# The Problems of Seismological Data Mining via Internet[1]

*Gitis V.G., Weinstock A.P.*
Institute for Information Transmission Problems,
Russian Academy of Sciences,
B.Karetnyi Lane, 19, 101447, Moscow, GSP-4, RUSSIA,
e-mail: gitis@iitp.ru,
http://gis.iitp.ru

## 1. Introduction

Internet opens wide possibilities for specialists to access the remote databases. Due to considerable interest of society to ecology and its importance for vital activity it is reasonable to supply Internet with databases on regional natural hazards and information tools for processing and analysis of geological and geophysical data. The tools have to help the specialists in spatial data mining of natural phenomena and processes.

GeoProcessor[2] [1] is an analytical Web GIS for presentation, modelling and analysis of environmental and in particularly seismological data. The system supports remote access to geological, geophysical and geographical databases and processing, modelling, analysis and spatial data mining. It helps to estimate and detect properties of geological environment using the set of the plausible inference methods, such as: similarity with precedents, similarity with expert expressions in fuzzy logic constructions, membership functions, and nonparametric regression.

The motivation of this paper is development of the seismological problem domain for information technology in scope the IST Project SPIN! [2]. The paper outlines the basic problems of seismic impact forecasting, gives specification of WWW users of seismological GI, considers four basic methods of data mining referring to seismological applications, and discusses the peculiarities of seismological data mining in conclusion.

## 2. Seismological problem domain

### 2.1. General methodology

There are three principal problems in seismic impact forecasting:
1. Seismic hazard assessment;
2. Earthquake prediction.
3. Induced hazard assessment.

General methodology for the solution of these problems consists of three main steps:
(i) Assessment of the relationship between seismo-tectonic attributes or between natural and man-made objects;

[2] http://www.iitp.ru/projects/geo, http://gis.iitp.ru/, http://borneo.gmd.de/and/geoprocessor

(ii)  Prediction of the target seismological attributes or detection of the target seismological objects;

(iii)  Representation of the result for cartographic exploration.

## 2.2.  Seismic hazard assessment

Seismic hazard assessment consists in two problems: estimation of seismic regime parameters and estimation of seismic shakeability (Fig. 1). Seismic regime according to Gutenberg-Richter model [3] is determined by the following 3 parameters: intensity of seismic flow, coefficient of decreasing linear relationship between the logarithm of a number of events and their magnitudes, and maximal possible magnitude ($M_{max}$) of expected earthquake. Shakeability defines seismic impact on the earth surface in the MSK-64, EMS-98 [4] intensity scales or in acceleration scale. Shakeability at the point is calculated as a sum of seismic impacts of all seismic sources taking into account spatial damping of seismic energy with distances from seismic sources.

The following initial data are used for seismic hazard assessment: earthquake catalogue, data about state and tendencies of evolution of geological environment, expert knowledge/hypotheses about the geodynamic regional models and expert solutions.

The most difficult and important problem is estimation of $M_{max}$ spatial distribution [5, 6]. Complexity of the problem is necessity to estimate rare and extreme event under the condition of incomplete information. Economy and social consequences due to underestimation or overestimation of $M_{max}$ define importance of the problem. Problem solution is based on historical facts about earthquake repetition and on suggestion about relationship between $M_{max}$ and seismo-tectonic earth crust properties, which are very slowly changed in time.
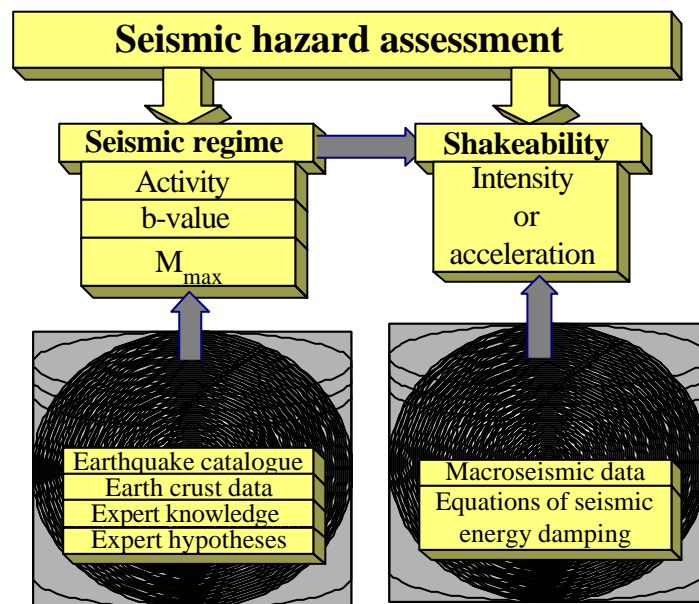


