Earthquake forecasting: statistics and information

Vladimir Gertsik[1,*], Mark Kelbert[1,2,3], Anatoly Krichevets[4]

[1] *Institute of Earthquake Prediction Theory and Mathematical Geophysics RAS, Moscow, Russia*

[2] *Swansea University, Department of Mathematics, Swansea, United Kingdom*

[3] *Higher School of Economics, Moscow, Russia*

[4] *Lomonosov MSU, Department of Psychology, Moscow, Russia*

ABSTRACT

The paper presents a decision rule forming a mathematical basis of earthquake forecasting problem. We develop an axiomatic approach to earthquake forecasting in terms of multicomponent random fields on a lattice. This approach provides a method for constructing point estimates and confidence intervals for conditional probabilities of strong earthquakes under conditions on the levels of precursors. Also, it provides an approach for setting a multilevel alarm system and hypothesis testing for binary alarms. We use a method of comparison for different algorithms of earthquake forecasts in terms of the increase of Shannon information. 'Forecasting' (the calculation of the probabilities) and 'prediction' (the alarm declaring) of earthquakes are equivalent in this approach.

1. Introduction

The methodology of selecting and processing of relevant information about the future occurrence of potentially damaging earthquakes has reached a reasonable level of maturity over the recent years. However, the problem as a whole still lacks a comprehensive and generally accepted solution. Further efforts for optimization of the methodology of forecasting would be productive and well justified.

One of our objectives here is to address some drawbacks of the approach to earthquake prediction based on pattern recognition theory. It is developed by "Keilis-Borok team" and presented in Keilis-Borok and Soloviev [2003]. The pattern recognition is used in forecast with the hope that 'training sets' in which strong earthquakes have higher values of precursors may help to decide whether or not a strong earthquake will happen in the future. As it is demonstrated by below published Krichevets's theorem, this hope has no basis.

A comprehensive review of the modern earthquake forecasting state of knowledge and guidelines for utilization can be found in Jordan et al. [2011]. Note

that all methods of evaluating the probabilities of earthquakes are based on a combination of geophysical, geological and probabilistic models and considerations. Even the best and very detailed models used in practice are in fact only 'caricatures' of immensely complicated real processes. (For example, the equations of the theory of elasticity are valid only for infinitesimal strains, but used even at the edges of the cracks where strains tend to infinity. It is also clear that we do not know about the inhomogeneity of elastic properties in the real medium, and we have no way of knowing about it with sufficient accuracy.)

A mathematical toolkit for earthquake forecasting is well presented by Harte and Vere-Jones [2005]. This work is based on the modeling of earthquake sequences in terms of the marked point processes. However, the mathematical technique used is quite sophisticated and does not provide direct practical tools to investigate the relations of the structure of temporal-spatial random fields of precursors to the appearance of strong earthquakes.

The use of the multicomponent lattice models (instead of marked point processes) gives a different novel way of investigating these relations by a more elementary way. Discretization of space and time allows us to separate the problem in question into two separate tasks. The first task is the selection of relevant precursors, i.e., observable and theoretically explained physical and geological facts which are causally related to a high probability of strong earthquakes. Particularly, this task involves the development of models of seismic events and computing probabilities of strong earthquakes in the framework of these models. Such probabilities are used as precursors in the second task.

The second task is the development of methodol-

ogy of working with these precursors in order to extract the maximum information about the probabilities of strong earthquakes. This is the main topic of this paper.

Our approach allows us to obtain the following results:

- Estimates of probabilities of strong earthquakes for given values of precursors are calculated in terms of the frequencies of historic data.

- Confidence intervals are also constructed to provide reasonable bounds of precision for point estimates.

- Methods of predictions (i.e., binary alarm announcement [Keilis-Borok and Kossobokov 1990, Keilis-Borok 1996, Holliday et al. 2007]) and forecasting (i.e., calculating probabilities of earthquakes [Kagan and Jackson 2000, Harte and Vere-Jones 2005, WGNCEP 2007, Jordan et al. 2011]) are equivalent in the following sense: the setting of some threshold for the probability of earthquakes allows to update the alarm level. On the other hand, the knowledge of the alarm domain based on historical data allows us to evaluate the probabilities of earthquakes. In a sense, the prediction is equivalent to hypothesis testing as well (see Section 11).

- In our scheme we propose a scalar statistic which is the ratio of the actual increment of information to the maximal possible increment of information. This statistic allows us to linearly order all possible forecasting algorithms. Nowadays the final judgement about the quality of earthquake forecasting algorithms is left to experts. This arrangement puts the problem outside the scope of natural sciences which are trying to avoid subjective judgements.

The foundation of our proposed scheme is the assumption that the seismic process is random and cannot be described by a purely deterministic model. Indeed, if the seismic process is deterministic then the inaccuracy of the forecast could be explained by the non-completeness of our knowledge about the seismic events and non-precision of the available information. This may explain, at least in principle, attacks from the authorities addressed to geophysicists who failed to predict a damaging earthquake. However, these attacks have no grounds if one accepts that the seismic process is random. At the end of the last century (February-April 1999) a group of leading seismologists organized a debate via the web to form a collective opinion of the scientific community on the topic: 'Is the reliable prediction of individual earthquakes a realistic scientific goal?' (see http://www.nature.com/nature/debates/earthquake/).

Despite a considerable divergence in peripheral issues all experts taking part in the debate agreed on the following main principles:

- the deterministic prediction of an individual earthquake, within sufficiently narrow limits to allow a planned evacuation programme, is an unrealistic goal;

- forecasting of at least some forms of time-dependent seismic hazard can be justified on both physical and observational grounds. The following facts form the basis of our agreement with this point of view. The string-block Burridge-Knopov model, generally accepted as a mathematic tool to demonstrate the power-like Gutenberg-Richter relationship between the magnitude and the number of earthquakes, involves the generators of chaotic behaviour or dynamic stochasticity. In fact, the nonlinearity makes the seismic processes stochastic: a small change in the shift force may lead to completely different consequences. If the force is below the threshold of static friction the block is immovable, if the force exceeds this threshold it starts moving, producing an avalanche of unpredictable size.

This mechanism is widespread in the Earth. Suppose that the front propagation of the earthquake approaches a region of enhanced strength of the rocks. The earthquake magnitude depends on whether this region will be destroyed or remains intact. In the first case the front moves further on, in the second case the earthquake remains localized. So, if the strength of the rocks is below the threshold the first scenario prevails, if it is above the threshold the second scenario is adapted. The whole situation is usually labelled as a butterfly effect: infinitesimally small changes of strength and stress lead to macroscopic consequences which cannot be predicted because this infinitesimal change is below any precision of the measurement. For these reasons determinism of seismic processes looks more doubtful than stochasticity.

The only comment we would like to contribute to this discussion is that the forecasting algorithms based exclusively on the empirical data without consistent physical models could hardly be effective in practice (see Section 13 for more details).

Finally, note that our approach may be well-applicable for the space-time forecasting of different extremal events outside the scope of earthquake prediction.

2. Critical analisis of the prediction based on pattern recognition

The main objective of this paper is to present a mathematical formalism of the earthquake forecasting alternative to the approach based on the pattern recognition theory. So, it is appropriate to summarize the latter approach.

An advantage of the formalism of the Keilis-Borok team [Keilis-Borok and Kossobokov 1990, Keilis-Borok 1996, Keilis-Borok and Solovoev 2003] is a separation of the forecast into two separate tasks: the selection of relevant precursors and the technique of their process-

ing. On the other hand, some drawbacks of the scheme should be pointed out.

2.1. Pattern recognition

The method is relaying on the pattern recognition technique that can provide only empirical but not theoretical foundation. The possibility to base a reliable forecast on the idea of supervised pattern recognition is highly questionable in principle. Indeed, the features selection based on the observations only and not related to the physics of earthquakes is not helpful and any hopes for valid 'training sets' are not grounded. A successful supervised recognition is possible only if the features has proved causal relation with pattern. This principle is illustrated by a simple but important theorem by A.N. Krichevets.

Theorem. *Let A be a finite set, $B_1 \cup B_2 \subset A$; $B_1 \cap B_2 = \emptyset$. We say that B_1 and B_2 are finite educational samples. Let $X \in A$; $X \notin B_1 \cup B_2$ be a new object. Then among all classifications, i.e., subsets $\{A_1, A_2\}$ such that $B_1 \subset A_1$, $B_2 \subset A_2$, $A_1 \cup A_2 = A$, $A_1 \cap A_2 = \emptyset$ satisfying condition that either $B_1 \cup X \subset A_1$ or $B_2 \cup X \subset A_2$ exactly a half classifies X as an object of sample B_1 and a half classifies X as an object from B_2.*

Proof. It is easy to define a one-to-one correspondence between classifications. Indeed, if $\{A_1, A_2\}$ $A_1 \cup A_2 = A$; is a classification such that $B_1 \subset A_1$, $X \in A_1$, $B_2 \subset A_2$, one maps it into the unique classification $\{A_1', A_2'\}$ such that $B_1 \subset A_1'$, $X \in A_2'$, $B_2 \subset A_2'$, where $A_1' = A_1 \setminus X$, $A_2' = A_2 \cup X$.

