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 460392-1

SMR/107 - 11

WORKSHOP ON PATTERN RECOGNITION AND ANALYSIS OF SEISMICITY

(5 - 16 December 1983)

RECOGNITION OF EARTHQUAKE PRONE AREAS

INTRODUCTORY LECTURE

V.I. KEILIS-BOROK
A.A. SOLOVIEV

THE PROBLEM

A SEISMIC REGION IS CONSIDERED.

GIVEN IS THE LEARNING MATERIAL -

- EXAMPLES OF AREAS, WHERE EPICENTERS OF STRONG ($M \geq M_0$) EARTHQUAKES:

  ARE KNOWN - CLASS D_0 ("DANGEROUS")

  ARE NOT KNOWN - CLASS N_0

DESCRIPTION OF EACH AREA IS ALSO GIVEN - GEOPHYSICAL, GEOLOGICAL AND GEOMORPHOLOGICAL PARAMETERS.

THE GOAL - to RECOGNIZE THE AREAS, WHERE EPICENTERS OF STRONG EARTHQUAKES

MAY BE SITUATED - CLASS D

MAY NOT BE SITUATED- CLASS N.

N_0 MAY INCLUDE D, BECAUSE IN SOME AREAS $M \geq M_0$ MAY BE POSSIBLE, THOUGH YET UNKNOWN- THAT IS WHY OUR PROBLEM HAS ARISEN.

D_0 MAY INCLUDE N DUE TO THE ERRORS IN CATALOG (IN EPICENTERS AND/OR MAGNITUDE).

THE RULE OF RECOGNITION IS OF EQUAL INDEPENDENT INTEREST.

CONSEQUTIVE STAGES

1. THE CHOICE OF REGION AND THRESHOLD M_0 (WE ARE LOOKING FOR AREAS WHERE EPICENTERS WITH $M \geq M_0$ MAY BE SITUATED)
2. SELECTION OF AREAS ("OBJECTS OF RECOGNITION")
3. MEASUREMENT OF PARAMETERS FOR EACH OBJECT. THE SET OF PARAMETERS REPRESENTS THE ANSWERS TO A QUESTIONAIRE; QUESTIONS ARE THE SAME FOR ALL OBJECTS; ANSWERS, OBVIOUSLY, MAY DIFFER.
4. SELECTION OF LEARNING MATERIAL: EXAMPLES OF AREAS WHERE EPICENTERS WITH $M \geq M_0$ ARE KNOWN (CLASS D_0) OR ARE NOT KNOWN (CLASS N_0).
5. DISCRETIZATION AND CODING OF PARAMETERS. AS A RESULT EACH OBJECT IS DESCRIBED BY A BINARY VECTOR (BINARY CODE OF THE ANSWERS TO THE QUESTIONAIRE)
6. PATTERN RECOGNITION:
   a) CONSTRUCTION OF THE RECOGNITION RULE,
   b) DIVISION OF ALL OBJECTS INTO D and N CLASSES.
7. NUMERICAL TESTS ON ADDITIONAL INFORMATION.
8. TRANSFER OF CRITERIA TO ANOTHER REGION.
9. TEST ON RANDOMIZED DATA.

STAGE 1: REGION AND THRESHOLD M_o

GENERAL BASIS - NEOTECTONICS AND SEISMICITY

SIZE OF THE REGION DEPENDS ON M_o: THOUSANDS KM FOR $M_o = 8$, HUNDREDS KM FOR $M_o = 6$

REGION SHOULD BE LARGE ENOUGH TO INCLUDE AT LEAST 15-20 AREAS D_o AND AT LEAST AS MANY AREAS N_o.

DATA: A SET OF MAPS - TOPOGRAPHY; TECTONICS; NEOTECTONICS; GRAVITY ANOMALIES; GEOLOGY ETC.

SCALE 1:5 mln FOR M=8, 1:2,5 mln FOR $M_o = ?$, 1:1 mln FOR $M_o = 6$

CATALOGS OF EARTHQUAKES - ALL AVAILABLE.

STAGE 2: OBJECTS OF RECOGNITION.