Fig. 1. Sesimic hazard assessment

## 2.3.  Earthquake prediction

The problem [7, 8, 9, 10, 11, 12, 13] consists in detection of effective earthquake precursors and in using the precursors for estimation of time, place and energy of expected earthquake

(Fig.2). There are no effective solutions of this problem till nowadays in spite of the great common activity of specialists of different disciplines. Earthquake catalogues, geo-monitoring time series, data about stationary seismo-tectonic properties of geological environment, and expert knowledge and hypotheses about earthquake precursors are used to solve this problem. Complexity of the problem consists in the great uncertainty in the model of earthquake preparation process, limited volume of the measurements, high level of noise in data, and presence of different effects, which trigger the earthquakes.
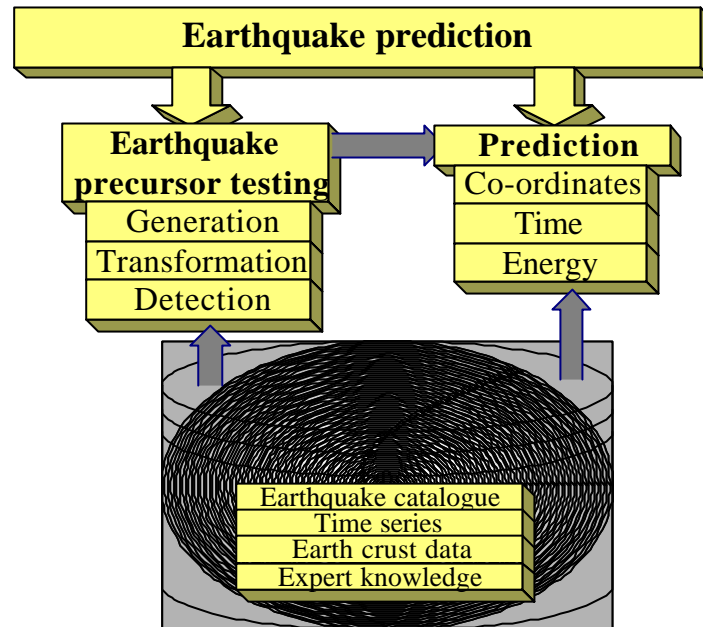


Fig. 2. Earthquake prediction

## 2.4. Induced hazard assessment

Earthquake can induce another natural and man-made catastrophes (Fig. 3). The model of induced hazard assessment was proposed in [14, 15]. It considers a system that consists of interacting subsystems, such as different elements of the environment, economical structures, or nodes of energetic or information networks. Each of these subsystems can be in several states: the normal state and a number of states representing some degree of damage. It is supposed that damage in a certain subsystem can cause damage in another one. The model allows to incorporate the following data: a set of hazard objects, expert knowledge about possible damage states of the objects, expert knowledge about paired links between the objects, estimations of probabilities of spontaneous catastrophes, the expert evaluations of probabilities for one catastrophe to induce directly another catastrophe. The model is presented as the oriented graph nodes of which are all possible catastrophes and arrows are probabilities of induced catastrophes.
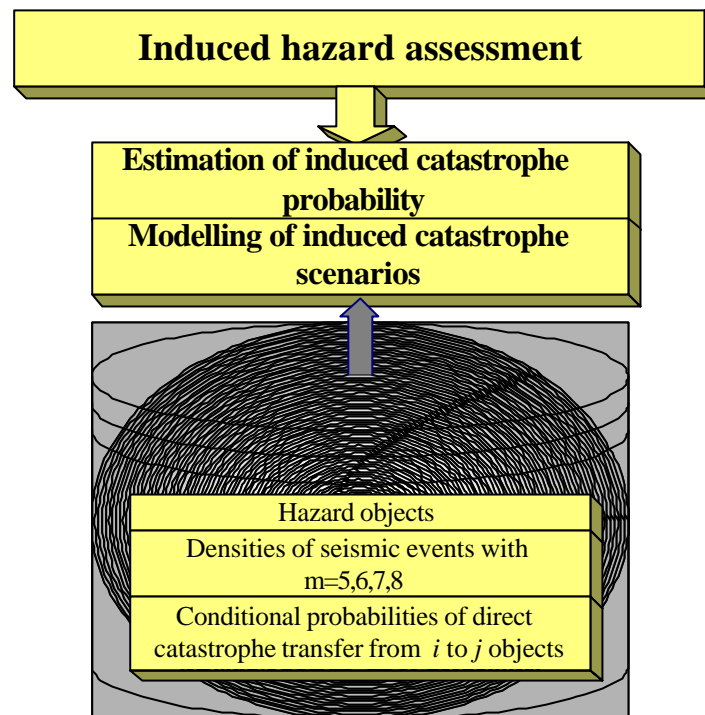
Fig. 3. Induced hazard assessment

## 3. Users of seismological Web GIS

Seismological Web GIS [17] have two classes of users: seismological GI suppliers and seismological GI consumers.

