A Method of Identification of Potential Earthquake Source Zones

K. N. Petrov, V. G. Gitis, and A. B. Derendyaev(✉)

The Institute for Information Transmission Problems, Moscow, Russia
stranger12@list.ru, gitis@iitp.ru, wintsa@gmail.com

Abstract. We propose a machine learning method for mapping potential earthquake source zones (ESZ). We use two hypotheses: (1) the recurrence of strong earthquakes and (2) the dependence of sources of strong earthquakes on the properties of the geological environment. To solve this problem, we know the catalog of earthquakes and a set of spatial fields of geological and geophysical features. We tested the method of identification of the potential ESZ with $m \geq 6.0$ for the Caucasus region. The map of the potential earthquake source zones and a geological interpretation of the decision rule are presented.

Keywords: Earthquake prediction · Spatial grid fields of features · Machine learning · One-class classification · Potential Earthquake Source Zones

1 Introduction

In areas of natural risk, a mapping of potential hazards of geological processes and phenomena is carried out. The level of danger depends on the energy characteristics of the source of exposure, the distance to it, and the attenuation parameters. This article discusses the problem of zone identification of potential sources of earthquakes. The complexity of the task lies in need to assess the places of the possible occurrence of rare events, and the importance is associated with social and economic consequences due to errors in missing the target and adding false zones to the analyzed territory on the map.

To construct a map of the earthquake source zones (ESZ), we use two hypotheses: (1) the recurrence of strong earthquakes and (2) the dependence of sources of strong earthquakes on the properties of geological environment. The first hypothesis is based on the known facts of an earthquake repetition. But the history of seismic observations is very short with respect to the speed of tectonic processes, and strong earthquakes occur relatively rarely. Therefore, the use of only the first hypothesis can lead to omissions of the zones due to the relatively short observation period [4,13,17]. The second hypothesis compensates for the effect of insufficient representativeness of existing seismic observations

The paper is supported by the Russian Science Foundation, project No20-07-00445.

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12252, pp. 405–414, 2020.
https://doi.org/10.1007/978-3-030-58811-3_29

[3,7,21]. However, usually, the data of the geological environment are incomplete and indirectly associated with the magnitude of the strong earthquakes. Incomplete information dramatically complicates the search for patterns based on empirical data. Besides, this makes it difficult to choose a reasonable model for the statistical estimation of the accuracy of a solution. Therefore, the result of solving the problem should allow specialists in the field to interpret the obtained forecasting rule.

In solving the problem, we are aware of the epicenters of the occurrence of targeted earthquakes with magnitudes above a certain threshold and a set of spatial grid based fields of geological and geophysical features. We suggest that the target earthquakes may recur in the vicinity of the earlier earthquake epicenters. The rest of the region is a mixture of points at which the possibility of the appearance epicenters of target earthquakes has not been established. The challenge is to map the zones in which epicenters of the targeted earthquakes are possible and provide a geological interpretation of the solution. In machine learning, this task is called a one-class classification task [2,12,15].

A method for identifying potential ESZ is discussed in Sect. 2. It generalizes two one-class classification algorithms: the algorithm of the method of the minimum area of alarm [9] and the algorithm of the preference method [10]. Section 3 discusses the verification of this method for determining ESZ with $m \geq 6.0$ for the Caucasus region.

2 Method

Let the seismotectonic properties of the analyzed zone be represented by a catalog of earthquakes and a set of geological and geophysical spatial fields. The challenge is to use this data for identifying potential earthquake source zones (ESZ) with target magnitudes that exceed a certain threshold. We suggest a recurrence of strong earthquakes, as well as the possibility of the emergence of new sources of earthquakes in areas with similar geological properties. Following the hypothesis of an earthquake recurrence, the potential ESZ should cover all the epicenters of the target earthquakes in the catalog. In order to extrapolate the area of ESZ to the territory where target earthquakes were not observed, it is necessary to determine the decision rule based on the hypothesis that zones with strong earthquakes depend on the properties of the geological environment. For this, machine learning methods are used.