THEY ARE DETERMINED ON:

MORPHOSTRUCTURAL SCHEME;

IF UNAVAILABLE - ON THE SCHEME OF MAJOR FAULTS.

OBJECTS MAY BE OF THE FOLLOWING TYPES:

DISJUNCTIVE KNOTS - STRUCTURES DEVELOPED AT THE INTERSECTION OF LINEAMENTS (THEIR MAIN FEATURE IS MOSAIC STRUCTURE AND DEVELOPMENT).

PARAMETERS ARE MEASURED WITHIN THE KNOTS

IF DISJUNCTIVE KNOTS ARE NOT DETERMINED -

INTERSECTIONS OF LINEAMENTS OR FAULTS

PARAMETERS ARE MEASURED IN THE CIRCLES, HAVING THE CENTERS IN THE INTERSECTIONS.

POINTS ON LINEAMENTS OR FAULTS

PARAMETERS ARE MEASURED IN SIMILAR CIRCLES.

OTHER TYPES OF OBJECTS WERE ALSO CONSIDERED. AREAS OF ONLY ONE TYPE SHOULD BE CONSIDERED SIMULTANEOUSLY. OBJECTS MAY COVER NOT THE WHOLE REGION.

PRESUMPTION: EPICENTER WITH $M \geqslant M_o$ MUST BE SITUATED ONLY WITHIN SOME OBJECTS.

THIS SHOULD BE TESTED: THIS MAY BE TRUE ONLY WITH SUFFICIENTLY LARGE M_o OR NOT AT ALL. IN THE LAST CASE THE CHOICE OF OBJECTS SHOULD BE CHANGED.

STAGE 3: PARAMETERS

WIDE VARIETY OF PARAMETERS IS RELEVANT:

NEOTECTONICS, TECTONICS, GRAVITY ANOMALIES, OTHER GEOPHYSICAL FIELDS ETC.

ONLY LIMITED SET MAY BE USED SIMULTANEOUSLY.

OTHERWISE THE DANGER TO OBTAIN RANDOM RECOGNITION WILL INCREASE.

ASTONISHINGLY LIMITED SET OF DATA HAPPENED TO BE SUFFICIENT: ISOSTATIC ANOMALIES OR TOPOGRAPHY AND VOLCANISM ETC. THIS IS BECAUSE THESE DATA REFLECT A WIDE SET OF STRUCTURAL AND DYNAMICAL CHARACTERISTICS OF THE LITHOSPHERE.

THE USE OF OTHER DATA (NOT ALL AT A TIME) IS A WORTHWHILE TASK.

STAGE 4: LEARNING MATERIAL

CLASS D_o: OBJECTS HAVING KNOWN EPICENTERS WITH $M \geq M_o$ WITHIN DISTANCE R_1.

CLASS N_o: OBJECTS HAVING NO KNOWN EPICENTERS WITH $M \geq M_o - \delta$ WITHIN DISTANCE R_2 ($R_2 \geq R_1$, $\delta = 0 - 0.2$).

CLASS X: OBJECTS WHICH CANNOT BE POSITIVELY ASSIGNED TO D_o or N_o OR OBJECTS RESERVED FOR NUMERICAL TESTS.

IN OTHER WORDS A SET OF ALL OBJECTS W IS A PRIORI DIVIDED INTO THREE NON-INTERSECTING SUBSETS:

$$W = D_o \cup N_o \cup (W \setminus (D_o \cup N_o))$$

THE LAST SET [illegible]

ARE LOOKING FOR DIVISION $W = D \cup N$.

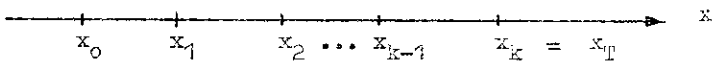
OUR GOAL IS:

USING LEARNING MATERIAL TO FIND - HOW TO RECOGNIZE OBJECTS D and N;

TO APPLY THIS RULE TO ALL OBJECTS.

STAGE 5: DISCRETIZATION AND CODING.