The objectives of seismological GI suppliers are the following:
1. Dissemination of knowledge on historical and current seismicity.
2. Training of actions on seismic disaster mitigation for individuals.
3. Support of administrative solutions for seismic risk mitigation.
4. Support of scientific research.

GI consumers can be divided in four groups: citizens, administration, students, and experts. The queries of seismological GI consumers are the following:

Citizens:
1. Current and historical seismicity
2. Documents on seismic hazard and seismic risk mitigation for individuals

Administration:
1. Seismic hazard and induced hazard assessment
2. Actions on seismic risk mitigation and insurance policy

Students:
1. Training in seismic hazards assessment
2. Training in earthquake precursor analysis
3. Training in estimation of seismological parameters

Experts:
1. Seismic hazard assessment and developing the seismic zone maps
2. Investigation of new earthquake precursors
3. Earthquake prediction in real time

4. Induced hazard assessment
5. Seismic risk assessment

# 4. Spatio-temporal seismic data analysis.

## 4.1. Seismological data and types of data analysis

Seismological database contains information about seismological and seismo-tectonic entities, attributes and relationships between entities and attributes. Entities are represented by geographical objects and 2D or 3D grid data. The objects include the polygons for representation of geological and administrative zones, the points for representation of earthquake catalogues, cities, dangerous man-made objects, seismological stations, the lines for representation of geological faults, lineaments, roads, topographical elements, the network of geo-monitoring data stations for measurement of seismo-tectonic time series. Grid data are used for spatial and spatio-temporal presentation of seismo-tectonic properties of geological media.

The following four types of GI analysis are used in seismological GIS for solution of the problems specified in the previous part of the paper.
- Estimation of the relationships between the attributes.
- Estimation of the relationships between the objects.
- Estimation and prediction of the target GI attribute.
- Detection of the target GI objects.

## 4.2. Estimation of the relationships between the attributes

There are two basic methods for analysing the relationships between the seismo-tectonic attributes:
(i) Correlation and factor analysis of the attributes.
(ii) Statistical and logical inference of the relationships between the attributes.

Correlation and factor analysis help to discover the peculiarities of seismo-tectonic structure and tendency of evolution for the region under study. Plausible inference methods are aimed to determine the relationships, which could be applied for the solving the forecast problems.

An example of the inferred relationship between the maximal magnitude of expected earthquake $M_{\max}$ and geological and geophysical attributes for the Caucasus region [5] is presented in graphical form in Fig. 4. The relationship is $F(\mathbf{x}) = 4.1 + \varphi_1(x_1) + \varphi_2(x_2) + \varphi_3(x_3)$,
where:
$\varphi_1(x_1)$, $\varphi_2(x_2)$, $\varphi_3(x_3)$ are increasing piece-linear functions;
$x_1$ is a spatial attribute of the faults, which is equal to half sum of the closeness to thrusts active in Cenozoic period ($y_1$) and closeness to strike-slip faults ($y_2$) active in the same period $x_{1t} = (y_{1t} + y_{2t})/2$, $y_{it} = 100(1 - r_t/R)$, $i$=1,2, $t$ is the raster point number, $r_t$ is the distance from the grid point $t$ to the nearest thrust (fault), $R = 50$ km;
$x_2$ is the absolute value of the post-Sarmatian vertical tectonic movement velocity gradient in $10^{-9}$/year;

