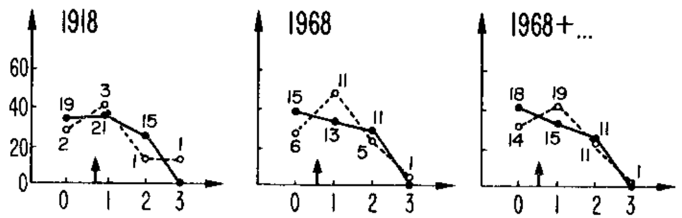
Fig. 2. For caption see p. 245.

14

13. n_1



14. n_2



15. H_{max}, km

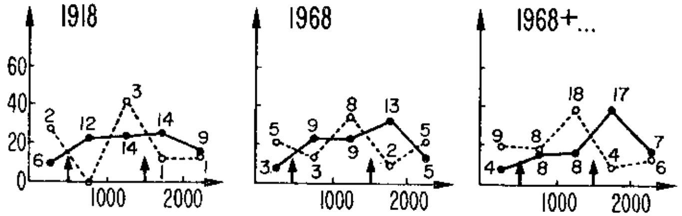
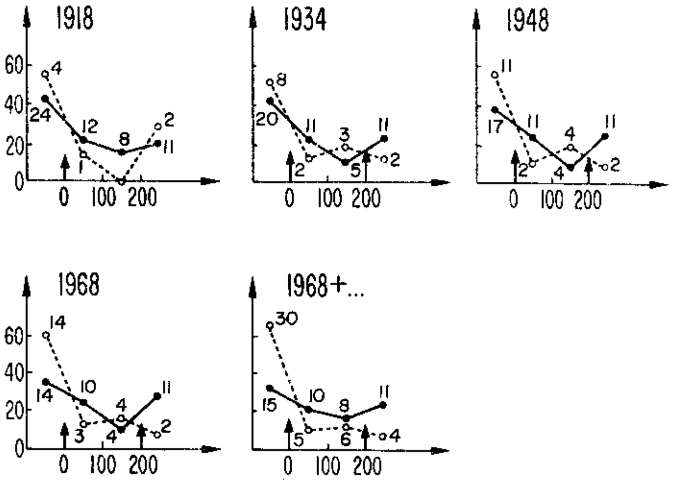


Fig. 2. For caption see p. 245.

15

16. H_{min}, m



17. n_3

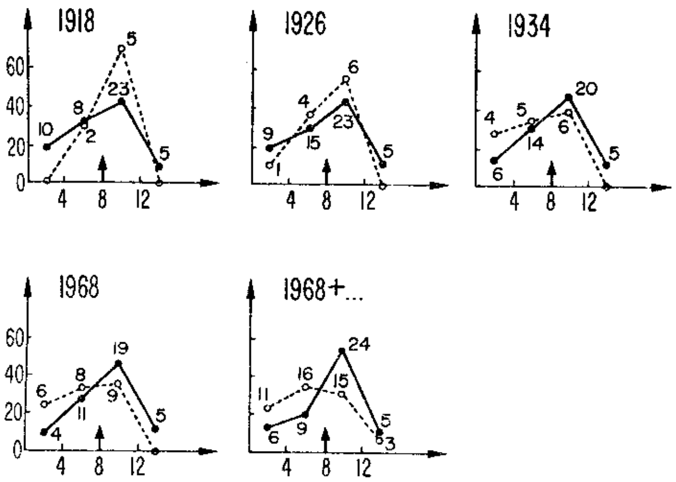
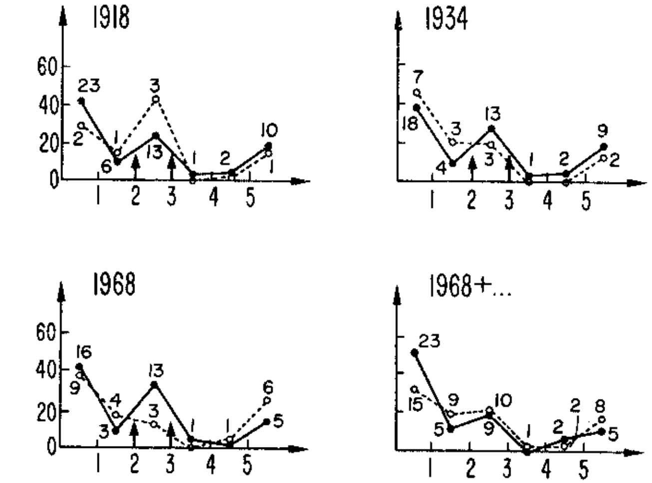


Fig. 2. For caption see p. 245.

16

18. n_4



19. n_5

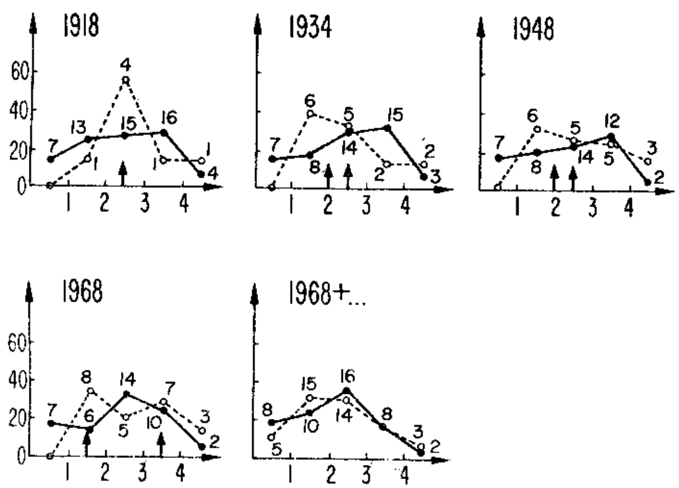


Fig. 2. For caption see p. 245.

17

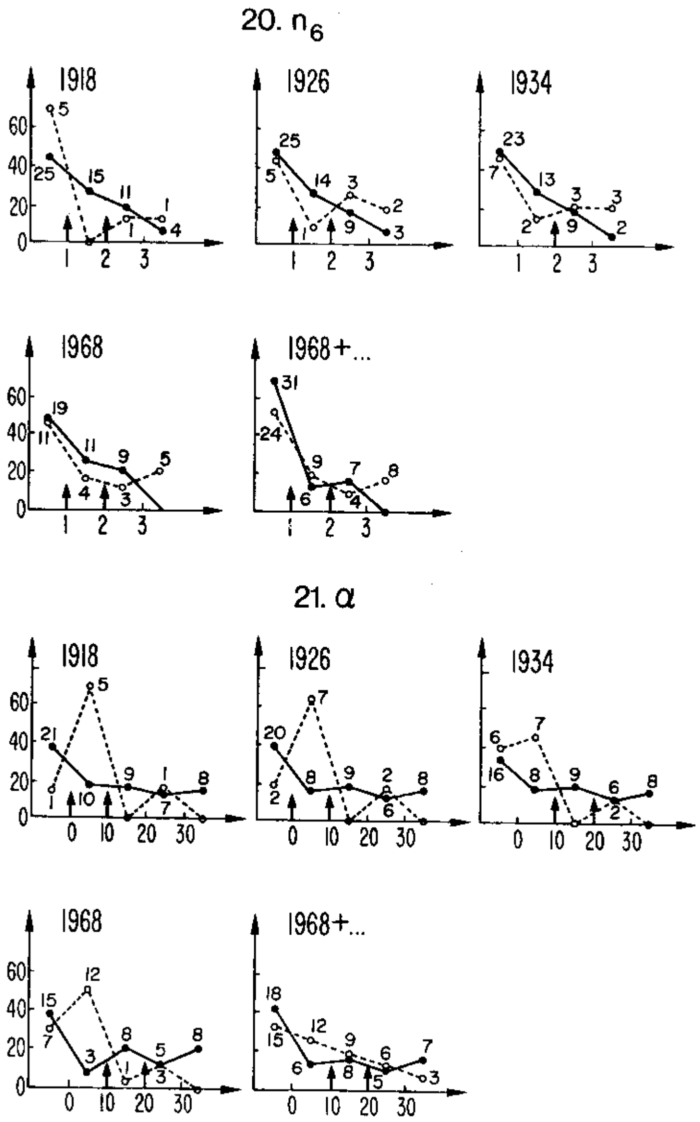


Fig. 2. For caption see p. 245.

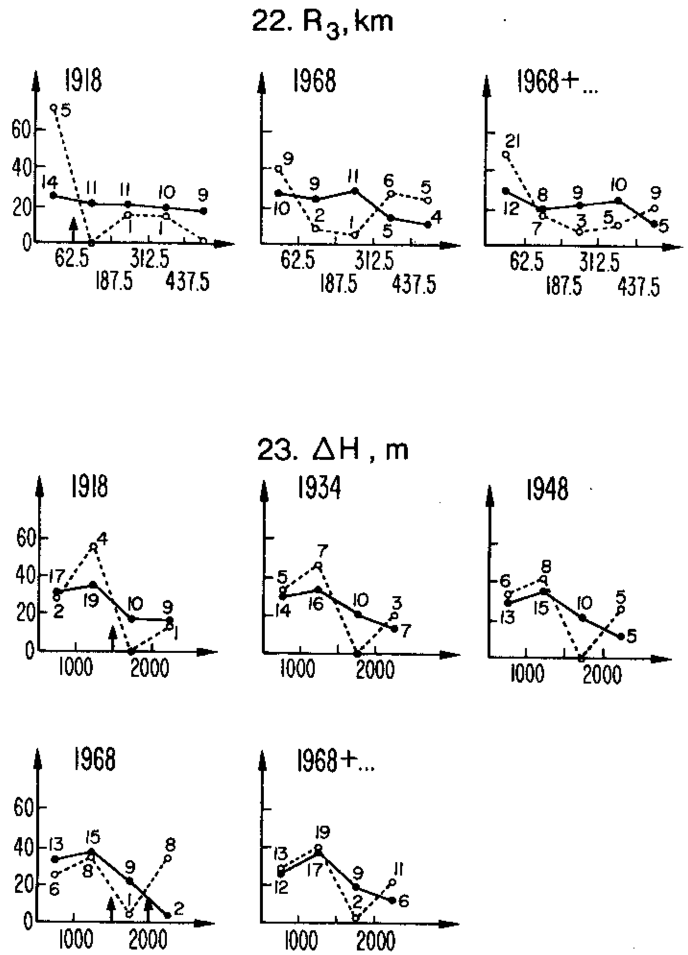


Fig. 2. One-dimensional distributions of parameters for points on major strike-slip faults of California shown in Fig. 1. 1968 = Last year in earthquake catalog; earthquakes up to this year inclusive are used in the separation of objects into classes; other dates are used to denote terminal years of EH experiment; 1968+... refers to the control experiment EF with objects recognized as D, after learning on full catalog, assigned to class I; h_{min}, etc. are the ordinal number and symbol for parameter as used in Table VI. The values of parameters are indicated on the horizontal axis; the dashed line is the histogram for points of the first kind (near known epicenters); the solid line is the histogram for points of the third kind (far from known epicenters); the assigned thresholds are shown by arrows; the number next to each point refers to the number of objects for which the values of parameters fall into the given interval shown on the abscissa; the ordinate is the same number expressed as a percentage of the total number of objects in the corresponding class.

TABLE VI

Parameters used for recognition

N	Number in Table V	Name of parameter	Symbol	Assigned threshold
1	1	maximal elevation	h_{max}	≤ 500 ≤1,250
2	2	minimal elevation	h_{min}	≤ 50 ≤ 500
3	3	elevation difference	Δh	≤ 500 ≤1,000
4	5	"gradient"	$\Delta h/l_1$	≤ 30 ≤ 70
5	6	relative area of soft sediments	q	≤ 10 ≤ 50
6	7	type of rocks	..	presence of I
7	8	distance to closest fault	r_1	≤ 12.5 ≤ 37.5
8	11	distance to closest end or intersection of faults	r_2	≤ 12.5 ≤ 37.5
9	12	distance to geothermal zone	r_3	≤ 25 ≤ 75
10	15	distance to the closest zone of divergence	r_4	≤ 100 ≤ 200
11	17	distance to the reference point (intersection of San Andreas and Big Pine faults)	r_5	≤ 125 ≤ 375
12	14	distance to closest water reservoir	r_6	≤ 0 ≤ 25
13	27	the number of unnamed faults on *Tectonic Map of U.S.* (USGS, 1962)	n_1	≤ 3 ≤ 5
14	20	the number of changes in types of relief	n_2	= 0
15	22	maximal elevation	H_{max}	≤ 500 ≤1.500
16	23	minimal elevation	H_{min}	≤ 0 ≤ 200
17	30	the number of contacts between rocks of different age on *Geological Map of North America* (USGS, 1965)	n_3	≤ 8
18	31	the number of parallel faults	n_4	≤ 2 ≤ 3
19	32	the number of faults	n_5	≤ 1 ≤ 3
20	33	the number of ends and intersections	n_6	≤ 1 ≤ 2
21	35	the angle between the fault and the dominant structural trend in the region (°)		≤ 10 ≤ 20
22*	13	distance to a region of large precipitation	R_3	≤ 62.5
23*	24	elevation differences	ΔH	≤1,500

The rule for coding the description of a point is the following: "1" means that the inequality in the last column is fulfilled.
* These parameters were used in control experiments only with learning based on earthquakes prior to 1934.

TABLE VII

Distinctive D and N features

(A) Based on entire earthquake catalog for points located on faults:

Feature number	Maximum elevation, h_{max} (meters)	Gradient $\Delta h/l_1$	Relative area of soft sediments, q %	Type of rocks	Distance to closest end or intersection of faults, r_2, km	Distance to nearest geothermal zone, r_3, km	Distance to nearest body of water, r_6, km	Minimum elevation, H_{min} (meters)	Number of contacts between rocks of different age on geologic map	Number of faults n_5	Angle between fault and dominant regional strike (deg.)
	1	4	5	6	8	9	12	16	17	19	21
D											
1					≤37.5	≤75					≤10
2				no I	≤37.5						≤10
3				no I	≤37.5					≤3	
4			>10		≤12.5					>1	
5					≤37.5			≤200	≤8		
6			>10		≤37.5			≤0			
7				no I	≤37.5	>25					
8			≤50		≤12.5	>25					
N											
1							>20		>8	≤3	
2					>37.5			≤0		≤3	
3				I		>25				≤3	
4	>500			I						≤3	
5					>12.5		>20		>8		
6					>12.5	>75			>8		
7				I		>75			>8		
8					>37.5	>25		≤200			
9			>10		>37.5			≤200			
10	>500				>37.5			≤200			
11	>500		>10		>12.5						

(B) Based on earthquake catalog extending through 1922:

Feature number	Maximum elevation, h_{max} (meters)	Minimum elevation, h_{min} (meters)	Elevation difference, Δh, (meters)	Gradient, $\Delta h/l_1$	Type of rocks	Distance to nearest fault, r_1, km	Distance to closest end or intersection of faults, r_2, km	Number of contacts between rocks of different age on geologic map	Distance to the zone of large precitation, r_3, km	Elevation difference ΔH, (meters)
	1	2	3	4	6	7	8	17	22	23
D										
1	≤1250				SM					≤1500
2			≤1000				>37.5	>8		
3			≤1000			>12.5		>8		
4	≤1250					≤12.5		>8		
5	≤1250			>30				>8		
N										
1	>500							≤8	≤62.5	
2			>500					≤8		
3	>500				no SM			≤8		
4	>500				no SM		>37.5			
5	>1250					>12.5				

Parameter number corresponds to Table VI

The points recognized as D with this threshold and $\tau = 1$ are shown in Fig. 3a. It is encouraging that these points do not occur erratically, but form a limited number of compact segments. This might be expected however, because of the spatial continuity of some of the parameters. We see from Fig. 3a that significant parts of major faults are recognized as N. These should be segments where epicenters will not occur, though faulting could extend into these regions from epicenters on adjacent segments of the fault. Among the N segments the major branch of the San Andreas fault north of San Francisco, between the points *62* and *26*, seems especially surprising as does the Garlock fault. D segments concentrate at five areas from north to south between points *61–25* (lat. 41–40°N), *27–7* (38.5–37°N), *35–43* (36–34.5°N), *45–97* (34.2–33°N), and *18–21* (33–32.2°N). We shall discuss these areas later.