THE VALUES OF EACH PARAMETER X LIE WITHIN CERTAIN RANGE (x_0, x_T). WE DIVIDE THIS RANGE INTO k INTERVALS BY POINTS x_i, i = 1,2, ... k-1.



DISCRETIZATION MEANS TO CONSIDER NOT THE VALUE OF THE PARAMETER BUT ONLY THE INTERVAL WHERE IT BELONGS.

USUALLY WE DIVIDE THE RANGE INTO TWO INTERVALS ("LARGE" AND "SMALL" VALUES) OR INTO THREE INTERVALS ("LARGE", "MEDIUM" and "SMALL" VALUES).

THE PURPOSE OF DISCRETIZATION IS TO FIND SUCH INTERVALS WHERE OBJECTS OF ONE CLASS OCCUR MORE OFTEN THAN OBJECTS OF ANOTHER CLASS.

DENOTE: P_D^i - % OF OBJECTS OF CLASS D_o IN THE i-th INTERVAL,

P_N^i - % OF OBJECTS OF CLASS N_o IN THE i-th INTERVAL

DISCRETIZATION IS SATISFACTORY, IF AT LEAST FOR SOME i

$$|P_D^i - P_N^i| > 15\%$$

OBJECTIVE DISCRETIZATION: EACH INTERVAL HAS ABOUT EQUAL NUMBER OF ALL THE OBJECTS TOGETHER (OR OF THE OBJECTS D_o AND N_o TOGETHER)

DATA-FITTED DISCRETIZATION: TO FIND THE INTERVALS, WHERE THE ABOVE CONDITION IS SATISFIED.

BEWARE OF TOO LARGE $|P_D^i - P_N^i|$ - IT MAY OCCUR BY CHANCE, BUT WILL STRONGLY AND UNDULY INFLUENCE THE RESULTS; BETTER SHIFT x_i TO GET SOME MIXING.

SOME PARAMETERS MAY BE INTEGERS: SEVERAL INTEGERS MAY BE JOINED INTO ONE INTERVAL.

ELIMINATE THE PARAMETER FROM FIRST RUNS, IF FOR ANY REASONABLE DISCRETIZATION $|P_D^i - P_N^i| < 10\%$ FOR ALL i. BUT (THIS PARAMETER) MAY BE USEFUL AT LATER STAGES - IN COMBINATION WITH OTHER PARAMETERS.

IF NO SATISFACTORY DISCRETIZATION IS OBTAINED FOR AT LEAST 7-10 PARAMETERS - CHANGE LEARNING MATERIAL.

CODING:

SUPPOSE A VALUE OF A PARAMETER BELONGS TO l-th INTERVAL, $x_{l-1} < x \leq x_l$. TWO WAYS OF CODING ARE THE FOLLOWING:

i =	1	2	l-1	l	l+1	k-1	k	
I-coding:	0	0 ...	0	1	0 ...	0	0	(k digits)
S-coding:	0	0 ...	0	1	1 ...	1		(k-1 digits)

STAGE 6: PATTERN RECOGNITION.

1. CORA-3 ALGORITHM (C)

THE ALGORITHM CONSISTS OF TWO STEPS:

SELECTION OF CHARACTERISTIC TRAITS (LEARNING)

AND VOTING

LEARNING. THE TRAIT IS REPRESENTED BY THE MATRIX

$$A = \begin{pmatrix} i_1 & i_2 & i_3 \\ \delta_1 & \delta_2 & \delta_3 \end{pmatrix}$$

i_1, i_2, i_3 - INTEGERS, $1 \le i_1 \le i_2 \le i_3 \le l$,

l - LENGTH OF THE BINARY CODE OF THE OBJECTS,

$\delta_j = 0$ or 1.

OBJECT WITH THE CODE $(\omega_1, \omega_2, \ldots, \omega_l)$, $\omega_j = 0$ or 1 HAS THE TRAIT A IF

$$\omega_{i_1} = \delta_1 \ , \ \omega_{i_2} = \delta_2 \ , \ \omega_{i_3} = \delta_3 \ .$$

FOR EXAMPLE THE TRAIT

$$A = \begin{pmatrix} 1 & 3 & 4 \\ 0 & 1 & 0 \end{pmatrix}$$

MEANS THAT THE FIRST AND THE FOURTH DIGITS IN THE CODE OF THE OBJECT ARE 0 AND THE THIRD DIGIT IS 1.