$x_3$ is the anomaly of the upper mantle $P$-wave run time variation in seconds,
$x_3(\lambda,\varphi) = z(\lambda,\varphi) - \dfrac{1}{\pi R^2} \iint\limits_{u^2+v^2 \leq R^2} z(\lambda-v,\varphi-u)dudv$, where $z(\lambda,\varphi)$ is the upper mantle P-wave run time variation in the point with geographic co-ordinates $\lambda$ and $\varphi$, $R = 30$ km.
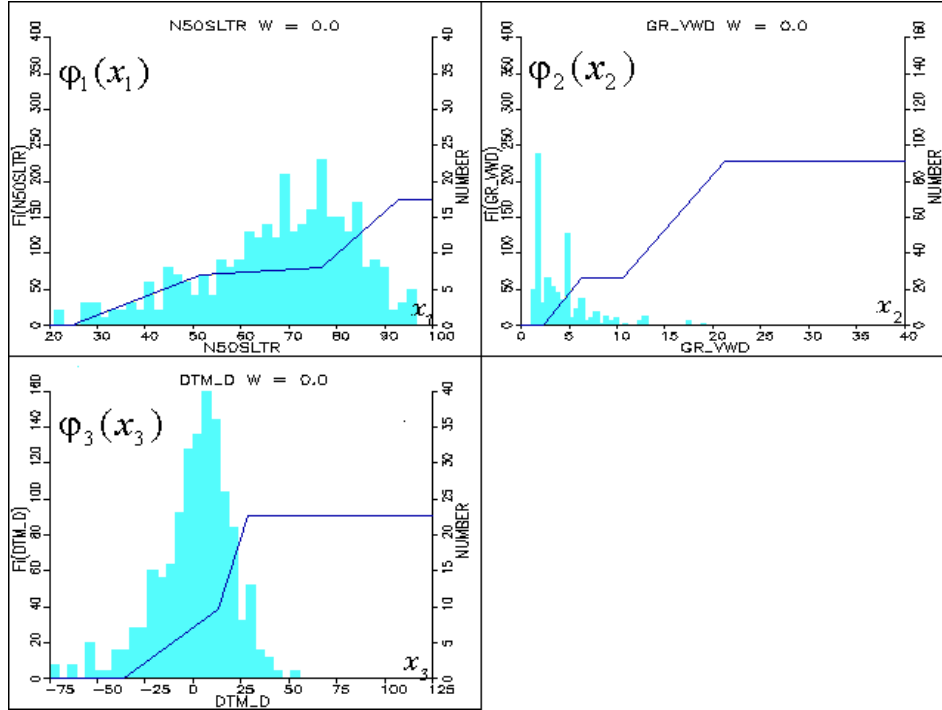


Fig. 4. Relationship between $M_{max}$ and thematic attributes for the Caucasus region,
$$F(\mathbf{x}) = 4.1 + \varphi_1(x_1) + \varphi_2(x_2) + \varphi_3(x_3)$$

### 4.3.    Estimation of the relationships between the objects

Three methods of analysis the relationships between the seismo-tectonic objects are the most typical:
(i)    Object classification and clustering.
(ii)   Estimation of geometric properties of seismic process.
(iii)  Induced hazard assessment (modelling of scenarios and estimation of probability of induced catastrophes).

Method of classification and cluster analysis are usually apply to select out homogeneous groups of earthquakes or to analyse earthquake migration. Geometric properties of seismic process describe spatio-temporal variations of the relationship between the earthquakes. The parameter of fractal dimension of the earthquakes (D-value) is usually measured for this aim. D-value is close to 1 in the area where earthquakes form the linear groups and it is close to 2 in the case of diffusive spatial earthquake distribution. The methods of induced hazard assessment use the paired links between natural and man-made objects in order to model scenarios of catastrophe evolution.

An example of catastrophe chain induced directly by the earthquake is presented in Fig. 5. The earthquakes with energy $e_1$, $e_2$, $e_3$ induce the landslides with body rocks $m_1$ and $m_2$. The

landslides create dams and water body. Probabilities of earthquakes are given by vector $\mathbf{p}^T=(\pi_1, \pi_2, \pi_3)$. Probabilities of $e_i$ earthquake induces a landslide $m_j$ is given by the 3x2 matrix $\mathbf{P}_{em}=(P(e_i,m_j)$. Two catastrophes of water body are possible: dam overfilling or dam crush. Let us $\mathbf{D}_o(m)$ is 2x2 diagonal matrix, probability of dam overfilling after the landslide $m_j$; probability of not overfilling for the dam is $\mathbf{D}_n(m)=\mathbf{E}-\mathbf{D}_o(m)$; probability of mudflow after overfilling is $\mathbf{Q}_o^T(m)=(\mathbf{Q}_o(m_1),\mathbf{Q}_o(m_2))$; probability of mudflow after the dam crush is $\mathbf{Q}_n^T(m)=(\mathbf{Q}_n(m_1), \mathbf{Q}_n(m_2))$; $U$ is probability of the structural damage after mudflow. Then the probability of indirect structural damage after the earthquake in matrix notation is given by the following expression:

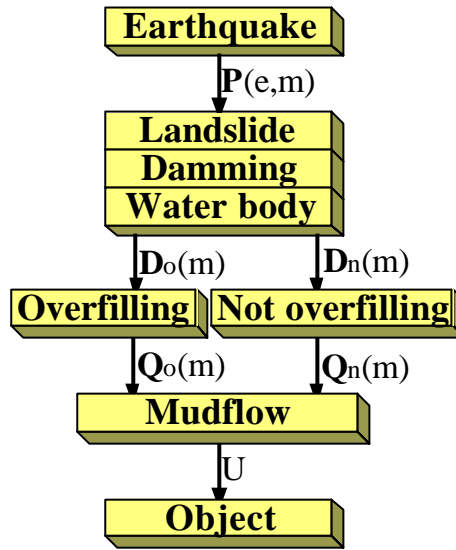$$L=\mathbf{p}^T\mathbf{P}_{em}(\mathbf{D}_o\,\mathbf{Q}_o+\mathbf{D}_n\,\mathbf{Q}_n)U$$



Fig. 5. An example of chain of catastrophes

## 4.4.   Estimation and prediction of the target GI attributes

(i)   Generation of new geographical object attributes or layers.
(ii)  Estimation of spatio-temporal distribution of seismo-tectonic attributes.
(iii) Forecasting of the seismic parameters.

Generation of new geographical object attributes or layers is carried out by the attributes and geometry of another layers. For example, it is possible to calculate seismological attributes of the polygon or point by the earthquake epicentres located at the polygon or point buffer zones.

An example of estimation of spatio-temporal distribution of seismo-tectonic attribute is the variation of minimal representative earthquake magnitude $M_{min}$ for the East Mediterranean region. The result presented in Fig. 6 was obtained in cooperation with G. Papadopoulos in scope of EC Copernicus Project ASPELEA (contract IC 15 CT97 0200). The less $M_{min}$ then the sensitivity of seismic network is higher. The variations of $M_{min}$ manifested that sensitivity of seismic network was significantly improved firstly in 1988-1991 years and the second time in 1997-2000 in the north-west area of the region.
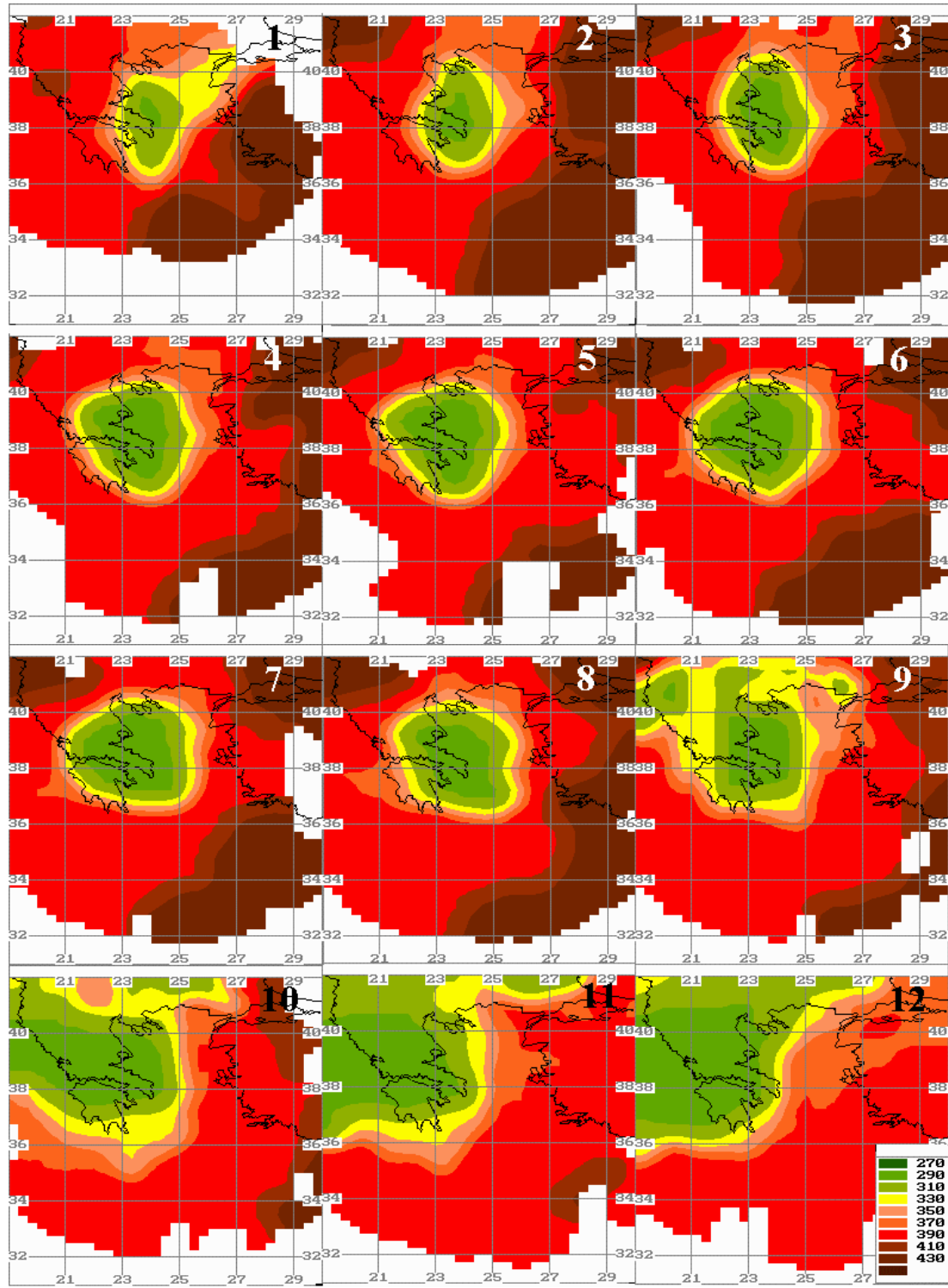
Fig. 6. The sections of 3D raster of $M_{\min}$ variation from 1967 to 2000 with 3-year interval