Corollary. *A supervised pattern recognition is impossible. After the training procedure the probability to classify correctly a new object is the same as before training.*

This result implies that a clairvoyance is involved in a successful forecast: one should know in advance (or guess correctly) a "relevant" set of precursors.

2.2. Voting procedure

As a decision rule for a forecast a voting procedure is adopted: an alarm is announced when the number of precursors taking abnormally large values exceeds certain thershold. This procedure is well-justified only if the pattern in question is unique. Generically, this is not true: if two or more possible scenarios are present, different precursors may be strongly affected in different scenarious. In this case a combination of abnormal values for precursors related to different scenarios may produce a false alarm.

2.3. Alarm domain

The values of precursors are computed at the point of some lattice, and are assigned a domain surrounding the point. The above scheme employs a circle with a center at a lattice site and a radius proportional to the minimal magnitude of the forecasted earthquake. However, this choice leads to a contradiction.

Indeed, suppose we announce an alarm for earthquakes with minimal magnitude 6 in a domain A_6. Note that outside the alarm domain we expected that the earthquake will not occur. But the announcement of the alarm with minimal magnitude 7 is also correct for this point and the alarm should be announced in the domain A_7 as well. By definition $A_6 \subset A_7$ and we expect an earthquake with a magnitude of at least 7 and do not expect an earthquake with à magnitude of at least 6 in the domain $A_7 \setminus A_6$. This is absurd.

Futhermore, the alarm area is too large with this approach.

2.4. Molchan's error diagram

Below we demonstrate that the procedure of alarm announcement is equivalent to the classical problem of hypothesis testing. So, the standard error diagram provides the dependence of the second type error (probability of missing a target) from the first type error (probability of false alarm).

In line with the general case the well-known Molchan's error diagram [Molchan 1990] displays the dependence of the first kind error probability versus the share of alarms.

H_0: the distribution of the precursor is equal to its unconditional distribution (earthquake "could either happen or not happen").

H_1: the distribution of the precursor is equal to its conditional distribution under the condition that the earthquake will happen.

Note that the rejection of hypothesis H_0 leads to absurd results: "earthquake will not happen and will happen".

Of course, Molchan's error diagram can be interpreted not in terms of hypothesis testing. Then there is no contradiction, but the standard error diagram is different from the Molchan's error diagram by a small value proportional to the share of cells occupied by the strong earthquakes (in our approach the spacetime is divided into disjoint cells).

2.5. The loss function, risk function and the economic criterion for the quality of the forecast

The Keilis-Borok team refers to the economic loss function for the quantitative evaluation of the quality of forecast. First, we calculate the expected loss correctly. A natural economic measure for a quality of binary forecast is the economic risk function (it should not be confused with seismic risk, there is the concept of the theory of statistical decisions) or damage r related to the earthquakes and the necessary protective measures. In mathematical statistics the risk function is

defined as the expectation of the loss function, which is an expected losses for different outcomes. In our case there are two types of losses: damage and expenses related to protection. For each cell of our grid (in our approach the space-time is divided into disjoint cells) the risk may be specied by the formula

$$r = \beta \Pr\{\text{no earthquake; alarm}\} + \alpha \Pr\{\text{earthquake; no alarm}\} + \gamma \Pr\{\text{earthquake; alarm}\}$$

here the symbol $\Pr\{\cdot\}$ denotes a probability, α stands for the average damage from a seismic event, β stands for the average expenses for protection after a seismic alarm is announced, γ stands for the average damage after the alarm, $\gamma = \alpha + \beta - \delta$, where δ is the damage prevented by the alarm. The coefficient in front of $\Pr\{\text{no earthquake; no alarm}\}$, obviously, equals 0, because in the absence both of a seismic event and an alarm there is no loss of any kind. Clearly, only the case when $\delta > \beta$ is economically justified, i.e., the gain from the prevention measures is positive. Obviously, δ should be less than $\alpha + \beta$, i.e., an earthquake cannot be profitable. Taking into account that α, β and γ depend on the geographical position of the cell, we write the total risk as the summation over all cells in the region of a given forecast. In the simplest case of the absence of the spacial component, when a single cell represents a region of forecast, the expression for the risk is simplied as follows $r = \alpha\lambda\nu + \beta\tau + \gamma\lambda(1-\nu)$, where τ is a spatio-temporal share of alarm, λ is a share of cells occupied by the earthquakes, ν is the share of missed targets. In Molchan's theory related to the one-dimensional case, the third component is missing.

However, the risk function r, which is very useful for economical considerations and as a basis for an administrative decision, could hardly be used as a criteria for the quality of seismic prediction. First of all, it cannot be computed in a consistent way because the coefficients α, β and γ are not known in practice, and hence no effective way of its numerical evaluation is known. The computation of these coefficients is a difficult economic problem and goes far beyond of the competence of scientists. On the other hand, the readiness of the authority to commit resources to solving the problem depends on the quality of the geophysical forecast. This situation leads to a vicious circle.

The second drawback of the economic risk as a criterion for the quality of prediction is related to the fact that it depends on many factors which have no relation to geophysics or earthquake prediction. These factors include the density of population, the number and size of industrial enterprises, infrastructure, etc. It also depends on subjective factors such as the willingness of authorities to use resources for prevention of the damage from earthquakes. The natural sciences could hardly accept the criteria for the forecast quality which depend on the type of state organization, priorities of ruling parties, results of the recent elections, etc.

In this paper, we introduce a scalar efficiency (information efficiency) forecast that only depends on information about future earthquakes contained in the precursors.

2.6. The measure of an area

In Molchan and Keilis-Borok [2008] the area of the alarm domain is defined in terms of a non-homogeneous measure depending on spacial coordinates. Specifically, the measure of an area is proportional to the number of earthquake epicenters from a sample catalog. This approach is used to eliminate the decrease of the share of alarmed sites τ with the extension of the domain when a purely safe and aseismic territory is included into consideration. It would be well-justied if the quantity τ could be accepted as an adequate criterion of the quality of forecast in its own right. On the other hand, it can be demonstrated that the our information efficiency converges to a non-zero value $1-\nu$ when the number of cells with an alarm is fixed but the total number of cells tends to infinity. An inhomogeneous area of the territory under forecast which is proportional to "activity" does not enable us to calculate the informational efficiency. Moreover, it possesses a number of unnatural features from the view of evaluating economical damage. A seismic event in the territory of low seismicity is more costly because no precautions are taken to prevent the damage of infrastructure. However, in this inhomogeneous area an alarm announced in an aseismic territory will have a smaller contribution than an alarm in a seismically active territory where the losses would be in fact smaller. We conclude that this approach 'hides' the most costly events and does not provide a reasonable estimate of economic damage.

Next, we build a forecasting method free from this kind of flaws.

3. Events and precursors on the lattice

In order to define explicitly estimates of probabilities of strong earthquakes we discretize the two-dimensional physical space and time, i.e., introduce a partition of three-dimensional space-time into rectangular cells with the space partition in the shape of squares and time partition in the shape of intervals. These cells should not intersect to avoid an ambiguity in computing the frequencies for each cell. In fact, the space cells should not be perfect squares because of the curvature of the Earth's surface, but this may be neg-

lected if the region of forecasting is not too large.

So, we obtain a discrete set Ω_K with $N = I \times J \times K$ points which is defined as follows. Let us select a rectangular domain A of the two-dimensional lattice with $I \times J$ points $x = (x_i, y_j)$, $x_i = a \times i$, $i = 1, ..., I$ and $y_j = a \times j$, $j = 1, ..., J$, a is the step of the lattice. A cell in Ω_K takes the shape of parallelepiped of height Δt with a square base. Clearly, any point in Ω_K has coordinates (x_i, y_j, t_k); $t_k = t + k\Delta t$, $k = 0, ..., K$.

We say that a seismic event happens if an earthquake with magnitude greater than some pre-selected threshold M_0 is registered. For any cell in our space-time grid we define an indicator of an event, i.e., a binary function $h(i, j, k)$. This function takes the value 1 if at least one seismic event is registered in a given cell and 0 otherwise. Suppose that for all points (x_i, y_j, t_k) the value of a vector precursor $\mathbf{f}(i, j, k) = \{f_q(i, j, k), q = 1, ..., Q\}$ is given. The components of the precursor $f_q(i, j, k)$, $q = 1, ..., Q$ are the scalar statistics constructed on the base of our understanding of the phenomena that precede a seismic event.

4. Mathematical assumptions

A number of basic assumptions form the foundation of the mathematical technique of earthquake forecasting. In the framework of mathematical theory they can be treated as axioms but are, in fact, an idealization and simplication with respect to the description of the real phenomena. Below we summarize the basic assumptions which are routinely used in existing studies of seismicity and algorithms of earthquake forecasting even the authors do not always formulate them explicitly. We accept the following assumptions or axioms of the mathematical theory:

(i) The multicomponent random process $\{h(i, j, k), \mathbf{f}(i, j, k)\}$, describing the joint evolution of the vector precursors and the indicator of seismic events, is stationary.