DENOTE:

$K_1(A)$ - THE NUMBER OF OBJECT OF CLASS D_o WHICH HAS THE TRAIT A

$K_2(A)$ - THE NUMBER OF OBJECT OF CLASS N_o WHICH HAS THE TRAIT A.

A IS A CHARACTERISTIC TRAIT OF CLASS D

("D - TRAIT") IF

$K_1(A) \ge K_1$, $\quad K_2(A) \le \widetilde{K}_1$

A IS A CHARACTERISTIC TRAIT OF CLASS N

("N - TRAIT") IF

$K_2(A) \ge K_2$, $\quad K_1(A) \le \widetilde{K}_2$

HERE K_1, $\widetilde{K}_1$, K_2 and $\widetilde{K}_2$ ARE A PRIORI CONSTANS.

CONSIDER TWO "D - TRAITS".

DENOTE:

S_1 - A SET OF OBJECTS OF CLASS D_o WHICH HAS THE FIRST TRAIT.

S_2 - A SET OF OBJECTS OF CLASS D_o WHICH HAS THE SECOND TRAIT.

THE TRAITS ARE EQUIVALENT IF S_1 AND S_2 COINCIDE.

THE FIRST TRAIT IS WEAKER THAN THE SECOND IF S_1 IS A SUBSET OF S_2.

DEFINITION FOR N - TRAITS IS SIMILAR.

THE RESULT OF LEARNING IS SELECTION OF CHARACTERISTIC TRAITS

WITHOUT WEAKER TRAITS AND

WITH ONLY ONE TRAIT FROM EACH GROUP OF EQUIVALENT TRAITS.

VOTING. EACH OBJECT HAS SOME NUMBER n_D OF "D-TRAITS" AND SOME NUMBER n_N of "N-TRAITS"; n_D and n_N ARE ≥ 0. THE OBJECT IS RECOGNIZED AS:

D if $n_D - n_N \geq \Delta$,

N if $n_D - n_N < \Delta$.

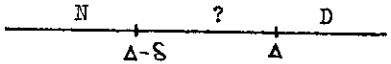
HERE Δ IS A GIVEN CONSTANT

NOTE: IT MAY BE MORE RELIABLE:

TO ASSIGN TO N ONLY THE OBJECTS WITH $n_D - n_N \leq \Delta - \delta$;

and

TO LEAVE UNASSIGNED THE OBJECTS WITH $\Delta > n_D - n_N > \Delta - \delta$



THE CLASSIFICATION DEPENDS ON THE CHOICE OF 5 PARAMETERS: K_1, $\tilde{K}_1$, K_2, $\tilde{K}_2$ and Δ .

2. LEARNING ALGORITHM (H)

CONSISTS OF TWO SIMILAR STEPS - DETERMINATION OF A KERNEL (LEARNING) AND VOTING.

LEARNING. EACH OBJECT IS A BINARY VECTOR. KERNEL IS A BINARY VECTOR OF THE SAME LENGTH; EACH COMPONENT OF THIS VECTOR IS "TYPICAL" FOR CLASS D.

DENOTE:

$q_D(i \mid 0)$ THE NUMBER OF OBJECTS OF CLASS D_o WHICH HAVE $\omega_i = 0$,

$q_D(i \mid 1)$ THE NUMBER OF OBJECTS OF CLASS D_o WHICH HAVE $\omega_i = 1$,

$q_N(i \mid 0)$ THE NUMBER OF OBJECTS OF CLASS N_o WHICH HAVE $\omega_i = 0$,

$q_N(i \mid 1)$ THE NUMBER OF OBJECTS CLASS N_o WHICH HAVE $\omega_i = 1$

LET US CONSIDER EACH COMPONENT IN TURN. COUNT RELATIVE NUMBER OF OBJECTS WHICH HAVE THIS COMPONENT EQUAL TO 1 AMONG CLASS D_o.