Another example refers to forecasting the properties of seismic process. It is an application of the RTL criterion [8] for detection of earthquake precursors made in scope of ASPELEA project. Cleaned from aftershocks earthquake catalogue with 1551 events occurred from 01.01.1964 till 06.09.1999 (one day before the earthquake) within the circle with 100 km radius and the centre in the Athens earthquake, $M$=5.9, 7.09.1999, latitude=38.15, longitude=23.62 was analysed. The RTL time series is presented in Fig. 7. The significant

positive anomaly before the Athens earthquake detects a set of foreshocks, which could be interpreted as earthquake precursor.
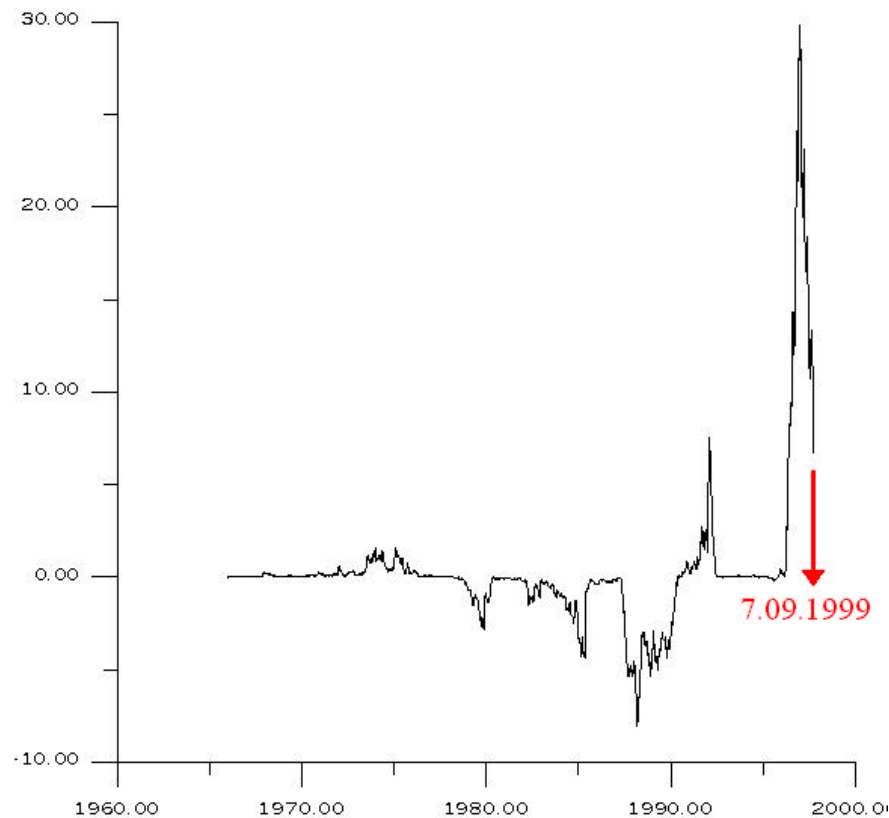


Fig. 7. Estimation of time series for the Athens earthquake, m=5.9, 7.09.1999, lat=38.15, long=23.62, by RTL -criterion

## 4.5.    Detection of the target GI objects

Two types of seismological object detection can be considered:
(i)    Delineation of seismic source zones.
(ii)   Earthquake prediction.