3.5. Control experiments

3.5.1. Experiment EH

In conducting this experiment the histograms of parameters are redetermined (see Fig. 2) because cataloged earthquakes after a specific year are not used in learning; their corresponding clusters are disbanded and their points are placed in class II. This may lead to the selection of new thresholds and even of other parameters. In practice, however, such changes will not occur at each step of the EH experiment, because the histograms are stable by definition and because the thresholds are chosen rather roughly. The histograms, corresponding to several steps, from 1918 to 1966 are compared in Fig. 2. The only changes that occur from histogram to histogram are the following: learning on the basis of earthquakes through 1918 we would assume different thresholds

TABLE VIII

Results of voting for points on major California faults – main variant

n_N \ n_D	0	1	2	3	4	5	≥6
≥7	$\tilde{70}$,$\tilde{74}$,$\tilde{80}$,$\tilde{83}$, $\tilde{89}$,$\tilde{90}$ 82,84,86,91						
6	$\tilde{69}$,$\tilde{88}$						
5	36,37,81						
4	$\tilde{65}$,$\tilde{66}$,$\tilde{73}$,$\tilde{94}$, $\tilde{95}$ 16,44, 52,54,76						
3	$\tilde{62}$,$\tilde{63}$,$\tilde{64}$,$\tilde{67}$, $\tilde{87}$ 17,22, 50,53,60						
2	9 26 $\tilde{71}$ $\tilde{92}$,$\tilde{96}$ 15,38	$\tilde{85}$	$\tilde{72}$				
1	41	51 $\tilde{68}$	$\tilde{75}$,7,33,45,46	55	6	20 8,25	
0	$\tilde{77}$, $\tilde{79}$	11 13 47 $\tilde{93}$ 42	*2 27* 30 34 *40 43* 48 49 $\tilde{100}$	*12* $\tilde{78}$	18 3 10 23 24 28 58 61	*1 19* $\tilde{98}$ *21* 31 $\tilde{99}$ *39 56* 59 *57* 29	*4* 5 14 *32* $\tilde{97}$ 35

Circled number = number of a point in a cluster; italicized number = number of a point in a cluster which has largest value of $n_D - n_N$ in its cluster; ordinary number with tilde = number of a point not in a cluster used in learning stage; ordinary number without tilde = number of a point not in a cluster, not used in learning stage.
$k_1 = 7, k_2 = 9,\ \bar{k}_1 = 4, \bar{k}_2 = 1$ and $k_3 = 0$.

for seven parameters (No.'s 6, 7, 11, 12, 16, 19 and 21) and select for recognition two more parameters (22 and 23). The changes in the values of the thresholds are indicated in Table X.

The results of recognition in the EH control experiment are shown in Table IX and in Figs. 3b and c. They are successful since by and large the results based on a partial catalog agree with those based on the full catalog (Fig. 3a). At least one point in each cluster was recognized as D before the corresponding earthquake occurred. Moreover, at least one point in each cluster was recognized as D at all steps since 1911 with three exceptions: clusters corresponding to the earthquakes of 1923, 1940 and 1942 were not recognized with learning through 1918 and 1923.
However, at least one object in these clusters was recognized as D before the actual earthquake for all other steps in the EH experiment. The recognition of the objects in classes II and III also appears to be stable.

3.5.2. Experiment EF

The results shown in Table IX also illustrate the stability of our procedures. The identification of only seven points changed. Four D_{II} points are reclassified as N_{II} and three N_{II} points are reclassified as D_{II}. Thus, our D objects will neither wander nor proliferate over all of California, but strong earthquakes are predicted for the limited number of D_{II} points identified.

3.5.3. Inverse EH experiment

In this experiment we eliminated earthquakes from learning in the order of their occurrence, the first to occur, the first eliminated, beginning with 1836. In this and the following experiments we did not redetermine the histograms of parameters although we verified the stability of recognition by varying the learning materials.

The results of this experiment are shown in Table XI. Recognition remains successful until the 1906 earthquake. The points near its epicenter are not

TABLE IX

Recognition results for control experiments EH and EF

Number of point on Fig. 1	Year of earthquake which defines cluster	Learning based on earthquakes from 1836 through the year								
		1911	1918	1922	1923	1934	1940	1942	1948	1948+...
		Clusters $\{D_1 + N_1\}$:								
14	1948	+	+	+	+	+	+	+	+	+
49		+	+	+	+	o	o	o	+	+
51		o	+	+	o	+	o	o	+	+
18	1942	+	+	+	o	+	+	+	+	+
55		+	o	o	o	o	o	+	+	+
56		+	+	+	o	+	+	+	+	+
19	1940	+	+	+	o	+	+	+	+	+
56		+	+	+	o	+	+	+	+	+
57		+	+	+	o	+	+	+	+	+
58		+	+	+	o	+	+	+	+	+
21	1934	+	o	+	o	+	+	+	+	+
20		o	+	+	+	+	+	+	+	+
23	1923	o	o	+	+	+	+	+	+	+
1		+	o	+	+	+	+	+	+	+
24		o	o	+	+	+	+	+	+	+
61		o	o	+	+	+	+	+	+	+
10	1922	o	o	o	+	+	+	+	+	+
9		+	o	+	o	o	o	o	o	o
39		o	+	+	+	+	+	+	+	+
13	1918	o	+	+	+	+	+	+	+	+
12		+	+	+	+	+	+	+	+	+
47		+	+	+	+	+	+	+	+	+
6	1911	+	+	+	+	+	+	+	+	+
31		+	+	+	+	+	+	+	+	+
32		+	+	+	+	+	+	+	+	+
34		+	+	+	+	+	o	+	+	+
2	1906	+	+	+	+	+	+	+	+	+
26		+	+	+	o	+	o	o	o	o
27		+	o	+	+	+	+	+	+	+
11	1857	+	+	+	o	+	o	o	+	+
40		+	+	+	+	+	+	+	+	+
43		o	+	+	+	+	+	+	+	+
3	1836	+	+	+	+	+	+	+	+	+
28		+	+	+	+	+	+	+	+	+
4		+	+	+	+	+	+	+	+	+

TABLE IX (*continued*)

Number of point on Fig. 1	Year of earthquake which defines cluster	Learning based on earthquakes from 1836 through the year								
		1911	1918	1922	1923	1934	1940	1942	1948	1948+...
4	1836	+	+	+	+	+	+	+	+	+
3		+	+	+	+	+	+	+	+	+
5		+	+	+	+	+	+	+	+	+
30		+	+	+	+	+	+	+	+	+
5	1836	+	+	+	+	+	+	+	+	+
4		+	+	+	+	+	+	+	+	+
30		+	+	+	+	+	+	+	+	+
32		+	+	+	+	+	+	+	+	+
					Class II, $\{D_2 + N_2\}$:					
62		+	o	+	o	o	o	o	o	o
63		o	o	+	o	o	o	o	o	o
64		o	o	o	o	o	o	o	o	o
65		o	o	o	o	o	o	o	o	o
66		+	o	+	o	o	o	o	o	o
67		+	o	+	o	o	o	o	o	o
68		+	+	+	o	o	o	+	+	+
69		o	o	o	o	o	o	o	o	o
70		o	o	o	o	o	o	o	o	o
71		o	o	o	o	o	o	o	o	o
72		o	o	o	+	o	o	o	+	o
73		o	o	o	o	o	o	+	o	o
74		+	o	+	o	o	o	o	o	o
75		+	o	o	o	o	+	+	+	+
77		+	o	o	o	+	o	+	+	+
78		+	o	o	o	o	o	+	+	+
79		o	+	+	o	+	+	+	+	o
80		o	o	o	o	o	o	o	o	o
83		o	o	o	o	o	o	o	o	o
85		o	+	o	o	o	o	o	o	+
87		o	+	o	o	o	o	o	o	o
88		o	+	o	o	o	o	o	o	o
89		o	+	+	o	o	o	o	o	o
90		o	o	o	o	o	o	o	o	o
91		o	o	o	o	o	o	o	o	o
92		+	+	+	o	o	o	o	o	o
93		+	+	+	+	+	o	o	+	+
94		o	o	o	o	o	o	o	o	o
95		o	o	o	o	o	o	o	o	o
96		o	o	o	o	o	o	o	o	o
97		o	o	+	+	+	+	+	+	+
98		o	+	+	+	+	+	+	+	+
99		+	o	+	+	+	+	+	+	+
100		+	o	o	o	o	o	o	+	+

TABLE IX (continued)

Number of point on Fig. 1.	Year of earthquake which defines cluster	Learning based on earthquakes from 1836 through the year								
		1911	1918	1922	1923	1934	1940	1942	1948	1948+...
					Class III (objects for voting but not for learning):					
7		o	+	+	o	+	+	+	+	+
8		+	o	o	+	+	+	+	+	+
15		o	+	o	o	o	o	o	o	o
16		o	+	o	o	o	o	o	o	o
17		o	+	+	o	o	o	o	o	o
22		o	o	+	o	o	o	o	o	o
25		o	o	o	+	+	+	+	+	+
29		+	+	+	+	+	+	+	+	+
33		+	+	+	+	o	o	+	+	+
35		+	o	o	+	+	+	+	+	o
36		+	o	+	o	o	o	o	o	o
37		+	+	o	o	o	o	o	o	o
38		+	o	o	o	o	o	o	o	o
41		o	+	o	o	o	o	o	+	o
42		+	+	+	+	+	+	+	+	o
44		o	+	+	o	o	o	o	o	o
45		+	+	+	o	o	o	o	+	+
46		+	+	+	o	o	o	o	+	+
48		+	+	+	+	+	+	+	+	+
50		o	+	+	o	o	o	o	o	o
52		o	o	o	o	o	o	o	o	o
53		o	o	o	o	o	o	o	o	o
54		o	o	o	o	o	o	o	o	o
59		+	+	+	+	+	+	+	+	+
60		o	o	+	o	o	o	o	o	o
76		o	o	o	o	o	o	o	o	o
81		o	o	o	o	o	o	o	o	o
82		o	o	o	o	o	o	o	o	o
84		o	o	o	o	o	o	o	o	o
86		o	o	o	o	o	o	o	o	o

+ = object recognized as D; o = object recognized as N; year refers to the data through which earthquakes in the catalog were used for learning; last earthquake used was in 1948; last column (1948+...) refers to experiment EF; blank lines separate the clusters, each identified by the year of the earthquake associated with the cluster; vertical lines show the time when the cluster was included in learning; to the left of this line the points of this cluster were included in class II during the learning stage.

recognized as D_{II}, after we transfer them in the learning stage to class II. However, after we also eliminate the 1911 epicenter, a D point is again recognized in each cluster. Thus, on the basis of learning on seven earthquakes (1918–1948) we recognized as dangerous the sites of the earthquakes for the period 1836–1911, except for 1906.

3.5.4. *Uniformity of the region*

Several areas were selected in Table IX, each corresponding either to a group of epicenters or to a part of a group. The epicenters in these areas were assumed to be unknown and the corresponding clusters were eliminated from learning. The results, shown in Table XII, indicate that recognition is reasonably

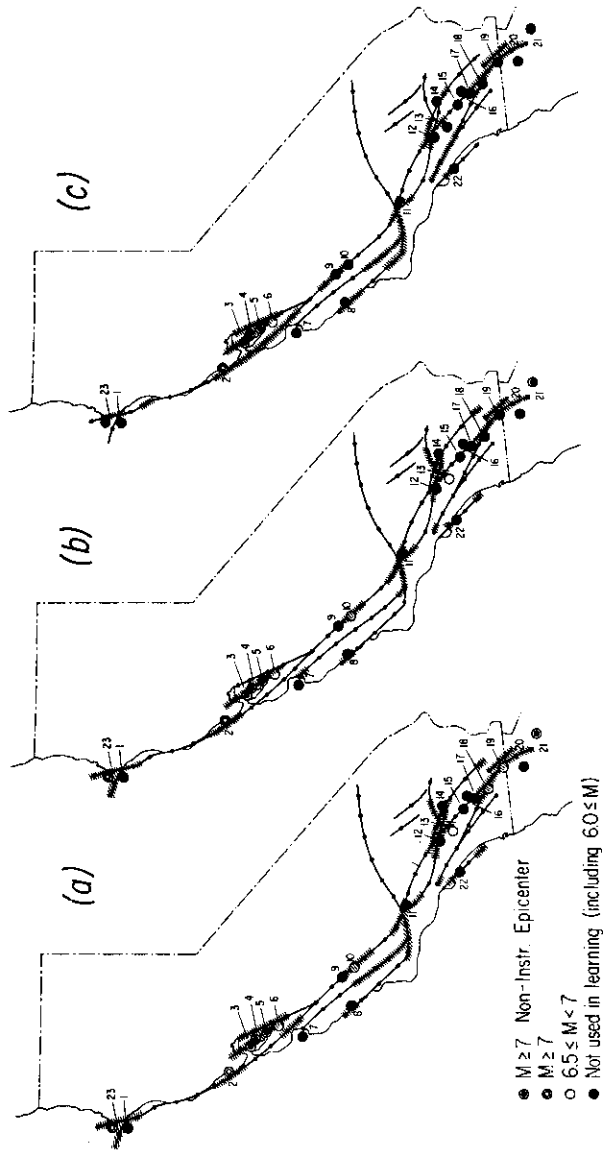


Fig. 3. Results of recognition for strike-slip earthquakes in California. Hachured regions of faults are recognized as dangerous. All earthquakes except those shown by solid circles were used in learning.
(a) All earthquakes $M \geq 6.5$ used in learning except those shown as solid.
(b) All earthquakes $M \geq 6.5$, prior to 1934 are used in learning.
(c) All earthquakes $M \geq 6.5$, prior to 1911 are used in learning.