$$\alpha_D(i \mid 1) = \frac{q_D(i \mid 1)}{q_D(i \mid 0) + q_D(i \mid 1)}$$ AND

AMONG CLASS N_o

$$\alpha_N(i \mid 1) = \frac{q_N(i \mid 1)}{q_N(i \mid 0) + q_N(i \mid 1)} ,$$

$i = 1, 2, \ldots, l.$

KERNEL OF CLASS D $K = (K_1, K_2, \ldots, K_l)$ IS DEFINED AS FOLLOWS:

$$k_i = \begin{cases} 1, & \text{if } \alpha_D(i \mid 1) \geq \alpha_N(i \mid 1), \\ 0, & \text{if } \alpha_D(i \mid 1) < \alpha_N(i \mid 1) \end{cases}$$

NOTE: IT MAY BE MORE RELIABLE TO ELIMINATE THE PARAMETERS FOR WHICH

$$|\alpha_D(i \mid 1) - \alpha_N(i \mid 1)| < \varepsilon ,$$

WHERE ε IS A SMALL CONSTANT.

VOTING. HAMMING's DISTANCE FROM AN OBJECT TO THE KERNEL OF CLASS D IS

$$\rho = \sum_{i=1}^{l} p_i |\omega_i - k_i| .$$

HERE p_i ARE THE WEIGHTS OF COMPONENTS

WEIGHT p_i CAN BE ANY NUMBER BETWEEN 0 and 1;

IN PARTICULAR (BUT NOT NECESSARILY) IT CAN BE CALCULATED AS:

$$p_i = \frac{|\alpha_D(i \mid 1) - \alpha_N(i \mid 1)|}{\max\limits_{1 \le i \le l} |\alpha_D(i \mid 1) - \alpha_N(i \mid 1)|}$$

OBJECT IS RECOGNIZED, AS AN OBJECT

OF CLASS D IF $\rho \le R$

OF CLASS N IF $\rho > R$

GENERAL COMMENTS TO ALGORITHMS C AND H

CLASSIFICATION DESERVES NUMERICAL TEST IF:

MAJORITY OF OBJECTS D_o IS RECOGNIZED AS D.

$$\frac{|D|}{|W|} \quad \text{or} \quad \frac{|D \wedge N_o|}{|N_o|} \quad \text{or} \quad \frac{|D \wedge (D_o \cup N_o)|}{|D_o \cup N_o|}$$

IS NOT TOO LARGE (SAY β , β IS AN A PRIORI CONSTANT)

$|\,|$ IS THE NUMBER OF OBJECTS IN A SET.

ADDITIONAL CONDITION: ONLY FEW OBJECTS HAVE ρ CLOSE TO R (FOR ALGORITHM H) OR $n_D - n_N$ CLOSE TO Δ (FOR ALGORITHM C), THE RECOGNITION RULE ALSO HAS TO MAKE SOME PHYSICAL SENSE; BUT DO NOT IMPOSE ON IT ALL COMMON BELIEVES - PREPARE TO ACCEPT THE UNEXPECTED.

ALGORITHM H IS SIMPLER AND SHOULD BE TRIED FIRST.

STABILITY OF THE RESULTS SHOULD BE TESTED BY VARIATIONS OF COMPUTATIONAL PARAMETERS

R and P_i (i = 1,2, ..., l) for ALGORITHM H,

K_1 $\widetilde{K}_1$, K_2, $\widetilde{K}_2$ and Δ FOR ALGORITHM C.

ON THE CHOICE OF K_1, $\widetilde{K}_1$, K_2 and $\widetilde{K}_2$: THEY SHOULD LEAD TO A COMPARABLE NUMBER OF D - and N - TRAITS (7 - 15 EACH).

IF NO SATISFACTORY CLASSIFICATION IS OBTAINED, ONE CAN TRY TO CHANGE: DISCRETIZATION (AND ACCORDINGLY CODING); OR LEARNING MATERIAL; OR QUESTIONAIRE; OR DEFINITION OF OBJECTS.

BUT EACH CHANGE INCREASES THE DANGER OF RANDOM DIVISION AND ACCORDINGLY WILL DEMAND MORE NUMERICAL TESTS.