This assumption provides an opportunity to investigate the intrinsic relations between the precursors and the seismic events based on the historical data. In other words, the experience obtained by analysing the series of events in the past, is applicable to the future as the properties of the process do not depend on time. In reality, this assumption holds only approximately and for a restricted time period. Indeed, plate tectonics destroys the stationarity for a number of reasons including some purely geometrical considerations. For instance, the movements of the plates leads to their collisions, their partial destruction and also changes their shapes. Nevertheless, the seismic process can be treated as a quasi-stationary one for considerable periods of time. At the time when the system changes from one quasi-stationary regime to another (say, nowadays, many researchers speak about the abrupt climate change) the reliability of any prediction including the forecast of seismic events is severely restricted.

(ii) The multicomponent random process $\{h(i, j, k), \mathbf{f}(i, j, k)\}$ is ergodic.

Any quantitative characteristic of seismicity more representative than a registration of an individual event is, in fact, the result of averaging over time. For instance, the Gutenberg-Richter law, applied to a given region relates the magnitude with the average number of earthquakes where the averaging is taken over a specific time interval. In order to associate with the time averaging a proper probabilistic characteristic of the process and make a forecast about the future one naturally needs the assumption of ergodicity. This exactly means that any averaging over a time interval $[0, T]$ will converge to the stochastic average when $T \to \infty$. In view of ergodicity one can also construct unbiased and consistent estimates of conditional probabilities of strong earthquakes under the condition that the precursors take values in some intervals. Naturally, these estimates are the frequencies of observed earthquakes, i.e., ratios of the number of cells with seismic events and prescribed values of precursors to the total number of cells with the prescribed values of precursors. (Recall that an unbiased point estimate $\hat{\theta}$ of parameter θ satisfies the condition $E[\theta] = \hat{\theta}$ and a consistent estimate converges to the true value θ when the sample size tends to infinity).

(iii) Any statement about the value of the indicator of a seismic event $h(i, j, k)$ in the cell (i, j, k) or its probability should be based on the values of the precursor $\mathbf{f}(i, j, k)$ only.

This assumption means that the precursor in the given cell accumulates all the relevant information about the past and the information about the local properties of the area that may be used for the forecast of the seismic event in this cell. In other words, the best possible precursor is used (which is not always the case in practice). As in the other cases, this assumption is only an approximation to reality, and the quality of a forecast depends on the quality of the selection and accumulation of relevant information in the precursors.

Below we present some corollaries and further specifications.

(iii-a) For any k the random variables $h(i, j, k)$, $i = 1, ..., I, j = 1, ..., J$ are conditionally independent under the condition that the values of any measurable function $u(\mathbf{f}(i, j, k))$ of the precursors $\mathbf{f}(i, j, k)$, $i = 1, ..., I$, $j = 1, ..., J$ are fixed.

In practice this assumption means that the forecast for the time $t_k = t_0 + k\Delta t$ cannot be affected by the values related to the future time intervals $(t_k, t_k + \Delta t]$. In reality

all of these events may be dependent, but our forecast does not use the information from the future after t_k.

(iii-b) The conditional distribution of the random variable h at a given cell depends on the values of the precursors **f** at this cell and is independent of all other variables.

(iii-c) The conditional probabilities $\Pr_{ij}\{h\,|\,u(f)\}$ of the indicator of seismic events h in the cell (i, j, k), under condition $u(f)$ in this cell do not depend on the position of the cell in space (the time index k related to this probability may be dropped due to the stationarity of the process).

In other words, the rule for computing the conditional probability Pr based on the values of precursors is the same for all cells, and the space indices of probability Pr may be dropped. This condition is widely accepted in constructions of the forecasting algoritms but rarely formulated explicitly. However, the probability of a seismic event depends to a large extent on the local properties of the area. Hence, the quality of the forecasting depends on how adequately these properties are summarized in the precursors. This formalism properly describes the space inhomogenuity of the physical space because the stationary joint distribution of $\Pr_{ij}\{h, f \le x\}$ for an arbitrary precursor f depends, in general, on the position of the cell in the domain A. Below we will use the distributions of precursors and indicators of seismic events in domain A that do not depend on the spatial coordinates and have the following form

$$\Pr_A\{h, f \le x\} = \frac{1}{I \times J} \sum_{(i,j) \in A} \Pr_{ij}\{h, f \le x\}$$

$$\Pr_A(x) \equiv \Pr_A\{f \le x\} = \frac{1}{I \times J} \sum_{(i,j) \in A} \Pr_{ij}\{f \le x\}$$

$$p_A \equiv \Pr_A\{h = 1\} = \frac{1}{I \times J} \sum_{(i,j) \in A} \Pr_{ij}\{h = 1\}$$

(iii-d) Note that assumption (iii) implies that the conditional probabilities $\Pr_{ij}\{h\,|\,u(f)\}$ are computed via the probabilities $\Pr_A\{h, f \le x\}$ only.

The properties listed above are sufficient to obtain the point estimates for the conditional probabilities of seismic events under conditions formulated in terms of the values of precursors. However, additional assumptions are required for a testing of the forecasting algorithm:

(iv) The random variables $f(i, j, k)$ are conditionally independent under condition that $h(i, j, k) = 1$.

Again, these conditions are not exactly true, however they may be treated as a reasonable approximation to reality. Indeed, if the threshold M_0 is sufficiently high then the strong earthquakes may be treated as rare events, and the cells where they are observed are far apart with a high probability. Any two events re-

lated to cells separated by the time intervals t are asymptotically independent as $t \to \infty$ because the seismic process has decaying correlations (the mixing property in the language of random processes). The loss of dependence (or decaying memory) is related to the physical phemonema such as healing of the defects in the rocks, relaxation of strength due to viscosity, etc. As usual in physical theories, we accept an idealized model of the real phenomena applying this asymptotic property for large but finite intervals between localizations of seismic events. The independence of strong earthquakes is not a new assumption, in the case of continuous space-time it is equivalent to the assumption that the locations of these events form a Poisson random field. (Note that the distribution of strong earthquake should be homogeneous in space a priori, because by condition there is no information about the heterogeneity outside precursors. The information about the location of epicenters should be contained in the precursors) The Poisson hypothesis is used in many papers, see, e.g. Harte and Vere-Jones [2005]. It is very natural for the analysis of the tails of the Gutenberg-Richter law for large magnitudes [Pisarenko et al. 2008]. Summing up, the development of the strict mathematical theory of earthquake forecasting does not require any additional assumption except those routinely accepted in the existing algorithms but usually not formulated explicitly.

5. The standard form for precursors

The correct solution of the forecasting problem given the values of precursors $f(i, j, k)$ is provided by the estimate of conditional probability $\Pr_{ij}\{h\,|\,f(i, j, k)\}$ of the indicator of a seismic event in the cell (i, j, k). In practice this solution may be difficult to obtain because the number of events in a catalog might be not sufficient.

Indeed, the range of value of a scalar precursor is usually divided into a number M of intervals, and only a few events are registered for any such interval. For a Q-dimensional precursor the number of Q-dimensional rectangles, covering the range, is already M^Q, and majority of them contains zero events. Only a small number of such rectangles contains one or more events, that is the precision of such an estimate of conditional probability is usually too low to have a practical value.

For this reason one constructs a new scalar precursor in the form of the scalar function of component of the vector precursor, and optimizes its predictive power. This approach leads to additional complication as the units of measurement and the physical sense of different components of a precursor are substantially different. In order to overcome this problem we use some transformation to reduce all the components of the precursor to a standard form with the same sense and range of values.

Let us transform all the precursors $f_q(i, j, k)$, $q = 1, ..., Q$ to variables with the values in $[0, 1]$ providing estimates of conditional probabilities. So, after some transformation F we obtain an estimate of $\Pr\{h \mid u(f(i, j, k) = 1\}$, where u is a characteristic function of some interval B, i.e., the probability of event $h(i, j, k) = 1$ under condition that this precursor takes the value $f(i, j, k) \in B$.