TABLE X

Change in thresholds of discretization in experiment EH

Number in Table VI	The name of parameter	Thresholds for learning through the year 1968	1918
6	type of rocks	I	not SM
7	distance to closest fault, r_1 (km)	$\leq$ 12.5	$\leq$ 12.5
		$\leq$ 37.5	$\leq$ 25
11	distance to the reference point (intersection of San Andreas and Big Pine faults), r_5 (km)	$\leq$125	$\leq$ 250
		$\leq$375	$\leq$ 375
12	distance to closest water reservoir, r_6 (km)	= 0	
		$\leq$ 25	$\leq$ 25
16	minimal elevation, H_{min} (m)	$\leq$ 0	$\leq$ 0
		$\leq$200	
19	the number of faults	$\leq$ 1	$\leq$ 2
		$\leq$ 3	
21	the angle between fault and dominant strike in the region, α (°)	$\leq$ 10	= 0
		$\leq$ 20	$\leq$ 10
22*	distance to a region of large precipitation, R_3	–	$\leq$ 62.5
23*	elevation difference, ΔH (m)	–	$\leq$1,500

* Parameter used only in learning with earthquakes prior to 1934.

stable to variations in learning material. They also indicate how homogeneously known earthquakes should cover a D area for successful recognition. At least one point near each eliminated epicenter was recognized as D in variants 1, 5, 7 and 8. These variants correspond to elimination of earthquakes in the southern and northern parts of the regions under study, namely, the Mendocino fault (variant 1), the intersection of the Garlock and San Andreas faults (variant 5), the San Jacinto Mountains (variant 7), and the Salton Sea depression (variant 8). With the other variants, 2, 3, 4 and 6, where earthquakes in the central area between San Francisco and Parkfield were eliminated, the corresponding D objects were not recognized.

3.5.5. Variations of the set of parameters

In this control experiment we eliminated each parameter (one at a time) from the data set and repeated the learning and voting. The results are shown in Table XIII. The order, in which the parameters are listed, corresponds to their increasing role in recognition. We see that r_2, the distance to a fault end or an intersection of faults, is most important. The importance of r_2 is confirmed in other ways. For example, to get a decent number of distinctive features with r_2 eliminated, we had to relax k_1 to 7 and $\bar{k}_2$ to 2. The objects separate poorly in the voting stage, and the results of recognition are unstable with respect to changes in the threshold $\bar{\Delta}$. The reason for the importance of this parameter is discussed in a later paragraph.

Recognition is stable with respect to the elimination of the other parameters, as Table XIII shows.

3.5.6. Variation of the algorithm

In this control experiment we use the algorithm CORA-3 instead of CLUSTERS. In this case class I is composed of the centers of the clusters; class II is unchanged. The results are very similar to those discussed before. This experiment confirms that a somewhat different algorithm for our recognition procedures leads to similar results. At the same time it confirms the advantage of the algorithm CLUSTERS for our particular case because the results obtained by CORA are much less stable to variations in the learning material, and especially to changes in the threshold $\bar{\Delta}$.

26 27

TABLE XI

Recognition results for inverse EH experiment (see Table IX for notation)

Number of point on Fig. 1	Year of earthquake which defines cluster	Year of first earthquake allowed for in learning						
		1836	1836	1836	1857	1906	1911	1918
14		+	+	+	+	+	+	+
49	1948	+	+	+	+	o	o	+
51		+	+	o	+	o	+	+
18		+	+	+	+	+	+	+
55	1942	+	+	+	o	+	+	+
56		+	+	+	+	+	+	+
19		+	+	+	+	+	+	+
56	1940	+	+	+	+	+	+	+
57		+	+	+	+	+	+	+
58		+	+	+	+	+	+	+
21	1934	+	+	+	+	+	+	+
20		+	+	+	+	+	+	+
23		+	+	+	+	+	+	+
1	1923	+	+	+	+	+	+	+
24		+	+	+	+	+	+	+
61		+	+	+	+	+	+	+
10		+	+	+	+	+	+	+
9	1922	o	o	o	o	+	+	+
39		+	+	+	+	+	+	+
13		+	+	+	+	+	o	+
12	1918	+	+	+	+	+	+	+
47		+	+	+	+	+	+	+
6		+	+	+	+	+	+	+
31	1911	+	+	+	+	+	+	+
32		+	+	+	+	+	+	+
34		+	+	+	+	+	+	+
2		+	+	+	+	+	o	o
26	1906	o	o	o	o	+	o	+
27		+	+	o	+	+	o	o
11		+	o	+	+	o	o	+
40	1857	+	+	+	+	+	+	+
43		+	+	+	+	+	+	+
3	1836	+	+	+	o	+	o	+
28		+	+	+	+	+	o	+
4		+	+	+	+	+	+	+
3		+	+	+	o	+	o	+
5	1836	+	+	+	+	+	+	+
30		+	+	o	o	o	o	+

TABLE XI (*continued*)

Number of point on Fig. 1	Year of earthquake which defines cluster	Year of first earthquake allowed for in learning						
		1836	1836	1836	1857	1906	1911	1918
5		+	+	+	+	+	+	+
4	1836	+	+	+	+	+	+	+
30		+	+	o	o	o	o	+
32		+	+	+	+	+	+	+
62		o	o	o	o	o	+	+
63		o	o	o	o	o	+	+
64		o	o	o	o	o	o	+
65		o	o	o	o	o	o	+
67		o	o	o	o	o	o	+
68		+	+	+	+	o	+	+
69		o	o	o	o	o	o	+
71		o	o	o	o	+	+	+
72		+	+	o	o	o	+	+
74		o	o	o	o	o	o	+
75		+	+	+	+	+	o	+
77		+	+	+	+	+	+	+
78		+	+	+	+	+	+	+
79		+	+	+	+	+	o	+
85		o	o	o	o	o	o	+
87		o	o	o	o	o	o	+
88		o	o	o	o	o	o	+
90		o	o	o	o	o	o	+
91		o	o	o	o	o	o	+
92		o	o	o	o	+	o	+
93		+	+	+	+	o	o	+
94		o	o	o	o	o	o	+
95		o	o	o	o	o	o	+
96		o	o	o	o	o	o	+
97		+	+	+	+	+	+	+
98		+	+	+	+	+	+	+
99		+	+	+	+	+	+	+
100		+	+	+	+	o	o	+
7		+	+	+	+	+	+	+
8		+	+	+	+	+	+	+
15		o	o	o	o	o	o	+
16		o	o	o	o	o	o	+
17		o	o	o	o	o	o	+
22		o	o	o	o	+	+	+
25		+	+	+	+	+	+	+
29		+	+	+	+	+	+	+
33		+	+	+	o	o	o	+
35		+	+	+	+	+	+	+
36		o	o	o	o	o	o	+
37		o	o	o	o	o	+	+
38		o	o	o	o	+	+	+
41		o	o	o	o	o	o	+
42		+	+	+	+	+	+	+

TABLE XI (*continued*)

Number of point on Fig. 1	Year of earthquake which defines cluster	Year of first earthquake allowed for in learning						
		1836	1836	1836	1857	1906	1911	1918
44		o	o	o	o	o	o	+
45		+	+	o	+	+	+	+
46		+	+	o	+	+	+	+
48		+	+	+	+	+	+	+
50		o	o	o	o	o	o	+
52		o	o	o	o	o	o	+
53		o	o	o	o	o	o	+
54		o	o	o	o	o	o	+
59		+	+	+	+	+	+	+
60		o	o	o	o	+	+	+
76		o	o	o	o	o	+	+
84		o	o	o	o	o	o	+

The earthquakes are eliminated cumulatively from learning in the order of their occurrence: first one, then two, etc. Year in top row refers to the first earthquake allowed for in learning. The rest of the notation is analogous to Table IX.

TABLE XII

Recognition in the control experiment "uniformity of the region" (elimination of groups of clusters from learning). Parentheses indicate the eliminated clusters. Otherwise the notation is the same as in Table IX

No. of point in Fig. 1	Number of variant 1	2	3	4	5	6	7	8
Clusters:								
14	+	+	+	+	+	+	(+)	(+)
49	+	+	+	o	+	+	(+)	(+)
51	+	o	o	o	o	+	(+)	(+)
18	+	+	+	+	+	+	+	(+)
55	o	o	o	+	+	+	+	(+)
56	+	+	+	+	+	+	+	(+)
19	+	+	+	+	+	+	+	(+)
56	+	+	+	+	+	+	+	(+)
57	+	+	+	+	+	+	+	(+)
58	+	+	+	+	+	+	+	(+)
21	+	+	+	+	+	+	+	(+)
20	+	+	+	+	+	+	+	(+)
23	(o)	+	+	+	+	+	+	+
1	(+)	+	+	+	+	+	+	+
24	(o)	+	+	+	+	+	+	+
61	(o)	+	+	+	+	+	+	+

No. of point in Fig. 1	Number of variant 1	2	3	4	5	6	7	8
10	+	+	+	+	+	(o)	+	+
9	o	o	o	+	o	(o)	o	+
39	+	+	+	+	+	(o)	+	+
13	+	o	o	+	o	+	(+)	+
12	+	+	+	+	+	+	(+)	+
47	+	+	+	+	+	+	(+)	+
6	+	+	+	(o)	+	o	+	+
31	+	+	+	(o)	+	+	+	+
32	+	+	+	(o)	+	+	+	+
34	+	+	+	(o)	+	+	+	+
2	(+)	(o)	(o)	(o)	+	+	+	+
26	(o)	(o)	(o)	(o)	o	o	+	+
27	(+)	(o)	(o)	(o)	+	+	+	+
11	+	+	+	+	(o)	+	+	+
40	+	+	+	+	(+)	+	+	+

TABLE XII (*continued*)

No. of point in Fig. 1	Number of variant 1	2	3	4	5	6	7	8
43	+	o	o	+	(+)	+	+	+
3	+	+	(o)	(o)	+	+	+	+
28	+	o	(o)	(o)	+	+	+	+
4	+	+	(+)	(+)	+	+	+	+
4	+	+	(+)	(+)	+	+	+	+
3	+	+	(o)	(o)	+	+	+	+
5	+	+	(+)	(o)	+	+	+	+
30	+	+	(o)	(o)	+	+	+	+
5	+	+	(+)	(o)	+	+	+	+
4	+	+	(+)	(+)	+	+	+	+
30	+	+	(o)	(o)	+	+	+	+
32	+	+	(+)	(o)	+	+	+	+
Group II:								
62	o	o	o	+	o	o	o	o
63	o	o	o	+	o	o	o	o
64	o	o	o	+	o	o	o	o
65	o	o	o	o	o	o	o	o
66	o	o	o	o	o	o	o	o
67	o	o	o	+	o	o	o	o
68	+	o	o	o	o	+	+	+
69	o	o	o	o	o	o	o	o
70	o	o	o	o	o	o	o	o
71	o	o	o	+	o	o	o	+
72	o	o	o	o	+	o	o	o
73	o	o	o	o	o	o	o	o
74	o	o	o	o	o	o	o	o
75	+	o	o	o	+	+	+	o
77	+	o	o	+	+	+	+	o
78	+	+	+	+	+	+	+	o
79	+	o	o	+	o	+	o	o
80	o	o	o	o	o	o	o	o
83	o	o	o	o	o	o	o	o
85	o	o	o	o	o	o	o	o
87	o	o	o	o	o	o	o	o
88	o	o	o	o	o	o	o	o
89	o	o	o	o	o	o	o	o
90	o	o	o	o	o	o	o	o
91	o	o	o	o	o	o	o	o
92	o	o	o	+	o	o	o	o
93	+	+	o	o	o	+	+	+
94	o	o	o	o	o	o	o	o
95	o	o	o	o	o	o	o	o
96	o	o	o	o	o	o	o	o
97	+	+	+	+	+	+	+	+
98	+	+	+	+	+	+	+	+
99	+	+	+	o	+	+	+	+
100	+	o	o	o	+	+	+	+
Group III:								
7	+	o	o	o	+	+	+	+
8	+	+	o	+	+	+	+	+
15	o	o	o	o	o	o	o	o
16	o	o	o	o	o	o	o	o
17	o	o	o	o	o	o	o	o
22	o	o	o	+	o	o	o	o
25	+	+	+	+	+	+	+	+
29	+	+	+	o	+	+	+	+
33	+	o	o	o	+	o	+	+
35	+	+	+	+	+	+	+	+
36	o	o	o	+	o	o	o	o
37	o	o	o	o	o	o	o	o
38	o	o	o	+	o	o	o	+
41	o	o	o	o	o	o	o	o
42	+	+	+	+	+	+	+	+
44	o	o	o	o	o	o	o	o
45	+	+	+	o	+	+	+	+
46	+	+	+	o	+	+	+	+
48	+	+	o	+	+	+	+	+
50	o	o	o	o	o	o	o	o
52	o	o	o	+	o	o	o	o
53	o	o	o	o	o	o	o	o
54	o	o	o	o	o	o	o	o
59	+	+	+	+	+	+	+	+
60	o	o	o	+	o	o	o	o
76	o	o	o	o	o	o	o	o
81	o	o	o	o	o	o	o	o
82	o	o	o	o	o	o	o	o
84	o	o	o	o	o	o	o	o
86	o	o	o	o	o	o	o	o

3.5.7. Recognition of randomly described objects

The binary codes, which describe points, are replaced by random binary numbers in this control experiment. Learning proceeds using the algorithm CLUSTERS. In a first step we use the same k_i and $\tilde{k}_i$ as with the real data described earlier. In this case 32 distinctive D features and no N features are found, an unsatisfactory result. This is not sufficient to

TABLE XIII

Control experiment: change of recognition due to the elimination of parameters

After elimination of the parameter	Recognition of the following points changed: D changed to N	N changed to D	Class in learning
h_{max}, r_6, n_3, or n_5 (only one at a time)		*85* or *96*	II
α	*72*		II
		41	III
r_3		*87* and *96*	II
		41	III
type of rocks	*72*	*96*	II
	25 and *35*		III
q	*51*		I
	72	*85* and *92*	II
	45 and *46*		III
H_{min}	*51*	*26*	I
	72		II
		22 and *60*	III
r_2	*49* and *6*	*9* and *26*	I
	68, 77, 79 and *100*	*66, 67, 71, 87, 62, 63, 64, 65* and *92*	II
	45 and *46*	*17, 22, 36, 38, 41, 52, 60* and *76*	III

reject the possibility of recognition with random data, because the values of k_i and $\bar{k}_i$ in the processing of real data are not predetermined; they are varied until sufficient and comparable numbers of distinctive D and N features are obtained. The only limitation is that k_i should be sufficiently large and $\bar{k}_i$ not too large. Therefore, k_i and $\bar{k}_i$ should be varied in this control experiment. With $k_1 = 12, k_2 = 9, \bar{k}_1 = 5, \bar{k}_2 = 2$ we find twelve distinctive features of D and fifteen of N. However, the separation of points by voting is rather vague, as is seen by comparing Tables VIII and XIV. Each feature occurs in fewer points; for example, each distinctive D feature is found at an average of fourteen points, whereas for real data the corresponding number is 21.

Also, the results are unstable, in that experiment EF results in twelve new D points in contrast to three points for real data. Furthermore the D segments of major faults are more scattered.

In this way the experiment with random data supports the results found with real data. We also learn that reliable results show up in a high degree of separation of the objects by voting.