In the fist case the problem consists in estimation of seismic source geometry by all available seismological, geophysical and geological data. In the second case the predictive problem consisting in assessing the time, coordinates and earthquake energy is solved with the help of preceding data only.

An example of earthquake prediction made retrospectively refers to Tangshan earthquake with magnitude $M$=7.8 occurred at north-eastern China on July 28, 1976 [13]. Daily time series of geophysical and hydrogeological parameters measured at 10 stations based on the observation period ranging from the beginning of 1972 till July 27, 1976 (24 hours before the Tangshan earthquake) were analysed. Four sections of 3D grid data corresponding to earthquake precursor and the seismic stations shown by triangles are presented in Fig 8. It also shows tectonic faults and epicentres of the future Tangshan earthquake and of the two strongest aftershocks with magnitudes 6.9 and 7.0. The anomaly corresponding to earthquake

9

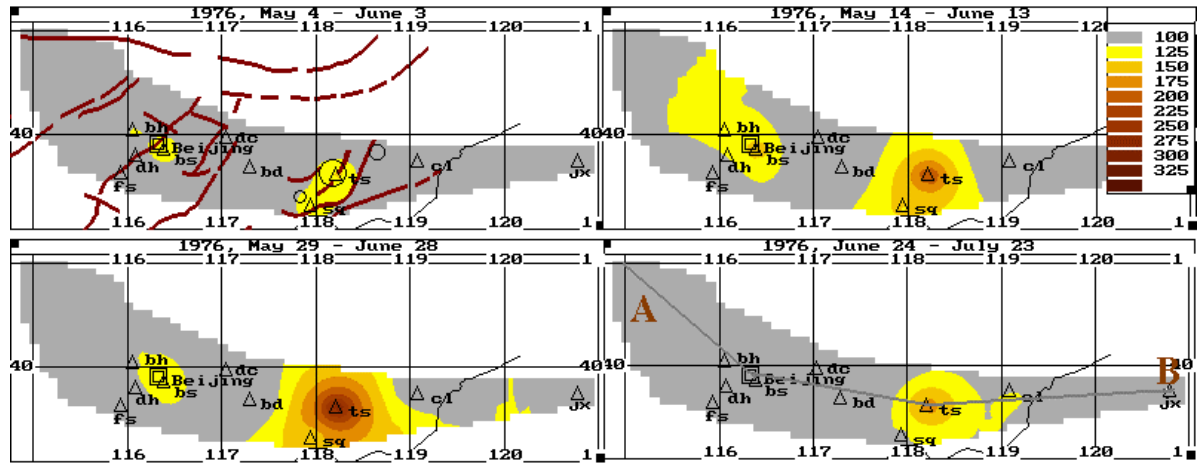precursor is evolved from Beijing and Tangshan areas and located at the vicinity of Tangshan epicentre.



Fig. 8. Precursor evolution in 60-7 days before the Tangshan earthquake, China, M=7.8, 28.07.1976. The value of anomaly is measured in values of standard deviations of stationary process  multiplied by 100

## 5.  Some peculiarities of seismological data mining

(i)      Two types of seismological predictive data mining are considered:

      A)   Prediction based on spatial and spatio-temporal geodata at the vicinity of the target object under study:  $y(\lambda, \varphi, t) = F(\mathbf{x}(\lambda, \varphi, t))$

      B)   Prediction based on the ratio between the objects:

$$y(n) = F(q_1, q_2, ..., q_N, u_{12}, ..., u_{nm}, ..., u_{(N-1), N})$$

(ii)     There are the following peculiarities in seismological predictive data mining:

      A)   Incomplete information (incomplete data, absence of mathematical model)

      B)   Uncertainty in data and expert knowledge (noise in data, ambiguity in expert knowledge)

(iii)    There are the following peculiarities in seismological GI:

      A)   Many types of spatio-temporal data (grid data, points, lines, polygons, images, time series, documents)