The transformation F of a scalar precursor $f(i, j, k)$ is defined as follows. Fix an arbitrary small number ε. Let L be a number of cells (i, j, k) such that $h(i, j, k) = 1$, and Z_l, $l = 1, ..., L$, be the order statistics, i.e., the values $f(i, j, k)$ in these cells listed in non-decreasing order. Define a new sequence z_m, $m = 1, ..., M$, from the ordered statistics Z_l by the following recursion: $z_0 = -\infty$, z_m is defined as the first point in the sequence Z_l, such that $z_m - z_{m-1} \geq \varepsilon$. Next, construct the sequence $z_m^* = z_m + (z_{m+1} - z_m)/2$, $m = 1, ..., M - 1$, and add the auxiliary elements $z_0 = -\infty$, $z_M = 1$. Define also a sequence n_m, $m = 1, ..., M$, where n_m equals to the number of values in the sequence Z_l, such that $z_{m-1}^* \leq Z_l < z_m^*$. Finally, define the numbers N_m, $m = 1, ..., M$ counting all cells such that $z_{m-1}^* \leq f(i, j, k) < z_m^*$, $m = 1, ..., M$. Observe that $\sum_{m-1}^M n_m = L$, $\sum_{m-1}^M N_m = N$, and use the ratios

$$\lambda = \frac{L}{N} \qquad (5.1)$$

as estimate of unconditional probability of a seismic event in a given cell

$$p_A \equiv \Pr_A\{h(i; j; k) = 1\} =$$
$$= \int_{-\infty}^\infty \Pr\{h(i; j; k) = 1 \mid x\} dP_A(x) \qquad (5.2)$$

The transformation F is defined as follows

$$g = F f(i, j, k) =$$
$$= \frac{n_m}{N_m} \text{ if } z_{m-1}^* \leq f(i; j; k) < z_m^*, m = 1, ..., M. \qquad (5.3)$$

This definition implies that transformation F replaces the value of a precursor for the frequency, i.e., the ratio of a number of cells containing a seismic event and the values of a precursor from $[z_{m-1}^*, z_m^*)$ to the number of cells with the value of precursor in this range. These frequencies are the natural estimates of conditional probabilities $\Pr\{h(i, j, k) = 1 \mid z_{m-1}^* \leq f(i, j, k) < z_m^*\}$, $m = 1, ..., M$, computed with respect to stationary distribution $P_A(x)$:

$$\Pr\{h(i; j; k) = 1 \mid z_{m-1}^* \leq f(i; j; k) < z_m^*\} =$$
$$= \frac{\int_{z_{m-1}^*}^{z_m^*} \Pr\{h(i; j; k) = 1 \mid x\} dP_A(x)}{\int_{z_{m-1}^*}^{z_m^*} dP_A(x)} \qquad (5.4)$$

(This conditional probability can be written as $\Pr\{h = 1 \mid u(f)\}$, where u is the characteristic function of interval $[z_{m-1}^*, z_m^*)$. The function g has a stepwise shape, and the length of the step is bounded from below by ε. It can be checked that there exists the limit $\tilde{g} = \lim_{\varepsilon \to \infty} \lim_{K \to \infty} g = \Pr\{h = 1 \mid u(f)\}$.

The estimates of conditional probabilities in terms of the function g are quite rough because typically the numbers n_m are of the order 1. As a final result we will present below more sharp but less detailed estimates of conditional probabilities and confidence intervals for them.

6. Combinations of precursors

There are many ways to construct a single scalar precursor based on the vector precursor $(F f_q, q = 1, ..., Q)$. Each such construction inevitably contains a number of parameters or degrees of freedom. These parameters (including the parameters used for construction of the precursors themselves) should be selected in a way to optimize the predictive power of the forecasting algorithm. The optimization procedure will be presented below, its goal is to adapt the parameters of precursors to a given catalog of earthquakes, that is to obtain the best possible retrospective forecast. However, this adaptation procedure creates a "ghost" information related with the specic features of the given catalog but not present in physical propertiess of real seismicity. This "ghost" information will not be reproduced if the algorithm is applied to another catalog of earthquakes. It is necessary to increase the volume of the catalog and to reduce the number of free parameters to get rid of this "ghost" information. Clearly, the first goal requires the considerable increase of the observation period and may be achieved in the remote future only. So, one concentrates on the reduction of numbers of degrees of freedom. The simplest variant including $Q - 1$ parameters is the linear combination

$$f^* = F f_1 + \sum_{q=2}^Q c_{q-1} F f_q. \qquad (6.1)$$

As a strictly monotonic function of a precursor is a precursor itself the log-linear combination is an equally suitable choice

$$f^* = \ln(F f_1) + \sum_{q=2}^Q c_{q-1} \ln(F f_q). \qquad (6.2)$$

Here c_q, $q = 1, ..., Q - 1$ are free parameters. The result of the procedure has the form $g = F f^*$.

7. Alarm levels, point and interval estimations

In view of Equation (5.3) the precursor g is the set of estimates for probabilities

$$\Pr\{h = 1 \big| z^*_{l-1} \le f(i;j;k) < z^*_l\}, l = 1, ..., L(f).$$

Its serious drawback is that typically $\{z^*_{l-1} \le f(i;j; k) < z^*_l\}$ correspond to single events, and therefore the precision of these estimates is very low (the confidence intervals discussed below may be taken as a convenient measure of precision). In order to increase the precision it is recommended to use the larger groups of cells containing a larger number of events, that is a more coarse covering of the space where the precursor takes its values. In a sense, the precision of the estimation and the resolving power are related by a kind of "uncertainty principle": the more precise estimate one wants to get the more coarse is the time-space range of their values and *vice versa*.

We adapt the following approach in order to achieve a reasonable compromise.

(1) For fixed thresholds a_s, $s = 1, ..., S+1$, $a_1 = 1$, $a_s < a_s+1$; $a_{S+1} = 0$, we define S possible alarm levels $a_{s+1} \le g(i;j;k) < a_s$ and subsets Ω_s, $s = 1, ..., S$, of the set Ω_K corresponding to alarm levels, i.e., Ω_s is a set of cells of Ω_K, such that $a_{s+1} \le g(i;j;k) < a_s$.

There are different ways to choose the number S of alarm levels and the thresholds a_s, $s = 2, ..., S$. Say, fix $S = 5$, and select $a_s = 10^{-\alpha(s-1)}$. This is a natural choice of the alarm level because at $\alpha = 1$ it corresponds to decimal places of the estimate of the conditional probability given by the precursor. The problem with $S = 2$, i.e. two-level alarm, may be reduced to the hypothesis testing and is discussed in more details below.

(2) Compute the point estimates of probabilities

$$\Pr\{h = 1 \big| a_{s+1} \le g(i,j,k) < a_s\}, s = 1, ..., S.$$

obtained via the distribution $P_\Omega(x)$ of a precursor g in the same way as in Equation (4.4). The property (iv) implies that for any domain Ω_s the binary random variables $h(i, j, k)$ are independent and identically distributed, i.e.

$$\Pr\{h = 1 \big| a_{s+1} \le g(i,j,k) < a_s\} \equiv p_s,$$

and the unbiased estimate of p_s takes the form

$$\theta = \frac{m_s}{n_s} \tag{7.1}$$

where n_s stands for the number of cells in domain Ω_s, and by m_s we denote the number of cells in Ω_s containing seismic events.

(3) The random variable m_s takes integer values between 0 and n_s. The probabilities of these values are computed via the well-known Bernoulli formula

$$\Pr(m_s = k) = \binom{n_s}{k} p_s^k (1 - p_s)^{n_s - k}.$$

Let us specify the confidence interval covering the unknown parameter p_s with the confidence level γ. In view of the integral Mouvre-Laplace theorem for n_s large enough the statistics

$$\frac{(\theta_s - p_s)\sqrt{n_s}}{\sqrt{p_s(1 - p_s)}}$$

is approximately Gaussian $N(0, 1)$ with zero mean and unit variance. Note that the values n_s increase with time. Omitting straightforward calculations and replacing the parameter p_s by its estimate θ_s we obtain that $\theta_s^- < \theta_s < +\theta_s^+$; where

$$\theta_s^- = \theta_s - \frac{t_\gamma \sqrt{\theta_s(1 - \theta_s)}}{\sqrt{n_s}},$$

$$\theta_s^+ = \theta_s + \frac{t_\gamma \sqrt{\theta_s(1 - \theta_s)}}{\sqrt{n_s}}$$

and t_γ is the solution of equation $\Phi(t_\gamma) = \frac{\gamma}{2}$. Here Φ stands for the standard Gaussian distribution function.

(4) As a result of these considerations we introduce 'the precursor of alarms' which indicates the alarm level: $\mathbf{R}(f(i, j, k)) = s(i, j, k)$. It will be used for calculations of point estimates and the confidence interval in the form

$$\{\theta_{s(i,j,k)}^- < \theta_{s(i,j,k)} < \theta_{s(i,j,k)}^+\}$$

This result will be used for a prospective forecasting procedure.

8. The information gain and the precursor quality

The construction of a 'combined' precursor R involves parameters from formula (6.1) or (6.2) as well as parameters which appear in the definition of each individual precursors f_q. It is natural to optimize the forecasting algorithm in such a way that the information gain related to the seismic events is maximal. In a one-dimensional case the information gain as a measure of the forecast efficiency was first intoduced by Vere-Jones [1998]. Here we exploit a modification of his ideas in the case of multi-dimensional space-time process. Remind the notions of the entropy and information. Putting aside the mathematical subtlety (see Kelbert and Suhov [2013] for details) we follow below an intuitive approach by Prohorov and Rozanov [1969]. The information containing in a given text is, basically, the length of the shortest compression of this text without the loss of its content. The smallest length S of the sequence of digits 0 and 1 (in a binary code) for counting N different objects satises the relations $0 \le S - \log_2 N \le 1$. So,

the quantity $S \sim \log_2 N$ characterizes the shortest length of coding the numbers of N objects.