3.6. Geological implications

To understand the geological implications of this work we pose the question: what makes a place D and not N? To answer this question, we review the parameters which make up the D and N features. We must review not only distinctive features, but also their equivalent and ϵ-equivalent features which were identified but not used in voting. Let us recall the definition of these features. Consider two features of some class F_1 and F_2. The sets of objects of the same class, which have F_1 or F_2, are called S_1 or S_2, respectively. F_1 and F_2 are equivalent, if S_1 and S_2 coincide. They are ϵ-equivalent, if only the ϵth fractions of the

TABLE XIV

Results of voting after learning on random data

n_N								
≥7	$\widetilde{64}, \widetilde{75}, \widetilde{79}$ $\widetilde{87}, \widetilde{95}, \widetilde{98}$ 42							
6	15							
5	36	$\widetilde{62}$, 76	$\widetilde{88}$	7				
4	$\widetilde{73}, \widetilde{91}$	$\widetilde{69}$	$\widetilde{67}$, 33, 35		47			
3	$\widetilde{78}, \widetilde{85}$ $\widetilde{92}$, 60	39 $\widetilde{72}$ 16, 29, 44	26 52	$\widetilde{77}, \widetilde{90}$	84			
2	$\widetilde{63}, \widetilde{66}$ $\widetilde{68}$	$\widetilde{71}$, 46 54	43 59 81	57	9 $\widetilde{97}$	$\widetilde{70}$, 8	6	
1	40	19 $\widetilde{74}$ 22, 38	1 $\widetilde{83}$ $\widetilde{100}$, 25, 37, 50	55 12 27 $\widetilde{93}, \widetilde{94}$	18 61 31 45		51 $\widetilde{99}$	
0		58 13 3 5 48, 86 $\widetilde{96}, \widetilde{89}$	4 $\widetilde{80}$ 82	23 53	14 21 *10* 17, 41	*32 34*	*56 20 24*	*49 2 11 28 30* 65
	0	1	2	3	4	5	6	≥7
	n_D							

$k_1 = 12, k_2 = 7, \bar{k}_1 = 5$ and $\bar{k}_2 = 2$ (k_i and $\bar{k}_i$ are relaxed compared to main variant).

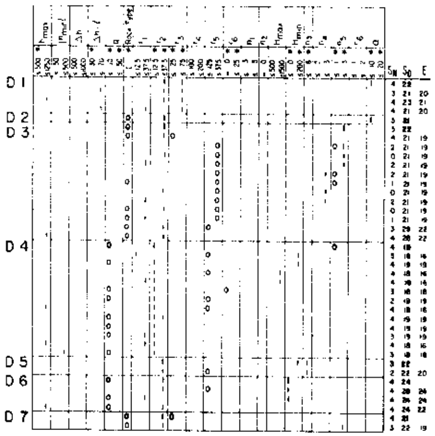
objects is S_1 and S_2 do not coincide (the fraction is relative to the total number of objects in S_1 and S_2; each object is counted once). Equivalent and ϵ-equivalent features, like distinctive features, carry significant information of geologic interest. They were not used in voting in order not to bias the vote by repeated contributions by some subsets of objects.

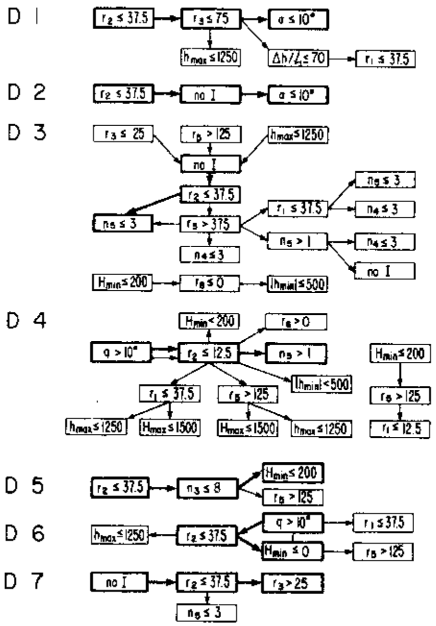
Although 21 parameters indicated in Table VI were considered in the learning stage, only eleven were selected for incorporation in distinctive features. Since we are examining features for their "geological content" we might as well broaden the number, using more parameters in the process. We do this by rerunning the algorithm, relaxing the k_i slightly. We also used objects recognized as D or N rather than points assigned to class I or II in this new learning stage aimed at finding more features. We will next discuss 177 distinctive, equivalent and ϵ-equivalent features, involving all 21 parameters of Table VI, found this way. Table XV summarizes these features.

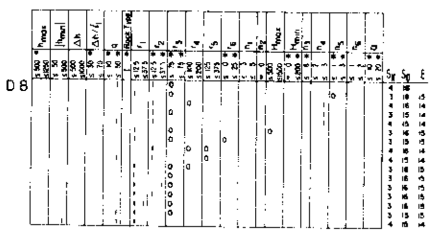
First consider the left-hand side part. Parameters are indicated in the uppermost row, with notation the same as in Tables V and VI. The second row indicates the thresholds for the parameters. All parameters used in the original learning stage (Table VI) are considered; parameters, which were selected for recognition in this earlier stage (Table VIIA) are marked by an asterisk. The features are divided into groups, separated by thin horizontal lines. Each group consists of one distinctive feature (from Table VIIA) and its equivalent and ϵ-equivalent features. A distinctive feature is indicated by a heavy number at the top of the group. The digit "1" means that the parameter must satisfy the threshold condition in row 2; "0" means that it must not. The numbers to the left of the table indicate the ordinal number of the distinctive feature in Table

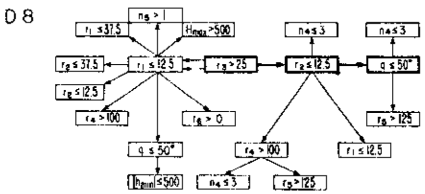
TABLE XV

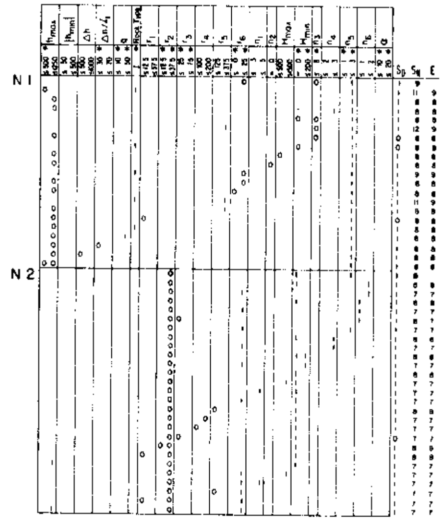
Distinctive features, examined for geological content for points located on faults. Left-hand side displayed in tabular form, right-hand side as flow diagram. (See paragraph 3.6 for explanation.)

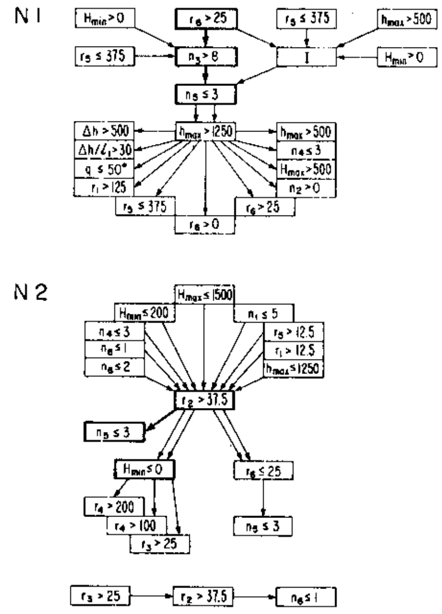


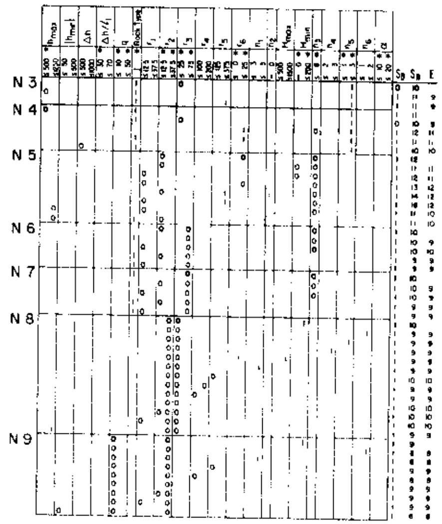












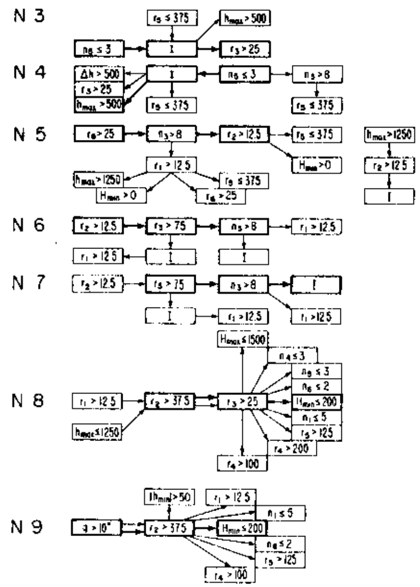
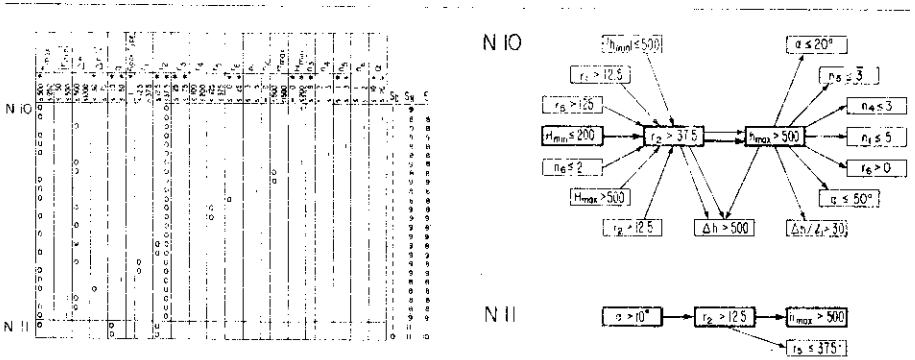


TABLE XV (*continued*)



VIIA. D and N indicate which type of feature it is. S_D and S_N to the right of the table indicate how many objects of class I and II, respectively, have this feature. E is the number of objects which have both this feature and the distinctive feature associated with it.

Now consider the right-hand side of Table XV. It shows another representation of the same groups which we call "the structure of the group". This form of representation becomes clear after comparison with the left-hand side of Table XV. Any group of three consecutive rectangles which are adjacent to one another in the direction of the arrows corresponds to some feature made up of a triplet of statements, which occurs on the left-hand side of Table XV (the direction of arrows can be reversed of course, but only simultaneously for all arrows). Rectangles, connected by double arrows, correspond to a feature consisting of a doublet of statements; the second arrow should be ignored in determining the triplet features. Distinctive features are indicated by heavy lines for the two arrows and three rectangles involved. We can now proceed to the geological interpretation of the features summarized in Table XV.

The distance to the closest intersection of faults, or to the closest end of a fault (r_2) is a dominant parameter for the following reasons:

(1) In 6 out of 8 D-groups all features, except one, contain the condition that r_2 is smaller than the threshold; in the two remaining exceptions, D3 and D8, half of the features has this condition.

(2) In 8 out of 11 N-groups at least two features have the condition that r_2 is larger than the threshold.

(3) In the groups D3 and D8, the exceptions mentioned in (1), all features which do not have limitations on r_2 have instead either one or both of the conditions: The distance r_1 from the closest major fault is small or the number n_5 of major faults nearby is > 1. Both of these conditions indirectly indicate proximity to an intersection.

(4) r_2 plays a dominant role in the test for stability of recognition (Table XIII).

This result (that r_2 is small for D and large for N) supports the hypothesis (Gelfand et al., 1972, 1973a, b, 1974a, b) that epicenters of strong earthquakes tend to be situated near the intersections of major lineaments. Because of this support for the hypothesis we are encouraged to apply it to California in Section 4.

The distance r_2 from the end of a fault or an intersection of major faults which we have just considered, though important, is not sufficient by itself in that not all ends and intersections are found in D areas. It may be seen, for example, from Fig. 2 that the situation $r_2 < 37.5$ km occurs not only for all D points, but also for 14 out of a total 44 N points (Fig. 2, histogram No. 8, 1968+). This is why r_2 must be used only in combination with other parameters in forming features.

The statistics of some of the parameters of D and N features are summarized in Table XVI. The table does not consider r_2 and r_1, which have already been discussed in order to highlight other features. In Table XVI it is interesting to compare features of three types:

D: points which are always close to the end of or an intersection of faults
NC: N points which may be close to the end of or an intersection of faults (r_2 not specified or ≥ 12.5 km only)
NF: N points, which are far from the end of and an intersection of faults ($r_2 \geq 37.5$ km)

The difference between D and NC is the difference between points D which are certainly near intersections and such points N which may be close to intersections. The difference between D and NF is the difference between D points and points far from ends and intersections, which are always N. We note first of all that some measure of elevation is small for most D points and large for NC points. For NF points the features show both large and small elevations, signifying that no clear dependence on height is established far from intersections.

An examination of geological maps of the region suggests that most of the other characteristics of D features in Table XVI may also be interpreted as an indirect indication of low elevation. These include the absence of igneous rocks, proximity to a large area of recent sediments, and the small number of contacts (n_3) on a geological map. At first sight the last parameter behaves contrary to expectations in that large n should be associated with intensive fracturing of the crust, which is a characteristic for D

TABLE XVI

Summary of the role of different parameters (excluding r_2 and r_1) in D and N features

Parameter		In how many D		In how many NC				In how many NF	
				(with $r_2 > 12.5$ km)		no condition on r_2		(with $r_2 > 37.5$ km)	
		groups	features	groups	features	groups	features	groups	features
Elevation is (h_{max}, or h_{min}, or H_{max}, or H_{min})	small	6	17	–	–	–	–	5	30
	large	1	1	2	5	1	18	2	19
Igneous rocks (I) are	present	–	–	3	7	1	5	2	7
	absent	3	10	–	–	–	–	–	–
q is	small	1	5			1	1	1	1
	large	2	11	1	2	–	–	1	9
Large water reservoir (r) is	close	2	2	–	–	–	–	1	11
	far	–	–	1	2	1	4	1	1
α is	small	2	2	–	–	–	–	1	1
	large	–	–	–	–	–	–	–	–
Number of contacts (n_3) on geological map is	small	1	2	–	–	–	–	–	–
	large	–	–	3	13	1	4	1	1
Reference point (r_5) is	close	–	–	2	3	1	3	1	2
	far	5	18	–	–	–	–	4	6

(Gelfand et al., 1972, 1973a, b, 1974b). However, only neotectonic fracturing is characteristic of D according to Gelfand et al. (1972, 1973a, b, 1974a, b), whereas n_3 represents the integrated tectonic history. A survey of places with large and small n_3 indicates that small n_3 tend to go with subsidence and low elevations.

The small distance r_6 to a large water reservoir (parameter 12, Table VI) is another characteristic of D. This indirectly implies relatively low elevation, and recalls some modern ideas on the role of water in triggering earthquakes. However, for many points small r_6 indicates proximity to the ocean, so that these interpretations are by no means firm.

Small values of the angle α between the strike of the fault and the dominant strike of the San Andreas system is characteristic of D. This suggests the reasonable notion that the strike-slip earthquakes considered in this paper tend to occur away from bends in the San Andreas fault system. Earthquakes with dip-slip components might be more characteristic of bends in the fault.

Large distance from the reference point (defined by the intersection of the San Andreas and the Big Pine faults) is characteristic of D, perhaps for the reason that earthquakes with dip-slip tend to occur in the Transverse Ranges which are adjacent to a bend in the San Andreas fault.