      B)   Huge earthquake catalogues involving up to tens and hundreds thousand events

(iv)    Principle peculiarity of seismological GIS architecture consists in integration of vector and grid data processing. This integration is essential for three following types of operations:

      A)   Estimation of spatio-temporal properties of geographical objects with the help transformations their geometric and thematic data into 2D and 3D grid data

      B)   Estimation of the attributes of geographical objects using evaluation of the functions of 2D and 3D grid data within the geographical object buffer zones

      C)   Detection of target geographical objects from 2D and 3D grid data

# References

1. Gitis, V., Dovgyallo, A., Osher, B., Gergely, T. GeoNet: an information technology for WWW on-line intelligent Geodata analysis. // Proceedings of the 4th EC-GIS Workshop. Hungary. 1998. JRC of EC. P. 124-135.
2. May, M. Spatial Knowledge Discovery: The SPIN! System // Proceedings of the 6th EC-GIS Workshop. Lyon, France. 2000. JRC of EC.
3. B. Gutenberg & C. H. Richter. Earthquake magnitude, intensity, energy, and accelaration, Bull. Seism. Soc. Am. 46, 105-145 (1956).
4. European Seismological Commission. Activity Report 1996-1998. Proceedings XXVI General Assembly, August 24-28, 1998. Tel Aviv, Israel: Geophysical Institute of Israel. 58 p.
5. Gitis, V. GIS Technology for the Design of Computer-Based Models in Seismic Hazard Assessment // Geographical Information Systems is Assessing Natural Hazards, A.Carrara and F Guzzetti (eds.). 1995. Kluver Academic Publishers. P. 219-233
6. Gitis, V., Vainchtok, A., Tatevosjan, R. Maximum expected magnitude assessment in GEO computer environment: case study // Natural Hazards 17. 1998. Kluver Academic Publishers. Netherlands. P. 225-250.
7. European Macroseismic Scale 1998 (EMS-98). Ed. G.Grunthal // Cahiers du Centre Europeen de Geodynamique et de Seismologie. Vol. 15. 1998 97 p.
8. Keilis-Borok, V.I. Intermediate-term earthquake prediction // Proc. Nat l. Acad. Sci. USA, Vol. 93. April 1996. P. 3748-3755.
9. Sobolev, G.A., Zavialov, A.D. Earthquake prediction. A concentration criterion for seismicically active faults. Earthquake prediction. National Report of IASPEI and IUGG, 1995 – 1998. M.: Geophysical Committee of RAS. 1999. C. 22-27.
10. Wyss, M., Console, R., Muggu, M. Seismicity rate change before the Iprina (M=6.9) 1980 earthquake // Bull. Seism. Soc. Amer., 1997, V.87. P. 318-326.
11. Zhang, Guomin, Zhang, Zhaocheng. The study of multidisciplinary earthquake prediction in China // Journal of Earthquake Prediction Res. 1992. V. 1, ? 1. H. 71-86.
12. Zschau, J. et al SEISMOLAP: A new approach to prediction // Proceeding of the International Conference on Earthquake prediction: state of art. Council of Europe, Strasbourg. 1996. P. 444-453.
13. Gitis, V.G., Osher, B.V., Pirogov, S.A., Ponomarev, A.V., Sobolev, G.A., Jurkov, E.F., A System for Analysis of Geological Catastrophe Precursors // Journal of Earthquake Prediction Research. Vol. 3. 1994. P. 540-555.
14. Ponomarev A.V., Sobolev G.A., Gitis V.G., Zhang Zchaocheng, Wang Guixuan, Qin Xinxi, Complex analysis of geophysical fields for detection of spatio-temporal earthquake precursors. Electronic Journal of UIPhE RAS, 4 (10). 1999. URL: http://www.scgis.ru/russian/cp1251/h_dgggms/4-99/komp-an.zip.
15. Gitis, V.G., Petrova, E.N., Pirogov, S.A. Catastrophe Chains: Hazard Assessment // Natural Hazards. 10. 1994. P. 117-127.
16. Gitis, V.G., Petrova, E.N., Pirogov, S.A. Expert knowledge approach to catastrophe Chains // Cahiers du Centre Europeen de Geodynamique et de Seismologie. Vol. 12. 1996. P. 67-72.
17. M. Mattarelli, C. Boehner and R.J. Peckham. Web Access to Earthquake Catalogues. Proc. EOGEO98. Salzburg, Feb. 1998. http://www.sbg.ac.at/geo/eogeo/authors/mattarelli/mattarelli.htm