Consider an experiment that can produce one of N non-intersecting events A_1, ..., A_N with probabilities q_1, ..., q_N, respectively, $q_1 + ... + q_N = 1$. A message informing about the outcomes of n such independent identical experiments may look as a sequence $(A_{i_1}, ..., A_{i_n})$, where A_{i_k} is the outcome of the experiment k. But for long enough series of observations the frequency n_i / n of event A_i is very close to its probability q_i. It means that in our message $(A_{i_1}, ..., A_{i_n})$ the event A_i appears n_i times. The number of such messages is

$$N_m = \frac{n!}{n_1!...n_N!}.$$

By the Stirling formula the length of the shortest coding of these messages

$$S_n \approx \log_2 N_n \approx -n \sum_{i=1}^{N} q_i \log_2 q_i.$$

The quantity S_n measures the uncertainty of the given experiment before its start, in our case we are looking for one of the possible outcomes of n independent trials. The specific measure of uncertainty for one trial is

$$\frac{1}{n} S_n = \frac{1}{n} S_n (q_1, ..., q_N) = -\sum_{i=1}^{N} q_i \log_2 q_i$$

known as Shannon's entropy of distribution q_1, ..., q_N (in physical literature it is also known as a measure of chaos or disorder). After one trial the uncertainty about the future outcomes decreases by the value $S = S_n - S_{n-1}$, this decrement equals to the information gain $I = S$, obtained as a result of single trial. The quantity

$$S(h) = -p_A \log_2 p_A - (1 - p_A) \log_2 (1 - p_A) \quad (8.1)$$

is the (unconditional) entropy of distribution for the indicator of the seismic event h in a space-time cell in the absence of any precursors. The conditional entropy $S(h \mid a_{s+1} \le g < a_s)$ under condition that in the cell (i, j, k) the alarm level s is set up equals

$$S\left(h \mid a_{s+1} \le g < a_s\right) = -p_s \log_2 p_s - (1 - p_s) \log_2 (1 - p_s)$$

The expected conditional entropy $S_\mathbf{R}(h)$ of indicator of seismic events where the averaging in performed by the distribution of precursors $\mathbf{R}$ takes the form

$$S_\mathbf{R}(h) = -\sum_{s=1}^{S} [p_s \log_2 p_s + \quad\quad (8.2)$$
$$+ (1 - p_s) \log_2 (1 - p_s)] P_A (a_{s+1} \le g < a_s).$$

We conclude that the knowledge of the precursor values helps to reduce the uncertainty about the future experiment by $S(h) - S_\mathbf{R}(h)$ which is precisely information $I(\mathbf{R}, h)$ obtained from the precursor. Taking into account Equations (8.1), (8.2) and the fact that

$$p_A = \sum_{s=1}^{S} p_s P_A (a_{s+1} \le g < a_s)$$

we specify the information gain as

$$I(\mathbf{R}, h) = \sum_{s=1}^{S} \left[p_s \log_2 \frac{p_s}{p_A} + \right.$$
$$\left. + (1 - p_s) \log_2 \frac{1 - p_s}{1 - p_A} \right] P_A (a_{s+1} \le g < a_s).$$

By analogy with the one-dimensional case [Kolmogorov 1965] the quantity $I(\mathbf{R}, h)$ may be called the mutual information about the random field h that may be obtained from observations of random field $\mathbf{R}$. It is known that the information $I(\mathbf{R}, h)$ is non-negative and equals to 0 if and only if the random fields h and $\mathbf{R}$ are independent. This mutual information $I(\mathbf{R}, h)$ takes its maximal value $S(h)$ in an idealized case of absolutely exact forecast. The mutual information is the information that the distributions of precursors contribute to that of the indicator of a seismic event. For this reason it may be considered as an adequate scalar estimate for the quality of the forecast.

The quantity $I(\mathbf{R}, h)$ depends on the cell size, i.e., on the space discretization length a and time interval Δt. We need a formal test to compare precursors defined for different size of the discretization cells. For this aim let us introduce the so-called 'efficiency' of precursors as the ratio of information gains

$$r(\mathbf{R}, h) = \frac{I(\mathbf{R}, h)}{S(h)}.$$

This efficiency varies between 0 and 1 and serves as a natural estimate of information quality of precursors. It allows comparing different forecasting algorithms and select the best one.

A natural estimate of $S(h)$ based on Equations (5.1) and (5.2) is defined as follows

$$\hat{S}(h) = -\lambda \log_2 \lambda - (1 - \lambda) \log_2 (1 - \lambda). \quad (8.3)$$

Taking into account Equation (8.1) and using an estimate of $P_A(a_{s+1} \le g < a_s)$ in the form of ratio $\tau_s = \frac{n_s}{N}$, we construct an estimate of $I(\mathbf{R}, h)$ as follows

$$\hat{I}(\mathbf{R}, h) = \sum_{s=1}^{S} \left[\theta_s \log_2 \frac{\theta_s}{\lambda} + (1 - \theta_s) \log_2 \frac{1 - \theta_s}{1 - \lambda} \right] \tau_s, \quad (8.4)$$

$$\hat{r}(R; h) = \frac{\hat{I}(\mathbf{R}, h)}{\hat{S}(h)}. \quad (8.5)$$

Remark. It seems reasonable to introduce a penalty

related to the number of superficious parameters in evaluating the quality of forecast pointing to the natural analogy with the Akaike test [Akaike 1974] and similar methods in information theory. In our context the main parameter of importance is $\hat{r}(R; h)$ and its limit as $t \to \infty$. This quantity does not involve the number of parameters directly. Probably, the rate of convergence depends on the number of parameters but this dependence is not studied yet.

9. The forecasting procedure

The number of time intervals, i.e., the number of observations N used in the construction of estimates increases with the growth of observation time. So, the computation procedure requires constant innovations. On the other hand some computation time is required to 'adapt' the model parameters to the updated information about seismic events via an iterative procedure. For these reasons we propose the following forecasting algorithm.

(1) Given initial parameter values at the moment $t_{K-1} = t_0 + (K-2)\Delta t$ we optimize them to obtain the maximum of efficiencys $\hat{r}(R; h)$ of a precursor in domain Ω_{K-1}. For this aim the Monte-Carlo methods is helpful: one perturbs the current values of parameters randomly and adapts the new values if the efficiency increases. The process continues before the value of efficiency stabilized, this may give a local maximum, so the precedure is repeated for a sufficient number of times. The choice of the initial value on the first step of the optimization procedure is somewhat arbitrary but a reasonable iteration procedure usually leads to consistent results. The optimization procedure takes the period of time $t_{K-1} < t \leq t_K$.

(2) Next, we construct the forecast in the following way. At the moment t_K the values of precursor $\hat{g}$ in each cell $(i, j, K + 1)$ is computed with optimized parameters. Based on these parameter values the alarm levels, the point estimates and confidence intervals are computed in each cell as well as the values of efficiency of precursors.

(3) The estimates of stationary probabilities of seismic events in the cell $\hat{\theta}(i, j)$ are defined as follows:

$$\hat{\theta}(i;j) = \frac{1}{K}\sum_{k=1}^{K}\theta_{s(i,j,k)},$$

they can be used for creation of maps of seismic danger (but not "hazard", this term is reserved) in the region.

10. Retrospective and prospective informativities

The efficiency of precursor which is achieved as a result of parameters optimization could be considered as retrospective as it is constructed by the precursors adaptation to the historical catalogs of seismic events. The prospective efficiency for the space-time domain containing the cell in the 'future' is based on the forecast. It is computed via formulas (8.3), (8.4), (8.5) with the only difference to the retrospective efficiency, however, approaches this value with time. In principle, the prospective efficiency is an ultimate criterion of the precursors quality and the retrospective efficiency could serve only for the preliminary selection of precursors and their adaptation to the past history of seismic events.

11. Testing of the forecasting algorithm

The efficiency of precursor could be computed exactly only in an idealized case of infinite observation time. However, its estimate may be obtained based on the observation over a finite time interval. So, if an estimate produces a non-zero value not necessarily the real effect is present. It may be simply a random fluctuation even if the precursor provides no information about the future earthquake. For this reason we would like to check the hypothesis H_0 about the independence of a precursor and an event indicator with a reasonable level of confidence. In case the hypothesis is rejected one has additional assurance that the forecasting is real, not just a "ghost". So, consider the distributions

$$P_A(x) = \frac{1}{I \cdot J}\sum_{(i;j)\in A}\mathrm{Pr}_{ij}\left\{g(i,j,k)\leq x\right\},$$

and

$$P'_A(x) = \frac{1}{I \cdot J}\sum_{(i;j)\in A}\mathrm{Pr}_{ij}\left\{g(i,j,k)\leq x \,\middle|\, h(i,j,k)=1\right\}.$$

The function $P_A(P_A^{-1}(y)) = y$ of variable $y = P_A(x)$ provides an uniform distribution $F^*(y) = \mathrm{Pr}\{\xi \leq y\}$ of some random variable ξ on $[0,1]$. Next, consider a distribution function $G(y) = P'_A(P_A^{-1}(y))$ on $[0,1]$, and use a parametric representation for abscissa $P_A(x)$ and ordinate $P'_A(x)$. If random fields g and h are independent the distribution functions $P_A(x) = P'_A(x)$ and $G(y)$ are uniform. So, the hypothesis about the absence of forecasting, i.e., about the independence of g and h is equivalent to the hypothesis H_0 that the distribution $G(y)$ is uniform.