Other parameters which play a role in D or N features occur. We defer offering possible explanations for them. Among these parameters are the distance r_3 to a geothermal zone, the number n_4 of parallel faults. Parameters which are present in the same way in both D and N features (for example $n_4 \leqslant 3$) are puzzling. Perhaps they were selected as features by chance, due to a random interplay of events, independent of their geological meaning. Perhaps they take on meaning only in combination with the other parameters they link up with, which are different for D and N features.

The qualitative conclusion which emerges is that D areas are characterized by proximity to the end or to an intersection of major faults in association with low relief and often with some kind of downward neotectonic movement, expressed in topography and geology. Apparently, we did not formulate an adequate single parameter for such movements, so that the indication shows indirectly through several different parameters.

Also the fact that minimal and maximal elevations are simultaneously small for many D points suggests that D points for these strike-slip earthquakes are often characterized by relative low relief or subsidence on a background of weak uplift. N points near the intersections or ends of major faults are characterized by higher elevations or uplift with less contrast in relief.

4. Intersections of major lineaments

4.1. The problem

The results of Section 3 (see paragraph 3.6) suggest the hypothesis that the epicenters of strong earthquakes in California are associated with intersections of major lineaments (Gelfand et al., 1972, 1973a, b, 1974a, b; Briggs and Press, in preparation). More specific formulation of this hypothesis as well as the definition of lineaments is given in the Appendix. The major lineaments of California and adjacent regions, based on a synthesis of published data (Atwood, 1940; Richter, 1958; USGS, no date, 1962, 1965; Wright and Frey, 1965; Cook, 1966; Hamilton and Myers, 1966; Hill, 1966; Thompson, 1966; Dickinson and Grantz, 1968; King, 1969; Hain, 1971), are shown in Fig. 4. Explanation and justification is provided in Ranzman (in preparation). The transverse lineaments in Fig. 4 may raise some doubts, since they are to a large extent based on geomorphic evidence. We believe that ERTS photographs of the region support these interpretations (Ranzman, in preparation).

The objects of recognition, according to the hypothesis, are now the intersections of the lineaments. The problem, formulated exactly as in paragraph 2.1, is to find such intersections (D_{I} and D_{II}) near which the epicenters of strong earthquakes may occur. However, new objects are defined. A larger area is considered, and dip-slip and strike-slip earthquakes are not distinguished. The last difference is not significant, since the area contains only two strong dip-slip earthquakes (1952, near intersection *165*; and 1971, near intersection *119*; the first epicenter is near that of a strong strike-slip earthquake). In the numerical experiments (paragraph 4.5) we eliminated these two earthquakes from learning. The distinctions between recognition of San Andreas and Basin–Range earthquake sources are considered in Briggs and Press (in preparation).

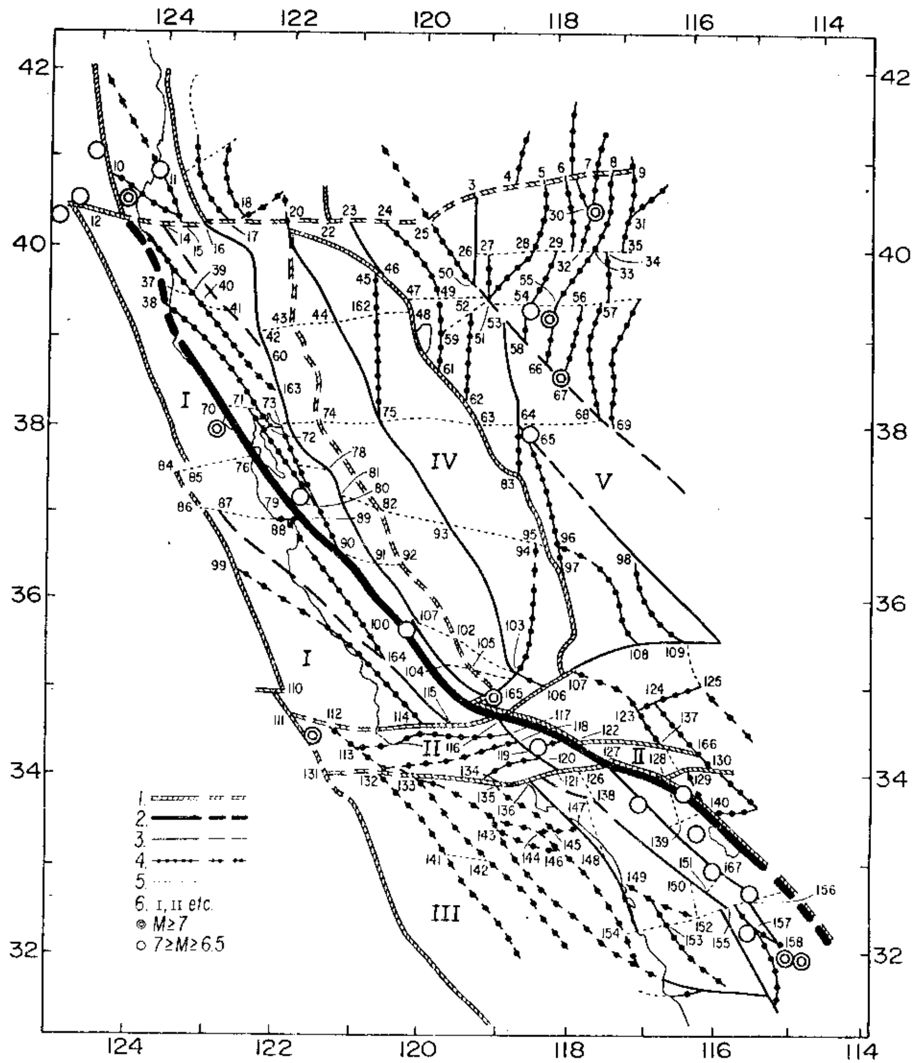


Fig. 4. Major lineaments of California (for definitions, see Appendix).
(1) Lineaments of first rank (boundaries of morphostructural countries).
(2) Major branch of San Andreas system; solid segments are reasonably certain, dashed segments are uncertain, underwater or buried under soft sediments.
(3) Lineaments of second rank (boundaries of morphostructural countries) megablocks.
(4) Lineaments of third rank (boundaries of morphostructural countries) blocks.
(5) Transverse lineament.
(6) The number of morphostructural country in the description of the scheme (second paragraph of Appendix); epicenters are shown as in legend.

TABLE XVII

Parameters of intersections

N	Parameter	Parameterization: learning for California	Parameterization: learning for Anatolia
1	absolute elevation, h_1 (m)	$\leqslant$ 200 $\leqslant$1,200	$\leqslant$ 200 $\leqslant$1,200
2*	maximal elevation, h_{max} (m)	–	
3*	maximal difference of elevations, Δh (m)	–	
4*	the distance between points with maximal elevations, l_1 (km)	–	
5*	$\Delta h/l_1$	–	$\leqslant$ 75 $\leqslant$ 125
6*	the number of lineaments, departing from intersection, n_1	–	$\leqslant$ 3
7*	morphology (most contrasting combinations in the forms of relief)	MM (mountain–mountain) MP (mountain–plain) and all the rest	
8*	types of tectonic structures	–	
9*	relative area of soft sediments, q (%)	–	
10	type of intersection	x or ⊥	
11	how pronounced is the lineament	strong feeble	strong feeble
12	the distance to the closest lineament, r_1 (km)	= 0 $\leqslant$ 25	= 0
13	the distance to the second closest lineament, r_2 (km)	$\leqslant$ 25 $\leqslant$ 37.5	$\leqslant$ 25 $\leqslant$ 37.5
14	the distance to the closest lineament of the first rank, $R_{1.1}$ (km)	–	
15	the distance to the second closest lineament of the first rank, $R_{1.2}$ (km)	–	
16	the distance to the closest lineament of the second rank, $R_{2.1}$ (km)	–	
17	the distance to the second closest lineament of the second rank, $R_{2.2}$ (km)	$\leqslant$ 50 $\leqslant$ 100	$\leqslant$ 50 $\leqslant$ 100
18	the distance to the closest lineament of the third rank, $R_{3.1}$ (km)	–	
19	the distance to the second closest lineament of the third rank, $R_{3.2}$ (km)	–	
20	the distance to the second closest longitudinal lineament, r_3 (km)	–	= 0 < 25
21	the distance to the second closest transverse lineament, r_4 (km)	–	< 62.5 < 100
22	the length of longitudinal lineament, l_2 (km)	–	
23	the length of the transverse lineament, l_3 (km)	–	
24	the distance to the closest intersection, r_5 (km)	= 0 $\leqslant$ 37.5	= 0 $\leqslant$ 37.5
25	the distance to the second closest intersection, r_6 (km)	–	$\leqslant$ 37.5 $\leqslant$ 50
26	the distance to the closest intersection inside some cluster, r_7 (km)	–	
27	maximal elevation, H_{max} (m)	$\leqslant$1,500	$\leqslant$1,500 $\leqslant$3,000
28	minimal elevation, H_{min} (m)	$\leqslant$ 100	
29**	maximal difference of elevations, ΔH (m)	–	$\leqslant$2,000 $\leqslant$3,000
30**	the distance between points with maximal and minimal elevations, l_4 (km)	–	

TABLE XVII (continued)

N	Parameter	Parameterization: learning for California	Parameterization: learning for Anatolia
31**	$\Delta H/l_4$	$\leqslant$ 40	
32**	relative area of soft sediments, Q_2 (%)	–	
33**	morphology	–	
34**	the number of intersections, n_2	$\leqslant$ 4 $\leqslant$ 7	$\leqslant$ 4 $\leqslant$ 7
35***	the weighted rank of lineament, M	$\leqslant$ 15 $\leqslant$ 17	$\leqslant$ 15 $\leqslant$ 17
36	the distance from San Andreas fault, l_5 (km)	–	–

Thresholds are indicated only for parameters used in learning.

* Parameter is determined in a circle of 12.5-km radius around intersection.

** Parameter is determined in a circle of 62.5-km radius around intersection.

*** $M = \sum_{j=1}^{n} (m_i/62.5)(62.5 - R_{ij})$ where i is the rank of the lineament; m_i is the weight of the rank ($m_1 = 5$, $m_2 = 4$, $m_3 = 3$); R_{ij} is the distance to the jth lineament of the rank i; n is the number of lineaments in the circle of 62.5-km radius around the intersection.

We apply recognition to the land part of California – the territory between the latitude of the Mendocino fault on the north, the Mexican border on the south and the axis of Great Valley on the east. In Fig. 4 this territory is bounded by lineaments *76–13–20–106–107–123–125* and the oceanic shore to the south of intersection *76*.

4.2. The data

4.2.1. Definitions

As in Section 3, we use for recognition only certain parameters which roughly characterize tectonic fracturing of the crust or the intensity and degree of contrast of neotectonic movements. This selection does not imply lack of significance to other factors, such as seismicity, geophysical anomalies, tectonic history, the structure of the crust, etc. The parameters used for recognition are listed in Table XVII. Parameters 6, 11, 21, 24–26, 34 and 35 characterize tectonic fracturing. Parameters 1–5, 7, 9 and 27–33 show the intensity and degree of contrast of the vertical neotectonic movements. Parameter 10 characterizes the presence of strike-slip faults. Some additional comments follow.

4.2.2. Conditions on lineaments

All lineaments are continuous and may change only at intersections; if the rank of a lineament changes at an intersection we consider it as two lineaments; if the strike of a lineament changes by more than ν degrees we consider it as two lineaments (in practice we took $\nu = 15°$).

4.2.3. Separate intersections

If the distance between intersections was less than δ we replaced them by one intersection at the midpoint (in practice we took $\delta = 12.5$ km or 5 mm on a map with scale 1:2,500,000; not more than three intersections were merged by this rule).

4.2.4. Vicinity of intersections

Parameter 1 is determined directly at the point of intersection. Parameters 2–11 are determined within the circle of radius $\epsilon_1 = 12.5$ km with a center at the intersection. Parameters 27–35 are determined in a larger circle of radius $\epsilon_2 = 62.5$ km.

4.2.5. Continuity of parameters

Since the lineaments are drawn approximately, all parameters are defined in such a way that they are continuous with respect to displacement of lineaments. They should also be continuous in the arbitrary parameters ϵ_1 and ϵ_2. Parameters 12–19 and 24–34 may become discontinuous if an intersection is formed by three or more lineaments, since after the displacement of any lineament several new intersections will appear and the considered characteristics will change sharply. To preserve the continuity we specify the definition of these parameters in the following way. Let us imagine that we displace the axes of lineaments slightly in such a way that they will not be parallel, will intersect only in pairs and that a T-intersection will not become a crossing. Each parameter converges to some limit when this displacement decreases. This limit is taken as the value of the parameter. In practice it is equivalent to the following rules (Table XVII):

Parameter numbers N = 12, 13 and 24: $r_1 = r_2 = r_5 = 0$ if more than two lineaments depart from intersection.
N = 34: n_2 = 1, 3 and 6 if the number of intersecting lineaments is 2, 3 and 4, respectively.
N = 25: $r_6 = 0$ if more than four lineaments intersect.
N = 14, 16 and 18: $R_{i,1} = 0$ if a lineament of rank i goes through the intersection.
N = 15, 17 and 19: $R_{i,2} = 0$ if more than one lineament of rank i goes through the intersection.

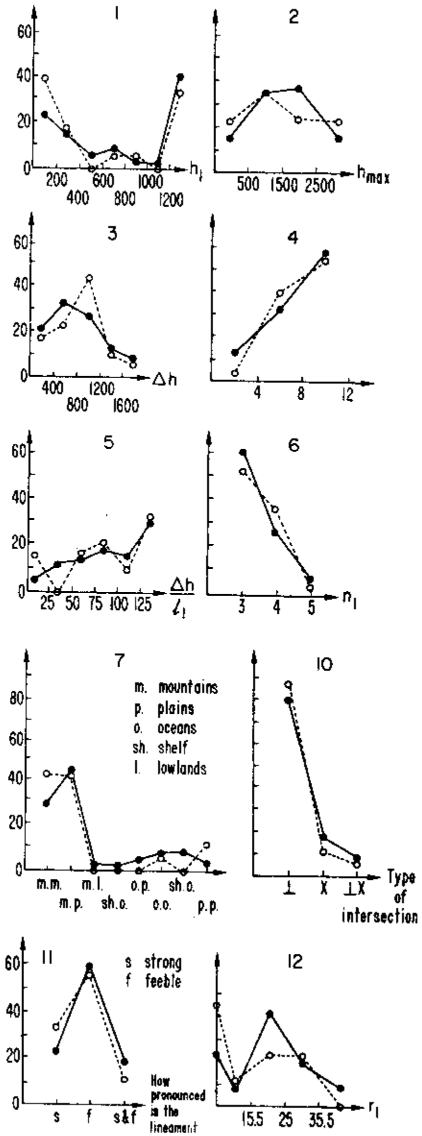
For example intersection *116* has $r_1 = r_2 = 0, n_3 = 6$; intersection *143* has $r_1 = 0, r_2 \neq 0, n_3 = 2$.