STAGE 7: NUMERICAL TESTS

THE TESTS INCLUDE SOME VARIATION OF THE OBJECTS, QUESTIONAIRE, NUMERICAL PARAMETERS ETC.

THE TEST IS POSITIVE IF THE RESULTS OF RECOGNITION ARE STABLE TO THESE VARIATIONS

SOME USUAL TESTS:

i) ELIMINATION OF OBJECTS FROM D_o and N_o - ONE AT A TIME. FORMAL CRITERIA OF STABILITY - SMALL CHANGE OF THE RATIO $\frac{m_D}{D_o}$ or $\frac{m_D + m_N}{D_o \quad N_o}$, HERE m_D AND m_N SHOW HOW MANY OBJECTS OF D_o and N_o RESPECTIVELY CHANGE THE CLASSIFICATION AFTER THEY WERE ELIMINATED FROM LEARNING.

ii) LEARNING ON THE SUBSETS OF THE OBTAINED SETS D and N.

iii) EH (EARTHQUAKE HISTORY): TRANSFER FROM D_o to N_o THE OBJECTS WHICH CORRESPOND TO THE LAST STRONG EARTHQUAKE AND HAVE NO OTHER STRONG EARTHQUAKES INSIDE; THEN TO THE TWO LAST EARTHQUAKES etc

iv) CHANGE OF THE QUESTIONAIRE, IN PARTICULAR ELIMINATION OF EACH QUESTION IN TURN.

USEFUL EXAMPLES OF NUMERICAL TEST CAN BE FOUND in [1].

STAGE 8: TRANSFER OF CRITERIA TO ANOTHER REGION

THE TEST CONSISTS OF APPLICATION OF EXACTLY THE SAME RECOGNITION RULES TO ANOTHER REGION. STAGES I-5 HAVE TO BE MADE FOR THIS REGION INDEPENDENTLY. HOWEVER THE CHARACTERISTIC TRAITS SHOULD BE APPLIED WITHOUT ANY CHANGES.

STAGE 9: TEST ON RANDOMIZED DATA

WHILE MAKING RANDOMIZATION IT SHOULD BE KEPT IN MIND THAT PARAMETERS MAY BE DISTRIBUTED NON-HOMOGENIOUSLY and/or MAY BE CORRELATED. ONE OF THE WAYS TO ALLOW FOR THIS FACT IS TO RANDOMIZE ONLY THE DISTRIBUTION OF OBJECTS BETWEEN D_o AND N_o.

WARNING: ALL THESE TESTS ILLUSTRATE - HOW RELIABLE ARE THE RESULTS OF THE PATTERN RECOGNITION. HOWEVER THEY DO NOT PROVIDE A PROOF IN THE STRICT STATISTICAL SENSE, SINCE THE LEARNING MATERIAL (THE CATALOG OF STRONG EARTHQUAKES) IS ALWAYS SMALL; THE DANGER OF SELFDECEPTION IS NOT COMPLETELY ELIMINATED. THAT IS WHY THE DESIGN AND IMPLEMENTATION OF NEW TESTS SHOULD BE PURSUED. STILL THESE TESTS ARE THE BEST AVAILABLE BASE FOR DECISION-MAKING (NOT SPEAKING OF A FINAL PROOF ON FUTURE EARTHQUAKES, WHICH MAY TAKE MANY YEARS).

TEST ON FUTURE HISTORY.

THE HISTORY IS LONGEST FOR THE CENTRAL ASIA, WHICH WAS STUDIED FIRST (THE RESULTS WERE PUBLISHED IN 1972). SINCE THAT TIME 4 EARTHQUAKES WITH $M \geq M_o$ ($M_o = 6.5$) HAD OCCURED IN THE REGION. THE EPICENTERS OF ALL THESE EARTHQUAKES ARE SITUATED WITHIN THE OBJECTS RECOGNIZED AS B; TWO OF THESE OBJECTS DID NOT BELONG TO D_o (i.e. STRONG EARTHQUAKES HAVE BEEN UNKNOWN THERE BEFORE).