The empirical distribution $G_L(y)$ related to $G(y)$ is defined as follows. Denote by u_l, $l = 1, ..., L$ the values of the function $g(i, j, k)$ sorted in the nondecreasing order and belonging to the cells where $h(i, j, k) = 1$. Let n_l be the numbers of cells such that $h(i, j, k) = 1$, $g(i, j, k) = u_l$. Denote by $m(u_l)$ the numbers of cells such that $g(i, j, k) < u_l$, and define the empirical distribution $G_L(y)$ as a stepwise function with $G_L(0) = 0$ and positive jumps of the size $\frac{n_l}{L}$ at points $y_l = \frac{m(u_l)}{N}$.

The well known methods of hypothesis testing requires that the function $G_L(y)$ has the same shape as for

independent trials, i.e., random variables u_l, $l = 1, ..., L$ are independent in view of axiom (iv). Naturally, we accept the precursors such that the hypothesis H_0 is rejected with the reasonable level of confidence. (Remind, that the hypothesis is accepted if and only if its logical negation could be rejected based on the available observations. The fact that the hypothesis cannot be rejected does not mean at all that it should be accepted, it only means that the available observations do not contradict this hypothesis. Say, the well-known fact that "The Sun rises in the East" does not contradict to our hypothesis, however it may not be considered as a ground for its acceptance.) For large values of L the Kolmogorov statistics [Kolmogorov 1933a] is helpful for this aim

$$D_L = \sup \left| G_L(y) - y \right|$$

with an asymptotic distribution

$$\lim_{L \to \infty} \Pr \left\{ \sqrt{L} D_L \le z \right\} = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}, z > 0,$$

or Smirnov's statistics [Smirnov 1939]

$$D_L^+ = \sup \left[G_L(y) - y \right]$$
$$D_L^- = -\inf \left[G_L(y) - y \right]$$

with asymptotic distribution

$$\lim_{L \to \infty} \Pr \left\{ \sqrt{L} D_L^+ \le z \right\} = \lim_{L \to \infty} \Pr \left\{ \sqrt{L} D_L^- \le z \right\} =$$
$$= 1 - e^{-2z^2}, z > 0.$$

The asymptotic expressions for these statistics can be used for $L > 20$ [Bolshev and Smirnov 1965].

12. The binary alarm and the hypothesis testing

The prediction is the form of forecast when an alarm is announced in a given cell without a preliminary evaluation of the probability of a seismic event. In this case we can estimate the probabilities of events, too. (If the alarm is announced in an arbitrary domain Ω we set up an alarm if at least half of the cell of our model is occupied by alarm.)

Let M be the number of cells in Ω which are in the state of alarm, M_0 be the number of cells where the seismic event is present but no alarm was announced (the number of 'missed targets'). Denote by $\tau = \frac{M}{N}$ the share of the cells with alarm announced, $\lambda = \frac{L}{N}$ the share of the cells with seismic events, and $\nu = \frac{M_0}{M}$ the share of missed targets. Let a random variable $\eta(i, j, k)$ equal 1 if an alarm is announced in the cell (i, j, k) and 0 otherwise. Obviously, the estimate of conditional probability $\Pr \{h(i, j, k) = 1 \,|\, \eta(i; j; k) = 1\}$ of the seismic event under the condition of alarm is $\frac{\lambda(1-\nu)}{\tau}$; and the esti-

mate of conditional probability $\Pr \{h(i, j, k) = 1 \,|\, \eta(i; j; k) = 0\}$ of the seismic event under the condition of no alarm is $\frac{\lambda\nu}{1-\tau}$.

If the alarm is announced according to the procedure described in Section 5 the threshold a_1 specifying the acceptable domain of values for $g(i, j, k)$ should be treated as a free parameter and selected by maximizing the information efficiency $\hat{r}(\eta; h)$. The estimate of information increase for given values of τ and ν equals

$$\hat{I}(\eta, h) = \lambda(1 - \nu) \log_2 \frac{1 - \nu}{\tau} + \lambda\nu \log_2 \frac{\nu}{1 - \tau} +$$
$$+ [\tau - \lambda(1 - \nu)] \log_2 \frac{\tau - \lambda(1 - \nu)}{(1 - \lambda)\tau} +$$
$$+ (1 - \tau - \lambda\nu) \log_2 \frac{1 - \tau - \lambda\nu}{(1 - \tau)(1 - \lambda)}.$$

The value of $\eta(i; j; k)$ characterizes the results of checking two mutually exclusive simple hypothesis:

H_0: the distribution of $f^*(i, j, k)$ has the form

$$P_A^0(x) = \Pr_A \left\{ f^*(i, j, k) \le x \,|\, h(i, j, k) = 0 \right\},$$

implying 'no seismic events', or

H_1: the distribution of $f^*(i, j, k)$) has the form

$$P_A^1(x) = \Pr_A \left\{ f^*(i, j, k) \le x \,|\, h(i, j, k) = 1 \right\},$$

implying the presence of seismic event.

Statistics for checking of these hypothesis is the precursor $g(i, j, k)$, and the critical domain for H_0 has the form $\{g(i, j, k) \ge a_1\}$. (If usual method of alarm announcement is used the relevant precursor plays the role of statistics and the critical domain is defined by the rule of the alarm announcement). The probability of a first type error

$$\alpha = \Pr \left\{ g(i, j, k) = 1 \,|\, h(i, j, k) = 0 \right\},$$

it is estimated as $\frac{\tau - \lambda(1-\nu)}{1 - \lambda}$. The probability of second type error

$$\beta = \Pr \left\{ g(i, j, k) = 0 \,|\, h(i, j, k) = 1 \right\},$$

is estimated as ν. (Note that due to condition (iii) any test used for checking these hypothesis should not depend on the coordinates of the cell).

The Neyman-Pearson theory allows to define the domain of images of all possible criteria: in coordinates (α, β) it is a convex domain with a boundary Γ which corresponds to the set of uniformly most powerful tests. This family may be defined in terms of the likelihood ratio $\Lambda(x) = \frac{p_A^1(x)}{p_A^0(x)}$ under condition that the distribu-

tions $P_A^1(x)$ and $P_A^0(x)$ have the densities $p_A^1(x)$ and $p_A^0(x)$.

$$\begin{cases} \eta(i,j,k) = 1 \ if \ \Lambda(x) > \omega \\ \eta(i,j,k) = 0 \ if \ \Lambda(x) < \omega \end{cases}$$

where ω denotes the threshold. In Gercsik and Kelbert [2004] we demonstrated that among all the tests with the images on the boundary Γ there exists three different best tests. Here the term "best" may be understood in three different senses, i.e., maximizing the variational, correlational and informational efficiency. The most relevant criteria is the informational efficiency $\hat{r}(\eta; h)$.

13. The choice of precursors

We use the term 'empiric precursor of an earthquake' for any observable characterisric derived from the catalog only, which provides for this catalog a reasonable retrospective forecast of seismic events, and which is not derived from basic physical concepts of seismicity (say, the periods of relatice calm, deviation of some basic characteristic from a long-time average , etc). In contrast, the physical precursors are derived from some physical processes and characterize physical quantities (stress fields, strength, concentration of cracks, etc.) or well-defined physical processes (i.e., phase transitions, cracks propagations, etc.) In the meteorological forecast the danger of using empirical precursors was highlighted by A. Kolmogorov in 1933 [Kolmogorov 1933b]. From that time the meteorological forecast relies on the physical precursors which are theoretically justified by the models of atmospheric dynamics. Below we will present Kolmogorov's argument adapted to the case of seismic forecast. This demonstrates that the purely empirical precursors work well only for the given catalog from which they are derived. However, their efficiency deteriorates drastically when they are applied to any other independent catalog.

Consider a group of k empirical precursors used for a forecast and selected from a set of n such groups. According to Kolmogorov's remark the number k is typically rather small. This is related to the fact that a number of strong earthquakes in catalog is unlikely to exceed a few dozen. As the values of precursors are random there exists a small probability p that the efficiency of the forecast exceeds the given threshold C. Then the probability of event $\hat{r}(R, h) \leq C$ equals $1 - p$, and the probability of event $\hat{r}(R, h) > C$ for at least one collection of precursors equals $P = 1 - (1 - p)^n$ and tends to 1 as $n \to \infty$. (According to Kolmogorov some arbitrariness of the assumption of independence is compensated by the large number of collections.)

Summing up, if the number of groups is large enough with probability close to 1 it is possible to find a group giving an effective retrospective forecast for a given catalog. In practice this is always the case as the number of empirical precursors could be increased indefinitely by variation of real parameters used in their construction. It is important to note that for such a group, which is highly efficient for the initial catalog, the probability that the efficiency is greater than C is still equal to p for any other catalog. In other words: the larger the number of the groups of empirical precursors the less reliable the forecast is. So, the collection of a large list of the empirical precursors is counter-productive. Much more reliable are the physical precursors intrinsicly connected with the physical processes which preserve their values with the change of sample. The probability to find such a set of precursors by pure empirical choice is negligible because they are very rare in the immense collection of all possible precursors.