4.2.6. The length of lineaments (parameters 22 and 23).

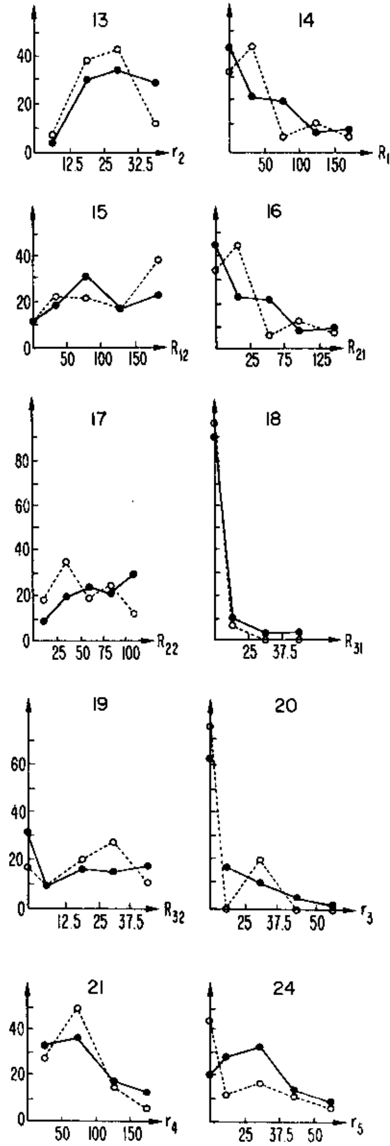
Suppose that all the longitudinal lineaments are placed in order of rank, with highest rank first. Lineaments of the same rank are placed in order according to their length, with greatest length first. After that ordering the first length will be l_2; l_3 is specified similarly.

4.2.7. One-dimensional distributions

The parameters specified in the preceding section can be determined from the scheme of lineaments (Fig. 4) and topographic and tectonic maps. Their one-dimensional distributions are shown in Fig. 5. Dashed lines correspond to the intersections closest to epicenters; solid lines correspond to all the remaining intersections. We see that the histograms for the



42.



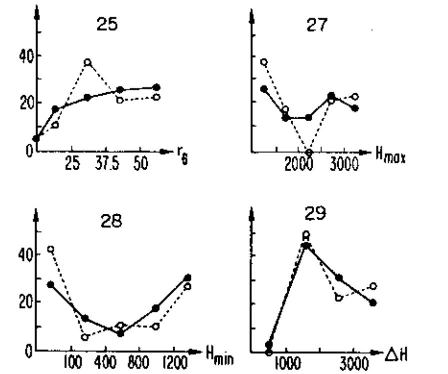


Fig. 5. Examples of one-dimensional distributions of parameters for intersections of lineaments shown in Fig. 4. The number and notation of parameters correspond to Table XVII. Otherwise, the notation is the same as in Fig. 2.

two groups are very similar for parameters 2–4, 7–9, 14–16, 18, 19, 22, 23, 28 and 30–33. These parameters were not used. The parameter l_5, distance from the San Andreas fault, was not selected for the opposite reason that it would dominate the characteristic features; the histograms separate too well. Since most known epicenters are close to this fault, the results of recognition would be trivial. Discrete values were assigned to the remaining parameters as shown in the third column of Table XVII.

Fig. 5 contains the histograms for the whole area shown in Fig. 4. Histograms based only on California data separate better, allowing thresholds such that the two groups have few or no common values of some parameters. This is undesirable for the reason indicated in paragraph 3.3; this is why we assumed thresholds corresponding to the data for the entire area in Fig. 4 although recognition was carried out for only a portion of the area.

4.3. Results

Algorithm CLUSTERS was used. Intersections within 50 km of an epicenter formed a cluster. The division of intersections between the clusters and classes II and III in learning is indicated in Table XIX.

4.3

TABLE XVIII

Distinctive D and N features for points located at intersections of lineaments

		1	7	12	13	17	24	27	28	31	34	35
		h_1	Morphology	r_1	r_2	$R_{2,2}$	r_5	H_{max}	H_{min}	$\Delta H / l_4$	n_2	M
D	1					>50						>17
	2	≤200										>17
	3		m,m						≤100			>15
	4		no m,m				≤12.5					>15
	5	>200					≤12.5					>15
	6		m,p		>25							>15
	7		m,m	≠0								>15
	8				>25				≤100		>4	
	9				>25				≤100	>40		
	10			≤25					≤100	>40		
	11		no m,m			≤50			≤100			
	12	≤200		≤25					≤100			
	13			≤25	>25			≤1500				
N	1							>1500				≤15
	2					>50					≤4	≤17
	3							>1500	>100			≤17
	4					>50		>1500				≤17
	5		no m,p				≤37.5					≤17
	6		no m,p			≤100						≤17
	7			≤25			≤37.5					≤15
	8	≤1200	no m,p									≤15
	9							>1500			≤4	
	10						≤37.5				≤4	
	11					≤100				≤40	≤4	
	12	≤1200	no m,m								≤4	

m - mountains p - plains

TABLE XIX

Results of voting for points at intersections of lineaments in California (see Table VIII for definitions of symbols)

n_N								
≥10	$\widetilde{40}$, $\widetilde{123}$, $\widetilde{124}$, $\widetilde{134}$, 145, $\widetilde{148}$, $\widetilde{149}$	$\widetilde{92}$						
9								
8								
7		$\widetilde{37}$, $\widetilde{125}$, $\widetilde{154}$						
6	$\widetilde{91}$	$\widetilde{38}$, $\widetilde{114}$						
5	10 $\widetilde{41}$, $\widetilde{42}$, $\widetilde{43}$, $\widetilde{82}$							
4	$\widetilde{39}$, $\widetilde{60}$, $\widetilde{135}$, 144, $\widetilde{153}$	143, 146	$\widetilde{107}$					
3	141, $\widetilde{147}$ 166	$\widetilde{132}$	$\widetilde{102}$, $\widetilde{104}$, $\widetilde{115}$	100		*158*		
2	$\widetilde{74}$ 150	$\widetilde{17}$, $\widetilde{152}$ 164						
1	142	88 133	$\widetilde{90}$, $\widetilde{130}$	137				
0	70 $\widetilde{76}$ 78 73 $\widetilde{106}$, $\widetilde{121}$	$\widetilde{15}$ 118 119 120	14 $\widetilde{18}$, $\widetilde{36}$, 138 $\widetilde{127}$ 136 89	*13* $\widetilde{72}$, $\widetilde{126}$ 151 153 167	$\widetilde{16}$ 158 *80* 81 *101* 116 *117* *122* 77	$\widetilde{20}$ *71* $\widetilde{73}$ 103 105 128 129 139	*155* *156* *157* *165*	*140*
	0	1	2	3	4	5	6	≥7

n_D

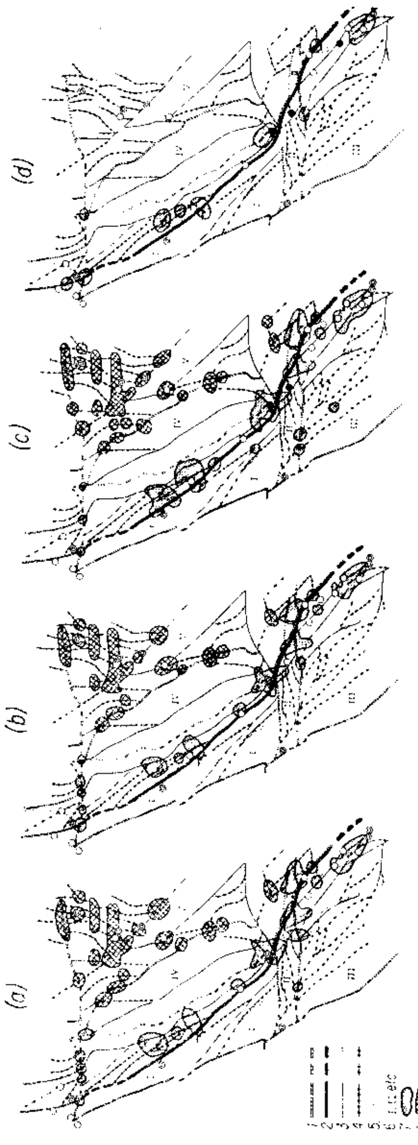


Fig. 6. Results of pattern recognition for intersections of major lineaments of California.
Dotted shading surrounds areas recognized as D for California. Crossed shading surrounds areas recognized as D for Sierra Nevada and adjacent parts of Basin and Range.
(a) Experiment EF.
(b) Main variant; all known earthquakes up to 1974 are used in learning.
(c) Experiment EH: all known earthquakes up to 1948 are used in learning.
(d) Experiment EH: all known earthquakes up to 1940 are used in learning.

TABLE XX

Recognition for control experiments EH and EF. The notation is the same as in Table IX. Uncertain ("neutral") voting: ± corresponds to $\Delta = \bar{\Delta}$; $\underline{\circ}$ corresponds to $\Delta = \bar{\Delta} - 1$

Number of intersection on Fig. 4	Year of earthquake which defines cluster	Learning based on earthquakes from 1836 through the year 1934	1940	1942	1948	1952	1968	1971	1973+...
		Clusters $\{D_1 + N_1\}$:							
119		o	o	o	o	o	o	o	o
120		o	o	o	o	o	o	o	+
118	1971	o	o	o	o	o	o	o	o
136		+	+	o	+	+	+	+	+
117		o	o	o	+	+	+	+	+
122		o	o	o	+	+	+	+	+
140	1968	o	+	+	+	+	+	+	+
139		o	o	o	+	+	+	+	+
165		o	+	+	+	+	+	+	+
116		o	o	o	±	+	+	+	+
103	1952	o	o	o	±	+	+	+	+
105		o	+	+	+	+	+	+	+
117		o	o	o	+	+	+	+	+
128		o	o	o	±	+	+	+	+
129		o	o	o	±	+	+	+	+
139	1948	o	o	o	+	+	+	+	+
140		o	+	+	+	+	+	+	+
166		o	o	o	o	o	o	o	o
167		o	o	+	+	+	+	+	+
151	1942	o	o	+	+	+	+	+	+
150		o	o	o	±	o	$\underline{\circ}$	o	$\underline{\circ}$
155		o	+	+	+	+	+	+	+
156		o	+	+	+	+	+	+	+
157	1940	o	+	+	+	+	+	+	+
155		o	+	+	+	+	+	+	+
158	1934	+	+	+	±	±	±	±	o
157		o	+	+	+	+	+	+	+
13		±	±	+	+	+	+	±	±
10	1923	±	±	o	o	o	o	o	o
14		$\underline{\circ}$	o	o	$\underline{\circ}$	±	±	±	o
101		$\underline{\circ}$	o	±	+	+	+	+	+
164	1922	+	±	$\underline{\circ}$	±	$\underline{\circ}$	±	o	o
100		o	o	·	$\underline{\circ}$	o	$\underline{\circ}$	o	o
158	1915	+	+	+	±	±	±	±	o

TABLE XX *(continued)*

Number of intersection on Fig. 4	Year of earthquake which defines cluster	Learning based on earthquakes from 1836 through the year 1934	1940	1942	1948	1952	1968	1971	1973+...
80		±	±	+	+	+	+	+	+
79		$\underline{\circ}$	$\underline{\circ}$	o	±	$\underline{\circ}$	±	o	o
89		±	±	+	+	+	+	±	+
88	1911	±	±	±	+	±	+	o	o
81		$\underline{\circ}$	o	+	+	+	+	+	+
77		$\underline{\circ}$	±	+	+	+	+	+	±
78		$\underline{\circ}$	$\underline{\circ}$	$\underline{\circ}$	±	$\underline{\circ}$	±	o	o
70	1906	$\underline{\circ}$	$\underline{\circ}$	$\underline{\circ}$	±	$\underline{\circ}$	±	o	o
71		+	+	+	+	+	+	+	+

Number of intersection on Fig. 4	Year of earthquake which defines cluster	Learning based on earthquakes from 1836 through the year 1940	1947	1948	1954	1954	1959	1973+...
		Class II $\{D_2 + N_2\}$:						
3		o	+	+	o	o	o	o
5		±	+	+	+	+	±	+
6		+	+	+	+	+	+	+
8		+	+	+	+	+	±	+
9		o	±	±	+	±	o	o
25		o	$\underline{\circ}$	±	+	±	o	+
26		o	+	+	+	+	+	+
27		o	+	+	+	+	+	+
46		+	+	+	+	±	±	±
50		+	+	+	+	+	±	+
57		+	+	+	+	+	±	+
59		+	±	+	+	+	±	+
61		o	+	+	+	o	o	o
68		+	+	+	+	+	±	+
69		o	+	+	+	+	±	+
94		o	±	o	o	o	o	o
95		+	+	+	+	+	±	+

The remaining intersections were recognized only as N.

The characteristic features are shown in Table XVIII.

The results of voting are shown in Tables XIX and XX, and in Fig. 6b.

The threshold for recognition was chosen as follows: $\bar{\Delta} = \Delta^*$ with Δ^* having the same definition as in paragraph 3.4.

We note that the recognition is successful in that most of the intersections are divided by voting into two rather distinctive groups, with only a small number of intersections given a "neutral" vote. Each cluster has at least one D intersection (Table XX). The majority of D and N intersections belong to class I (clusters) and to class II, respectively. The results are also encouraging for the following reasons.

(1) Non-instrumental data for the 19th century

show four earthquakes in the area of interest with estimated magnitude 7 or more. Epicenters of all these earthquakes are close to intersections recognized as D.

(2) Among D are the intersections closest to epicenters, with only two exceptions. One of the exceptions is the San Francisco earthquake of 1906. Intersection *70* which is closest to the 1906 epicenter was recognized as N; however, intersection *71*, which is slightly farther away, was recognized as D. The second exception is the epicenter of the 1971 earthquake, which is of the dip-slip variety. The closest intersection *119* is recognized as N; intersection *136* is recognized as D in the same cluster.

4.4. Interchangeability of distinctive features from different regions

Let us apply to California the distinctive features obtained for Anatolia, and vice versa. In this way we check our ability to recognize intersections near which epicenters of strong earthquakes actually fall, using criteria developed elsewhere. This is an important control experiment, since the features in the two

TABLE XXI

Distinctive features for recognition in Anatolia (Gelfand et al., 1974a)

		1	5	6	10	11	12	13	17	20	21	24	25	27	29	34	35
		h_l	$\Delta h/l_i$	n_l	intersection	lineament strength	r_1	r_2	$R_{2,2}$	r_3	r_4	r_5	r_6	H_{max}	ΔH	n_2	M
D	1											12.5-37.5					>9
D	2			=3								>12.5					>9
D	3					s or f										>2	
D	4				⊥											>2	
D	5								≤100				≤100			>2	
D	6				⊥				≤100				≤100		≤3000		
D	7									≤50					≤3000		
D	8								≤100				≤100		>2000		
D	9									>12.5			≤100	≤3000			
D	10								≤100				≤100	>2000			
D	11								≤100			>12.5	≤100				
D	12		>150									>12.5	≤100				
D	13					s or f				≤50		>12.5					
N	1			>3		s and f										≤4	
N	2	>200								>50						≤2	
N	3											>37.5			>3000		
N	4							>25						>3000	>2000		
N	5					f							>62.5	>3000			
N	6						>25	>62.5						>2000			
N	7					s and f							>62.5				
N	8	>200								>50			>62.5				
N	9	>200						>62.5					>62.5				
N	10	>200											>62.5				
N	11	>200								>50		>37.5					
N	12	>1000	≤300			f						>37.5					
N	13					f	>25		≤100								
		1000 ↓ 1200	150 ↓ 75; 300 ↓ 125				25 ↓ 12.5	62.5 ↓ 37.5					62.5 ↓ 37.5; 100 ↓ 50	2000 ↓ 1500		2 ↓ 4; 4 ↓ 7	9 ↓ 15

s = strong f = feeble

regions were obtained completely independently. Success would have important geological implications.

The characteristic features, obtained in Gelfand et al. (1974a, b) for Anatolia, are shown in Table XXI. They were applied to California with one natural exception: the parameter thresholds were reset to values appropriate for California.

The results of voting, shown in Fig. 7c, are quite good. We first note that the intersections in California receive many votes; the average ($n_D + n_N$) is 8, despite the fact that the distinctive features were derived from Anatolian data. The intersections are distinctly divided by voting into two groups, so that recognition is stable with respect to the choice of $\bar{\Delta}$.

Second, with $\bar{\Delta} = 0$ (the same threshold used in Gelfand et al. (1974a, b), for Anatolia) we recognize as D all intersections closest to each epicenter, even number *70* near the 1906 epicenter.

Third, we recognize as D all the same intersections, as in paragraph 4.3 as well as a few additional ones.

We can reverse the procedure and apply to Anatolia the characteristic features found for California (Table XVIII). Again, we changed the thresholds assigned to the parameters. The results are analogous. The intersections have many votes and are distinctly divided by voting. With $\bar{\Delta} = 0$ we recognize the intersections near each known epicenter (Fig. 8) except one (near intersection *33*). D areas are in good agreement with those found in Gelfand et al. (1974a, b), where learning was based on Anatolian data.

The ability to transfer the criteria of high seismicity from Anatolia to California seems natural, since most of the Anatolian earthquakes are connected with strike-slip movements, as is the case in California. However, the complete recognition of intersections near all known epicenters and the good agreement with results based on Californian data is astonishing and implies some generality to the distinctive features which were found. This particular control experiment argues against the criticisms that the pattern recognition results were fortuitous, or simply a restatement of the idea that future epicenters occur near previous ones.

For the Sierra Nevada and Basin and Range (regions *IV* and *V* in Fig. 4) the Anatolian criteria give negative results in that no intersections are recognized as D in the vicinity of 2 out of 5 epicenters. This failure is expected since the eastern California - Nevada regions are dominated by dip-slip focal mechanisms. This result underlines the success for California.

4.5. Control experiments

Experiment EH (Figs. 6c, d and Table XX) shows quite satisfactory results. Learning from earthquakes through the year 1940, we would have recognized at least one dangerous intersection in the cluster around the epicenter of each subsequent strong earthquake. After the 1942 earthquake, we would miss only one cluster, namely that corresponding to the 1971 earthquake. This cluster would be recognized by learning on all shocks up to and including the 1948 earthquake. To save time we did not redetermine the histograms of parameters at each step as we did in paragraph 3.5. The parameters and their thresholds were the same for each step. Therefore the results of this experiment only illustrate the stability of recognition with respect to the learning material.

Experiment EF (Fig. 6a) also shows the stability of recognition in that only four more intersections were recognized as D (*120*, *15*, *130* and *133*); and two intersections were transferred to N (*14* and *158*).

To test the sensitivity of results to the particular algorithm used, CORA-3 was applied. Class I contained the intersections closest to an epicenter, class II included all the rest. The threshold $\bar{\Delta} = 0$ was chosen. The results of recognition are shown in Fig. 9b. They basically confirm the results of paragraph 4.3. Experiment EH was successful up to 1942.

In another test we assumed other thresholds for parameters, obtained from histograms based on California data only. The results of recognition by algorithm CLUSTERS are in satisfactory agreement with paragraph 4.3. The control experiment results were good, though somewhat degraded. Thus our results are stable to changes in the thresholds, but the thresholds assumed in paragraph 4.3 are preferable.

4.6. Clarifying experiments

4.6.1. Epicenters or total fault breaks?

Does the epicenter have some special meaning, or is it just a random point on a fault break, say the place where by chance the stress first reaches the critical fracture strength? To check this we replaced the epicenters by the intersections on the entire fault

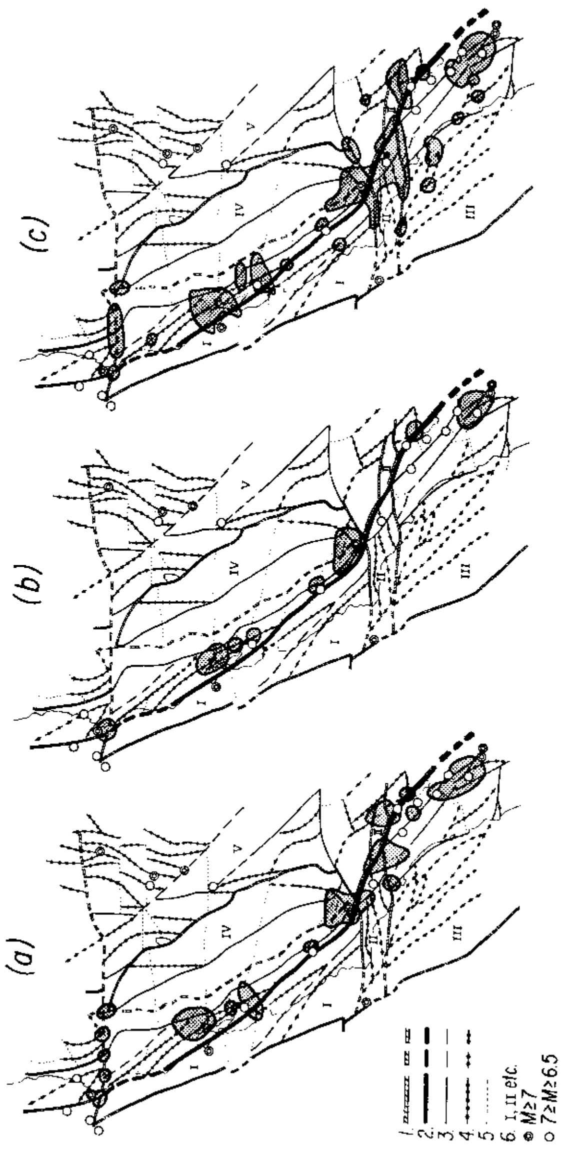


Fig. 7. Comparison of D areas for intersections of major lineaments for California, using different learning materials. Notation is the same as in Fig. 4.
(a) Main variant; same as Fig. 6b (see paragraph 4.3).
(b) Two dip-slip earthquakes eliminated (see paragraph 4.6).
(c) Learning for Anatolia (Table XXI).

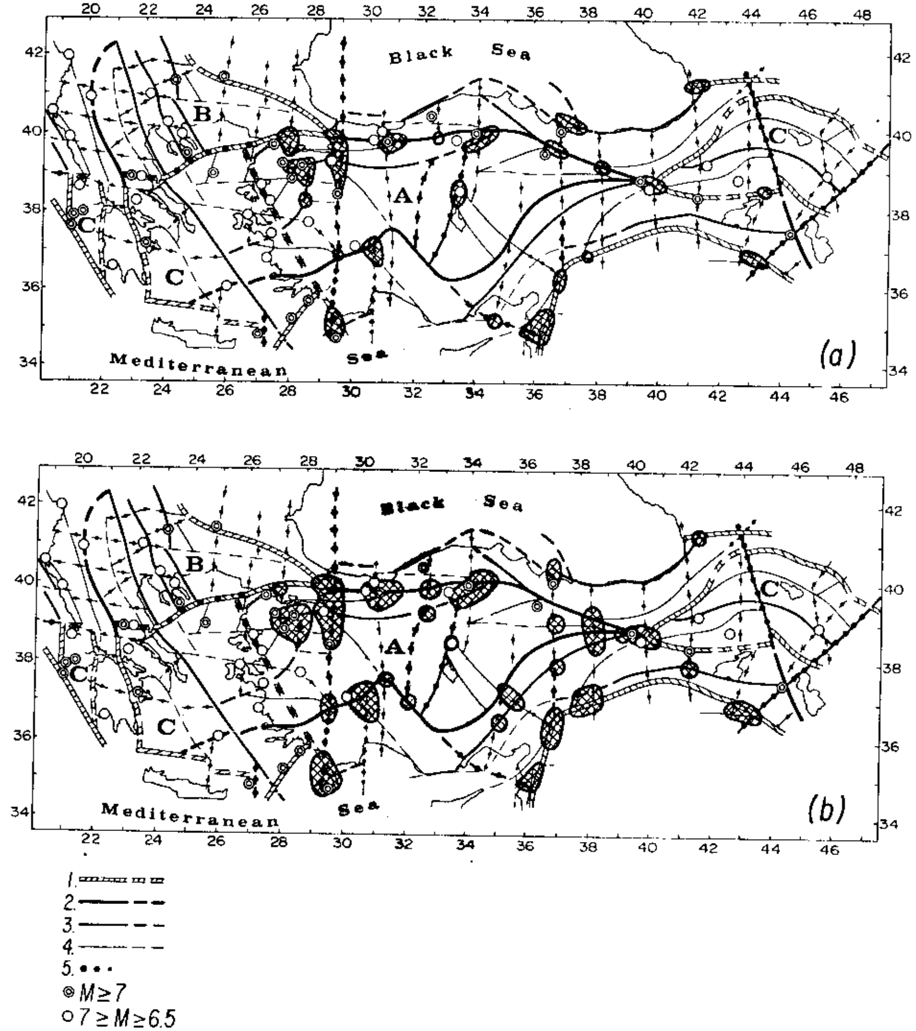


Fig. 8. Major lineaments of Anatolia.
(a) Recognition of D intersections with learning based on Anatolian data (Table XXI, after Gelfand et al. (1974a)).
(b) Recognition of D intersections by learning based on Californian data (Table XVIII).
Legend: lineaments (after Gelfand et al., 1974a): *1* = bounding the entire region (rank 1); *2* = bounding morphostructural countries (rank 1); *3* = bounding megablocks (rank 2); *4* = bounding blocks (rank 3); *5* = transverse.

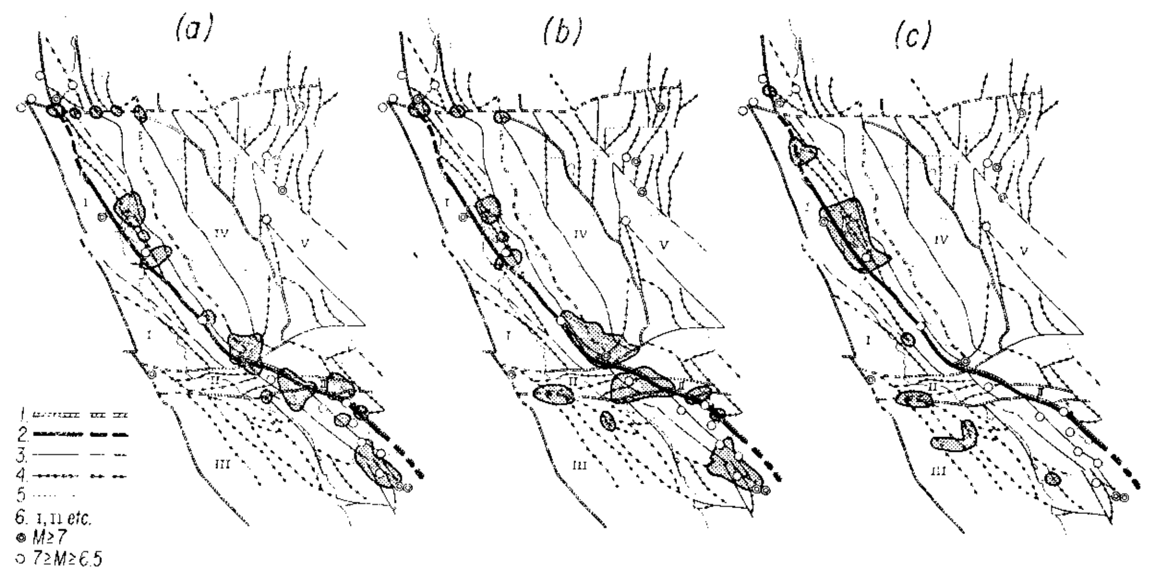


Fig. 9. Comparison of D areas for intersections of major lineaments for California. The notation is the same as Fig. 4.
(a) Main variant (same as Fig. 6b, see paragraph 4.3).
(b) Learning by CORA-3.
(c) Learning on fault breaks (paragraph 4.6).

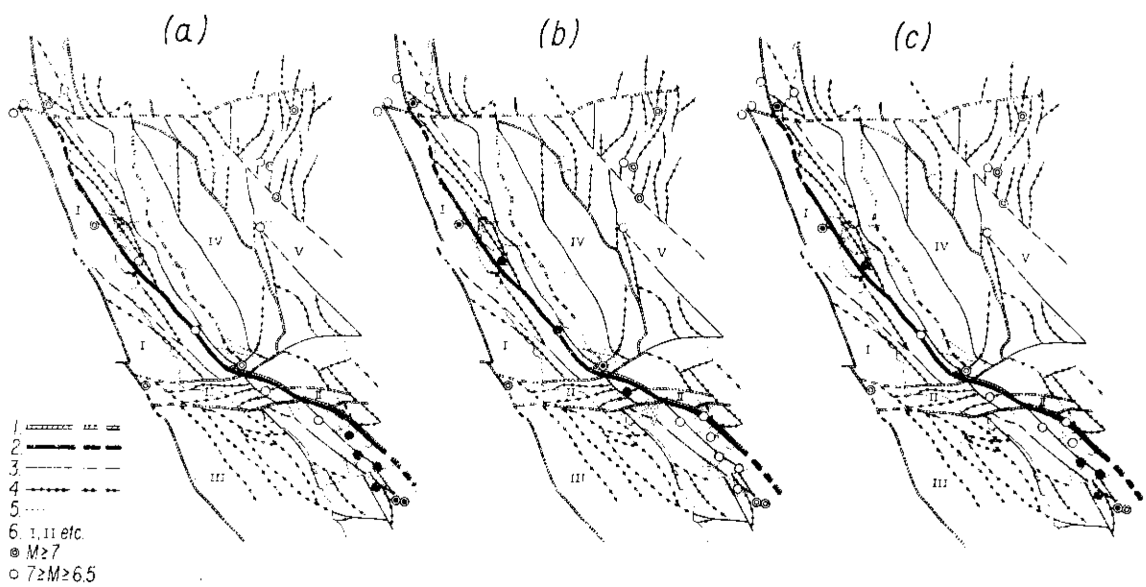


Fig. 10. Comparison of D areas for intersections of lineaments. Control experiment consists of eliminating groups of earthquakes. Epicenters eliminated from learning are filled in solid colors; the corresponding D areas are shaded. Lineaments defined in same way as Fig. 4. The experiment is aimed at recognition of intersections around the eliminated epicenters.

break of the San Francisco earthquake, 1906 (between intersections *13* and *89*). Clusters were formed around each of these intersections, exactly as was done around epicenters in the preceding sections. The results of recognition, shown in Fig. 9c, are unsuccessful. No intersections are recognized as D near 9 known epicenters out of 12. The three exceptions are the epicenters near the fault break, which are predetermined by learning. The next experiment indicates that this failure is due not to an uneven coverage of the territory by the objects of class I, but to the replacement of epicenters by the fault break.