14. Demonstration of algorithm

A preliminary version of the forecast algorithm described above was used in the paper [Ghertzik 2008] for Califorrernia and the Sumatra-Andaman earthquake region. These computations serve as a demonstration of the efficiency of the method but their actual results should be taken with a pinch of salt because the selection of precursors does not appear well-justied from the modern point of view: the number of free parameters (34) to be adapted in the precursor "stress indicator" is too large. For this reason the method was not fully exploited. Nowadays, the progress is resumed with a particular emphasis on the selection of a new set of precursors strongly related to the critical phenomena in geophysics.

The following selection of precursors was used:

(1) The stress indicator is a scalar predictor characterizing the local level of the stress field. It is conceivable as a stress tensor component. The predictor increases due to tectonic movements, decreases with viscosity and with increasing fracturing, and is affected by earthquake sources considered as cracks.

(2) The damage indicator is a predictor characterizing the stress-induced microfracturing level. The predictor increases with the value of applied stresses and decreases as cracks are healed.

(3) The third predictor is the concentration Zhurkov-Sobolev criterion of seismically active faults based on the calculation of the parameter concentration (density) seismogenic ruptures and modified in such a way that it can be calculated for a local cell.

(4) The far-order predictor is the integral of the radial correlation function for seismic energy released in a spatiotemporal cell. Its construction is based on the idea of the strong earthquake as a phase transition be-

tween consolidation and fracture states of a medium. The phase transition is preceded by the appearance of a far order, an abrupt increase in the correlation radius for the energy of shocks released in a cell.

For detailed description of precursors we refer to original, for it is only an illustration but not the topic of this work. It is useful, however, to summarize the main results of this paper.

California region. The catalog Global Hypocenter Data Base CD-ROM NEIC/USGS, Denver, CO, 1989, together with data from the site NEIC/USGS PDE (ftp://hazard.cr.usgs.gov) for earthquakes with magnitudes $M \geq 4.0$ with epicenters between $113°$-$129°$ of western longitude and $31°$-$43°$ of northern latitude was used for parameter adaptation. The initial time t_0 was selected by subtracting from the time of actual computation, 08.03.2006, an integer number of half-year intervals such that t_0 fits the first half of the year 1936. (The final time 08.03.2006 could be considered as an initial moment for constructing half year forecast forward up to the date of the latest earthquake available in the catalog). During the computation the time interval from the first half of 1936 to the first half of 1976 was used for relaxation of the zero initial data used for precursors. After this date the catalog for the earthquakes with magnitude $M \geq 4.0$ was used to estimate the probabilities of strong earthquakes with magnitude $M \geq 6.0$ up to the moment of actual forecast. Note that the adapted restriction to include into consideration only earthquakes with magnitudes $M \geq 4.0$ is a severe restriction. It decreases the precision of precursor computation and therefore, if a prediction is successful, increases the degree of condence in the predictor choice. We choose $a = 150$ km as a size of the spacial lattice, and $\Delta t = 6$ months as a time-step. Retrospective forecast was performed with 5 alarm levels defined by the thresholds $a_s = 10^{-\alpha(5-s)}$, $s = 1, ..., 4$ and $\alpha = 0.75$. (Due to too short time step no alarms was registered on the lowest level when parameter $\alpha = 1$ was selected). In order to reduce the influence of the boundary conditions the large square covering all the
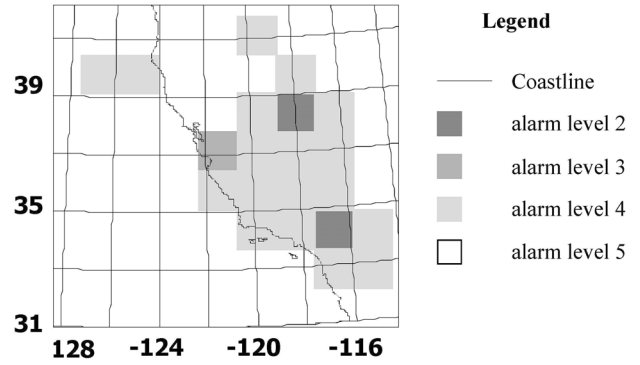

Figure 1. Mapped areas of five levels of half-year forward forecasting for California on March 8, 2006.

seismic events in the catalog used in the computations was reduced by two layers of elementary cells from each boundary. As a result of optimization the forecast information efficiency of 0.526 was achieved, i.e., the forecasting algorithm applied to the given catalog extracted from it about 53% of all available information about seismic events. It seems that this result could be only partially explained by a lucky selection of precursors: another contributor to the high efficiency of the algorithm is the adaptation of the parameters to the features of the specific catalog. The influence of this artificial information may be reduced only with the increase of the observation interval. Accepting the rule of binary alarm announcement in the cells from group 1 and 2 from 5 levels possible one obtains that the space-time share of alarmed cells is 3.4% and the share of missed targets is 18.2%. This result is comparable with the best forecasts available in the literature and obtained by other methods (in the cases when the quantitative parameters of algorithms are presented in the publications). When the forecast was constructed in the future we obtained that the estimate of probability of a strong earthquake anywhere in the area under study is 0.174, and the maximal point estimate of an event in any individual cell is 0.071. As a whole the seismic situation in California did not look too alarming. Indeed, there were no strong earthquakes in California in the next half-year.

Alarm level	Number of cells	Number of events	Left boundary of interval	Point estimate	Right boundary of interval
1 ■	43	25	0.4186	0.5814	0.6998
2 ■	155	11	0.0376	0.0710	0.1172
3 ■	457	6	0.0063	0.0131	0.0291
4 ■	1152	2	0.0005	0.001	0.0071
5 □	4049	0	0	0	0.001

Table 1. Number of events and alarm cells for each level of the retrospective analysis for California.

Table 1 presents the number of events and the total number of alarm cells for each level of the retrospective analysis, as well as their ratios, which are well-grounded point estimates of the probability that at least one earthquake with a magnitude $M \geq 6.0$ will occur in a cell at a given level of alarm. The 5% condence intervals are also given in the table for each of these probabilities. The true values of event probabilities lie within these intervals with a 95% probability (condence coefficient). In all, 5856 cells participated in the prediction, and 44 of these cells (0.75% of all cells) contained events.

SAE region. We have conducted a retrospective forecast of strong earthquakes with magnitudes $M \geq 7.0$ for the whole region where the Sumatra-Andaman earthquake (SAE) happened on 26.12.2004 with magnitude $M = 9.3$. The catalog Global Hypocenter Data Base CD-ROM NEIC/USGS, Denver, CO, 1989, together with the data from the site NEIC/USGS PDE (ftp://hazard. cr. usgs.gov) for earthquakes with magnitudes $M \geq 5.5$ with epicenters between $84.3°$-$128°$ of eastern longitude and $20°$-$26°$ of northern latitude was used for the parameters adaptation. The initial moment of time t_0 was selected by subtracting from the time of actual computation, 10.11.2004, an integer number of half-year intervals such that t_0 fits the first half of the year 1936. (The final time was selected in such a way that the next half-year period covers SAE and its powerful aftershock). During the computation the time interval from the first half of 1936 to the first half of 1976 was used for relaxation of the zero initial data used for precursors. After this date the catalog was used to estimate the probabilities of strong earthquakes with magnitude $M \geq 7.0$ up to the moment of actual forecast. (For magnitude $M \geq 7.5$ the number of seismic events was not sufficient for reliable forecast because the 5%-confi-

dence intervals strongly overlapped). In this case the restriction to include into consideration only earthquakes with magnitudes $M \geq 5.5$ was adapted. As before, it decreases the precision of precursor computation and therefore, if the prediction is successful, increases the degree of confidence to the predictor choice. We selected the size of the spacial grid as $a = 400$ km and the size of time-step $\Delta t = 6$ months. Retrospective analysis was conducted following the same scheme as in the previous case. In order to reduce the influence of the boundary conditions the large square covering all the seismic events in the catalog used in the computations was reduced by two layers of elementary cells from each boundary. As a result of optimization the forecast information efficiency was 0.549, i.e., the forecasting algorithm extracted around 55% of all available information about seismic events when applied to the given catalog. In the case of binary alarm announcement the space-time share of alarmed cells was 3.1% and the share of missed targets was 8.3%. This result is comparable with the best forecasts available in the literature and obtained by the other methods (in the cases when the quantitative parameters of algorithms are presented in the publications). In the case of forward forecast the two most powerful earthquakes, i.e., SAE and its major aftershock, happened in the alarm zone of the second level, and two other events with smaller magnitudes in the zone of alarm level 4.

The results of the half-year forward prediction beginning from November 10, 2004, and $M \geq 7.0$ earthquakes that occurred in this period are plotted in Figure 2. As seen from the figure, the two strongest events (the SAE and its strongest aftershock) occurred in the level-2 alarm zone, while two weaker events occurred in the level-4 alarm zone. It is noteworthy that, if monitoring
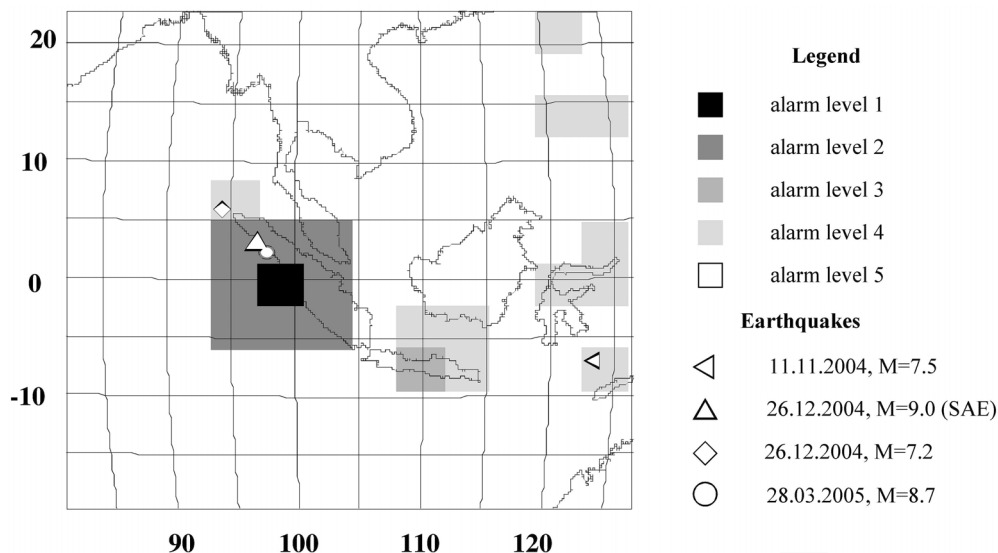

Figure 2. Mapped areas of five levels of half-year forward forecasting for the SAE region on November 10, 2004.

Alarm level	Number of cells	Number of events	Left boundary of interval	Point estimate	Right boundary of interval
1 ■	36	18	0.3300	0.500	0.6700
2 ■	222	15	0.0400	0.0676	0.1052
3 ■	220	2	0.0015	0.0091	0.0281
4 □	602	1	0	0.0017	0.0070
5 □	7272	0	0	0	0.0006

Table 2. Number of events and alarm cells for each level of the retrospective analysis for SAE region.

in this region were conducted according to the proposed scheme, a nine-cell square cluster containing one cell of the level-1 alarm would be regarded as hazardous, giving every reason for anxiety in relation to the impending SAE.

Table 2 presents the number of events and the total number of alarm cells for each level of the retrospective analysis, as well as their ratios, which are well-grounded point estimates of the probability that at least one earthquake with a magnitude M 7.0 will occur in a cell at a given level of alarm. The 5% condence intervals are also given in the table for each of these probabilities. In all, 8352 cells participated in the forecasting and 36 of these cells (0.43% of all cells) contained events.

Figure 3 plots the dependence of the fraction of target misses on the spatiotemporal fraction of alarms for points at which the boundaries between alarm levels divide the set of all cells into two classes: union of more anxiety cells and union of less anxiety cells. It is an estimate for Molchan's error diagram.
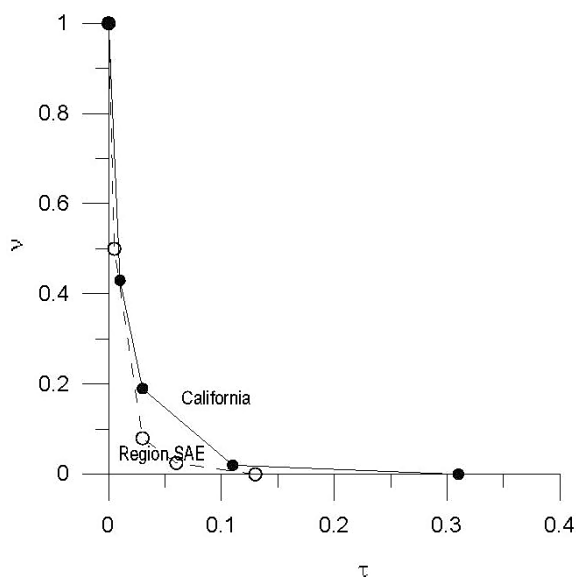

Figure 3. Empirical Molchan's diagram: the share of missed targets vs the share of alarms.

15. Conclusions

This article demonstrates that the prediction method based on the pattern recognition does not have sufficient foundation and contains a number of errors. We proposed here the alternative approach that is free of this errors and enhances the computing capability of the forecasting. Discretization of spacetime, using elements of the theory of multicomponent random processes and axiomatization of the problem allows to drastically simplify the procedure of earthquake forecasting and to obtain the following results:

(1) The method of announcement to declare a multi-level alarm was found.

(2) The method of calculating the point estimate of probability of an event for each alarm level was found.

(3) The method of calculating interval estimates of event probability for each level of anxiety was found.

(4) The equivalence of the announcement of a two-phase alarm and solving the problem of testing hypotheses was shown.

(5) The method of testing an algorthim of forecast was found.

(6) The scalar criterion for comparison of different algorithms of forecast was found.

(7) The equivalence of the method of the announcement of alarms and method assessing the probability of event was shown.

It appears to the authors, that a satisfactory mathematical formalism of earthquake forecasting based on arbitrary set of precursors is fully fledged. The most important result we believe is the discovery of the scalar quality criterion of forecast algorithms. The criterion allows to find out which one is the best from existing algorithms and to focus on finding the most effective precursors.

References

Akaike, H. (1974). A new look at the statistical model identification, IEEE T. Automat. Contr., 19 (6), 716-723.

Bolshev, L.N., and V.N. Smirnov (1965). Mathematical Statistical Tables, V.A. Steklov Matematical Institute, Academy of Sciences, Moscow, USSR, 464 p. (in Russian).

Gercsik, V., and M. Kelbert (2004). On comparison of hypothesis tests in Bayesian framework without a loss function, Journal of Modern Applied Statistical Methods, 3 (2), 399-405.

Ghertzik, V. (2008). Physical concepts of fracture and prediction of probabilities of strong earthquakes, Phys. Solid Earth, 44 (3), 22-39.

Harte, D., and D. Vere-Jones (2005). The entropy score and its uses in earthquake forecasting, Pure Appl. Geophys., 162 (6-7), 1229-1253.

Holliday, J.R., C. Chen, K.F. Tiampo, J.B. Rundle, D.L. Turcott and A. Donnellan (2007). A RELM Earthquake forecast based on pattern informatics, Seismol. Res. Lett., 78 (1), 87-93.

Jordan, T.H., Y.-T. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, G. Papadopoulos, G. Sobolev, K. Yamaoka and J. Zschau (2011). Operational earthquake forecasting: state of knowledge and guidelines for utilization, Annals of Geophysics, 54 (4); doi:10.4401/ag-5350.

Kagan, Y.Y., and D.D. Jackson (2000). Probabilistic forecasting of earthquakes, Geophys. J. Int., 143, 438-453.

Keilis-Borok, V.I., and V.G. Kossobokov (1990). Premonitory activation of earthquake flow: algorithm M8, Phys. Earth Planet. In., 61. 73-83.

Keilis-Borok, V.I. (1996). Intermediate-term earthquake prediction, P. Natl. Acad. Sci. USA, 93 (9), 3748-3755.

Keilis-Borok, V.I., and A.A. Soloviev (2003). Nonlinear Dynamics of the Lithosphere and Earthquake Prediction, Springer-Verlag, Berlin-Heidelberg, 338 p.

Kelbert, M., and Y. Suhov (2013). Information Theory and Coding by Example, Cambridge Univ. Press, Cambridge, 530 p.

Kolmogorov, A.N. (1933a). Sulla determinazione empirica di una legge distribuzione, Giornale dell'Istituto Italiano degli Attuari, 4, 83-91 (in Italian).

Kolmogorov, A.N. (1933b). On the suitability of statistically obtained prediction formulas, Zh. Geoz., 3, 7882.

Kolmogorov, A.N. (1965). Three approaches to the definition of the notion of information amount, Probl. Inf. Transm., 1 (1), 3-11.

Molchan, G.M. (1990). Strategies in strong earthquake prediction, Phys. Earth Planet. In., 61, 84-98.

Molchan, G.M., and V.I. Keilis-Borok (2008). Earthquake prediction: probabilistic aspect, Geophys. J. Int., 173, 1012-1017.

Pisarenko, V.F., A. Sornette, D. Sornette and M.V. Rodkin (2008). New approach to the characterization of M_{max} and of the tail of the distribution of earthquake magnitudes, Pure Appl. Geophys., 165 (5), 847-888.

Prohorov, Yu.V., and Yu.A. Rozanov (1969). Probability theory, basic concepts. Limit theorems, random processes, Springer-Verlag, 401 p.

Smirnov, N.V. (1939). On deviations from the empirical distribution curve, Mat. Sb., 6 (48), 1, 3-24.

Vere-Jones, D. (1998). Probability and information gain for earthquake forecasting, In: Problems of Geodynamics and Seismology, Computational Seismology, Moscow, 30, 248-263.

WGNCEP, Working Group on Northern California Earthquake Potential (2007). The Uniform California Earthquake Rupture Forecast, Version 3 (UCERF3) Project Plan: U.S. Geological Survey Open-File Report 1-176, 2011.

*Corresponding author: Vladimir Gertsik,
Institute of Earthquake Prediction Theory and Mathematical Geophysics RAS, Moscow, RF; email: gercsik@mitp.ru, gertzik@yandex.ru.