4.6.2. Uniformity of the region

The analogous experiment was described in Section 3. The purpose is to see if we can recognize dangerous areas even if the epicenters in their vicinity are eliminated during the learning stage. Three cases were considered as shown in Fig. 10. The number of known strong earthquakes in the three cases is 6, 6 and 7, respectively. According to our experience with EH experiments, this number is marginal for successful learning. Nevertheless, the results shown in Fig. 10 are quite good in that at least one point was recognized as dangerous in almost all clusters not used in learning. The only exceptions are: two clusters near intersections *101* and *13* in Fig. 10b; one cluster near *158* in Fig. 10a. This confirms the conclusion reached in the analogous experiment in Section 3.

4.6.3. Effect of dip-slip earthquakes

Two of the earthquakes, used for learning, are of dip-slip variety (years 1952 and 1971; epicenters near intersections *165* and *119*). The clusters, corresponding to these earthquakes, were eliminated from learning in this experiment. The distinctive features did not change noticeably. The results of recognition (with the same threshold as in main variant, $\bar{\Delta} = \Delta^*$) are shown in Fig. 7b. We see that the number of D intersections decreased significantly, from 25 in the main variant to 17. Only three points outside of clusters are recognized as D (*72, 73* and *104*), slightly expanding already known clusters. Fig. 6 shows that the main effect was the elimination of a group of D points in the Transverse Ranges, a geologically reasonable result.

4.7. Comparison with Section 3

D areas recognized for intersections and for points on major faults (Figs. 3 and 6) basically coincide in that the same five groups of D-points show up in both cases. The slight differences are due to the fact that the study of intersections covers a larger territory.

The characteristic features (Table XVII) show that D and N intersections are often characterized by low and high elevations (h_1, H_{min} and H_{max}), respectively. This reinforces the idea derived from the results in Section 3 that D intersections are often associated with neotectonic subsidence against a background of weak uplift. Two more observations may be made.

(1) Additional evidence of the importance of vertical movements at D intersections are the large gradients ($\Delta H/\ell_4$) and the mountain–plain contacts which show up in many D features, with opposite conditions holding for N features.

(2) The fracturing of the crust is more pronounced for D than for N intersections. This is shown by the occurrence of lineaments of high rank in D features and low rank in N features. Also many N features show small values of n_2 – the number of intersections within 62.5-km distance.

5. Conclusions

Using pattern recognition, we have divided California into two types of areas: D areas where epicenters of strike-slip earthquakes with $M \geqslant 6.5$ can occur and N areas where these epicenters cannot occur. Distinctive features for each of these areas were identified. The D areas are characterized by proximity to intersections or ends of major faults and by relatively low elevation or subsidence against a background of weaker uplift. The N areas often show higher elevation or stronger uplift but less contrast in relief. D areas concentrate in five places. These results were tested against numerous control experiments and seem to be stable and reliable.

It is surprising that the relatively simple and crude parameters which make up the distinctive features suffice to achieve these results. Perhaps large earthquakes occur on faults which penetrate significantly into the lithosphere and fault movements on such a scale can be recognized in simple ways. Perhaps parameters based on systems of faults and other types of lineaments summarize more geophysical and geological information than their simplicity implies. Perhaps the result that epicenter locations occur at identifiable places along faults and not at arbitrary locations on faults implies the existence of a special property or agent in the epicentral region which acts to initiate rupture. This special property might be one which concentrates stress to weaken rock, or to "lubricate" a fault.

Our results are not useful as they stand for earthquake hazard mitigation since they identify epicenters and not the entire length of fault rupture, which may extend well beyond a D area. However a search for short-term precursors may be promising in D areas and future efforts in earthquake control by fluid injection and withdrawal could make use of D and N information.

Over the years many efforts have been made to characterize earthquake-prone regions by the occurrence of certain relationships in the field. For example, "earthquake country" has been characterized by topographic and geomorphic forms, by intersecting faults and by other ways in the geological literature. Our procedure seems to be a more precise formulation of many ideas which have been vaguely recognized in the past. It allows one to get more definite results, providing a method for recognizing differences between geologically different areas rather than describing each area by itself. It employs past experience, intuition and/or hypotheses; but it uses a logical framework to formulate explicitly the pertinent and reliable parts of these elements. The procedure consists of five steps.

(1) Preliminary separation of objects into groups.

(2) Separation of promising parameters by histogram analysis.

(3) The combination of parameters into features, singlets to triplets, and the analysis of all possible features to identify distinctive ones.

(4) The use of distinctive features in recognition.

(5) The analysis of distinctive features for their geological or geophysical meaning.

(6) The design of control experiments to prevent self-deception.

Statistical procedures are alternative methods to pattern recognition but in this case they are difficult to apply because the number of strong earthquakes is inadequately small.

54

In the future we intend to apply pattern recognition to time prediction rather than space prediction and to other geophysical applications.

Acknowledgements

This study is a product of the U.S.A./U.S.S.R. Working Group on Earthquake Prediction. The participation of Drs. L. Knopoff and F. Press was supported by grants with the U.S. Geological Survey, Department of the Interior.

Appendix – Major morphostructures and lineaments of California and adjacent regions

1. Basic definitions

Morphostructures are structures in the earth's crust (or its upper part) expressed in present day topographic relief. Tectonic and geomorphic maps often describe structures, based only on either geological or geomorphic evidence. For the study of earthquakes, however, actively developing structures are important. These structures are expressed in topographic relief; geomorphic evidence may therefore be important for their reconstruction, together with geological data.

Rank of morphostructure. The following subdivision seems adequate for tectonically and seismically active regions. Each region is divided into morphostructural countries (I), countries into megablocks (II), and megablocks into blocks (III). The rank assigned to each type of morphostructure is indicated by brackets.

Countries differ by the type of orogenesis and/or by the large-scale tectonic features.

Megablocks inside a country differ either by the dominant type of relief, or by the average parameters of the relief (such as heights of peaks, dominant strikes, relative area occupied by basins and ranges, etc.); or by the pattern of the main elements of relief.

Blocks inside megablocks differ in neotectonic history; for example, large separate ranges or depressions in contrast to a complex of smaller ranges, depressions and valleys can be outlined as separate blocks.

Morphostructural provinces. In some large countries, such as the Pamir and Tien-Shan regions, where the history of relief is long and complicated, an intermediate morphostructure a province – was introduced (Gelfand et al., 1973a). A province is a group of megablocks which have similar preorogenic histories, and accordingly, similar general features of relief. The same first rank, as for countries, was assigned to provinces. (For example, an internal depression inside a platform may be transformed after activation into a large intermountain basin and regarded as a province.)

55

Lineaments. The boundary zones between morphostructures are called lineaments. They have the same rank as the morphostructures which they separate. Lineaments of the first rank have widths from about 10 to 50 km, perhaps more; they may be thousands of kilometers in length. A lineament is the surface trace of an active fault (or flexure), penetrating deeply into the earth's crust and perhaps through the entire lithosphere. Different parts of a lineament may be expressed in different ways: geological, geomorphic or both: the evidence for a lineament may not necessarily be as direct as faults shown on a tectonic map. Let us describe some of the evidence for lineaments, which may differ for longitudinal and transverse lineaments.

Longitudinal lineaments are approximately parallel to the main elements of relief (such as ridges, basins, elongated intra-mountain depressions, etc.). Usually they form the boundaries between these elements. They are often expressed directly by zones of tectonic faults. Detailed tectonic maps show them as a system of narrow, elongated blocks. However, some parts of longitudinal lineaments are not traced by faults, but by geomorphic evidence: direct contact of contrasting forms of relief such as steep mountain front and flat plain, linear form of valley between ridges, etc.

Transverse lineaments form a large angle with the dominant strike of the main elements of relief and tectonics. They are traced mainly by terminations of these elements, en-echelon patterns, relative displacements of elements, sharp changes in height or strike, sharp changes in types of relief. A regular linear pattern of these features is evidence for a transverse lineament. Transverse lineaments may also show as patterns in the occurrence of volcanoes, the highest peaks or saddle points in parallel ridges, etc. Sometimes the evidence for transverse lineaments may show simply as a straight-line river on the slope of a ridge, which does not flow in the direction of steepest descent of the slope. or distinct, short ranges, or narrow intrusive bodies with transverse strike, or transverse faults of limited extent, shown on tectonic maps. Transverse lineaments are not always marked by the clear traditional evidence of long narrow ridges, valleys and tectonic faults. The evidence for these lineaments may be indirect, subtle and discontinuous. However, the general pattern of evidence allows one to draw transverse lineaments with reasonable certainty, especially with the aid of satellite photography. The name "transverse" could be misleading in that the main distinction is not the strike, but evidence of the kind just described. For example, the Transverse Ranges of California, though perpendicular to the dominant strike of the region, are nevertheless classified as a longitudinal lineament, since the structure carries its own faults, ridges and valleys.

Disjunctive knots. Specific morphostructures of disjunctive knots are formed around the intersections of lineaments. They are zones of especially intensive fracturing and contrasting neo-tectonic movements. Following along a lineament one can notice the appearance of a disjunctive knot by a set of specific phenomena, connected with a general increase of neo-tectonic activity and with manifestations of intersecting lineaments: the types of relief become more diverse (types corresponding to each lineament show up) and appear in new combinations; the heights of ridges and height differences change drastically; boundaries of contrasting forms of relief assume a broken-line form; rivers flow along faults and their valleys straighten; the orientation of valleys corresponds to the direction of both intersecting lineaments; fault activity increases, the lineaments inside the knots are usually expressed by faults; the topographic and geological maps show mosaic patterns; a grid of surface faults expressed tectonically or geomorphically is not uncommon.

As a rule a knot is wider than the lineaments from which it is formed. The linear dimensions of knots depend on the rank of the component lineaments and may vary from 20 to 200 km or more. One knot may include several intersections of lineaments.

It is clear from this description that though the knots have rather distinct boundaries the determination of these boundaries demands quite specialized research – in the field and on detailed maps. In the absence of such research we must consider in our problem "intersections" instead of knots. An intersection is the point of intersection of the axes of lineaments. It is a much more formal object than a knot but it is easier to handle. Each intersection belongs to some knot; but it is unknown a priori which intersections belong to the same knot. Algorithm CLUSTERS was especially devised to deal with this difficulty.

2 The scheme of major lineaments (Fig. 4)

The area of interest is divided into morphostructures of three ranks: countries, megablocks and blocks (see first paragraph of this Appendix). Their boundaries are the lineaments of corresponding rank, shown in Fig. 3. The five following countries are represented:

I = Coast Ranges; *II* Transverse Ranges; *III* = southern California-northern end of the Peninsula Ranges and the entire Mexican section of the Cordilleras; *IV* = Sierra-Nevada; *V* = western part of the Basin and Range Province.

Country *I* belongs to the Pacific orogenic belt, with active Cenozoic folding. The four other countries belong to the Cordilleran orogenic belt where the folding was mostly finished in the Mesozoic area. These four countries differ in Pliocene Quaternary orogenic development. Faults of the San Andreas system cross countries, *I*, *II* and *III*. Its width is about 60 km in the north, near San Francisco Bay and up to 160 km, if not more, in southern California. The details are described elsewhere (Ranzman, in preparation).

References

Atwood, W., 1940. The Physiographic Provinces of North America. Ginn, Boston, Mass., 535 pp.

Bongard, M.M., Vaintsveig, M.I., Guberman, Sh.A., Izvekova, M.L. and Smirnov, M.S., 1966. The use of self-learning programs in the detection of oil containing layers. Geol. Geofiz., 6: 96–105 (in Russian).

Briggs, P. and Press, F., in preparation. Pattern recognition applied to earthquake epicenters in California and Nevada.

Coffman, J.L. and Von Hake, C.A., 1973. Earthquake history of the United States. Natl. Oceanic Atmos. Adm. (U.S.), Publ., 41-1, 208 pp.

Cook, K.L., 1966. Rift system in the basin and range province. In: The World Rift System, Proc. UMC Symp., Ottawa, Ont., September 6–8, 1965, Geol. Surv. Can., Pap., 66-14: 246–275.

Dickinson, W.R. and Grantz, A., 1968. Proc. Conf. on Geological Problems of San Andreas. In: Geological Sciences, Vol XI, Stanford University, Palo Alto, Calif., 374 pp.

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

Gelfand, I.M., Guberman, Sh.A., Izvekova, M.L., Keilis-Borok, V.I. and Ranzman, E.Ia., 1973a. Recognition of places where strong earthquakes may occur, I. Pamir and Tien-Shan. In: Computational Seismology, Vol. 6 (in Russian).

Gelfand, I.M., Guberman, Sh.A., Kaletskaja, M.S., Keilis-Borok, V.I., Ranzman, E.Ia. and Zhidkov, M.P., 1973b. On the transfer of criteria of high seismicity from Central Asia to Anatolia and adjacent regions. Dokl. Akad. Nauk S.S.S.R., 20, N2 (in Russian).

Gelfand, I.M., Guberman, Sh.A., Kalezkaja, M.S., Keilis-Borok, V.I., Ranzman, E.Ia., Rotwain, I.M. and Zhidkov, M.P., 1974a. Recognition of places where strong earthquakes may occur, II. Four regions of Asia Minor and S-E Europa. In: Computational Seismology, Vol. 7 (in Russian).

Gelfand, I.M., Guberman, Sh.A., Keilis-Borok, V.I., Ranzman, E.Ia., Rotwain, I.M. and Zhidkov, M.P., 1974b. Recognition of places where strong earthquakes may occur. III. The case when the boundaries of disjunctive knots are unknown. In: Computational Seismology, Vol. 7 (in Russian).

Gutenberg, B. and Richter, C.F., 1954. Seismicity of the Earth and Associated Phenomena. Princeton University Press, Princeton, N.J., 310 pp.

Hain, W.Y., 1971. Regional Geotectonics, Vol. I. Moscow University, Moscow (in Russian).

Hamilton, W. and Myers, W.B., 1966. Cenozoic tectonics of the Western United States. In: The World Rift System. Proc. UMC Symp., Ottawa, Ont., September 6–8, 1965, Geol. Surv. Can., Pap., 66-14: 291–306.

Hill, M.L., 1966. The San Andreas system, California and Mexico. In: The World Rift System, Proc. UMC Symp., Ottawa, Ont., September 6–8, 1965, Geol. Surv. Can., Pap., 66-14: 239–245.

King, P.B., 1969. The tectonics of North America. A discussion to accompany the *Tectonic Map of North America, Scale 1 : 5.000,000.* U.S. Geological Survey, Washington, D.C.

National Oceanic and Atmospheric Administration, 1965–1973. Bulletins of preliminary determination of epicenters. Boulder, Colo.

Ranzman, E.Ia., in preparation. Major morphostructures and lineaments of California.

Richter, C., 1958. Elementary Seismology. Freeman, San Francisco, Calif., 768 pp.

Rothe, J.P., 1969. The seismicity of the Earth. UNESCO, Paris, 336 pp.

Thompson, G.A., 1966. The rift system of the Western United States. In: The World Rift System, Proc. UMC Symp., Ottawa, Ont., September 6–8, 1965, Geol. Surv. Can., Pap., 66-14: 280–289.

USGS, no date. Topographic Map of the United States. Scale 1 : 1,000,000, Washington, D.C.

USGS, 1962. Tectonic Map of the United States. Scale 1 : 2,500,000, Washington, D.C.

USGS, 1965. Geologic Map of North America. Scale 1: 5,000,000, Washington, D.C.

Wright, H. and Frey, D., 1965. The Quaternary of the United States. Princeton University Press, Princeton, N.J., 922 pp.