The difficulty in finding a decision rule lies in incomplete data. Our set of examples of targeted earthquakes is not full due to the relatively short observation period. In addition, among the target earthquakes presented in the catalog, there are examples in which available geological and geophysical fields do not have geological and geophysical properties that explain their high magnitudes. In the first case, examples of environmental properties inherent in the areas of target earthquakes are not included in the training set. In the second case, examples of target earthquakes do not have properties that would distinguish places of their source zones from areas where target earthquakes are impossible. This

explains the errors in decision-making: errors in missing places where targeted earthquakes are possible, and errors in the reassessment of danger due to an incomplete description of the properties of the geological environment.

The method of identifying the potential ESZ consists of 3 stages.

1. Calculate a decision rule and select a target earthquakes for which the values of the feature fields adequately represent the seismotectonic features of the focal zones.
2. Calculate potential ESZ and generate a text explanation of the decision rule.
3. Supplement potential ESZ with remaining epicenters of target earthquakes.

The method uses decision rules based on monotonic functions. These rules can be applied in problems in which there is reason to assume that for a certain range of values the dependence of the predicted property on the features or the degree of confidence in the belonging of the objects in question to a certain class changes monotonously when the value of the function changes. Without loss of generality, we assume that the predicted property $S(\mathbf{x})$ increases or does not change (but does not decrease) with an increase in the value of any of the characteristics of the object, i.e., $\frac{\partial S(\mathbf{x})}{\partial x_i} \geq 0, i = 1 \ldots I$. For example, earthquake-risk zones are often characterized by heterogeneity of the geological environment, closeness to zones of active geological faults, abnormal velocities of modern vertical movements, high gradients of gravitational anomalies, etc. It can be assumed that the reliability of the fact that a map point belongs to a seismic zone is greater, all other things being equal, the greater the value of at least one of the listed feature fields.

Let the seismotectonic properties of the analyzed area are represented by spatial grid-based fields of features $\mathbf{F}_i, i = 1, \ldots, I$, in a single coordinate grid with a step $\Delta x \times \Delta y \times \Delta t$ and a sample of earthquakes $q = 1, \ldots, Q$ with the target magnitudes $m \geq M$. The values of these fields at the nodes of the grid $n = 1, \ldots, N$ correspond to the vectors of the I-dimensional feature space $\mathbf{f}^{(n)} = \{f_i^{(n)}\}$. The points of the feature space $\mathbf{f}^{(n)} = \{f_i^{(n)}\}$, which correspond to the nodes of the grid of feature fields closest to the epicenters of the target earthquakes, will be called precedents. The task is to find the field $\mathbf{\Phi}(\mathbf{F}_i)$, which determines the potential ESZ.

The data model contains two assumptions.

1. *Anomalilty condition*: epicenters of earthquakes with target magnitudes relate to zones in which the values of the feature fields are improbable and close to maximum (or minimum). To simplify the explanation, we assume that the anomalies refer only to the *largest* values of the features fields.
2. *Monotonicity condition*: points of the feature space that are componentwise greater than or equal to the precedent $\mathbf{f}^{(q)}$ of the earthquake q can also be the precedents for similar target events, that is, if $\mathbf{f}^{(n)} = \{f_i^{(n)}\} \geq \{f_i^{(q)}\}$, then $\mathbf{f}^{(n)}$ is also a precedent for a similar event.

Points of the feature space $\mathbf{f}^{(n)}$ which are componentwise greater than or equal to the precedent $\mathbf{f}^{(q)}$, will be called the *base* points of the event q. This

points belong to the half-interval $O^{(q)}$ with the vertex at the point $\mathbf{f}^{(q)}$. Each point $\mathbf{f}^{(n)}$ corresponds to one of the grid nodes of the features fields. Denote by $W^{(q)}$ a set of all grid nodes corresponding to the base points of the event q. The number of such nodes is denoted by $L^{(q)} = |W^{(q)}|$. We will call the ratio $v^{(q)} = L^{(q)}/L$ *the alarm volume of the event* q, where $L^{(q)}$ is the number of grid nodes of the potential ESZ field corresponded to the base points of the q event, L is the number of all grid nodes of the area of analysis.

The quality of the solution is determined by two indicators: (1) *probability of detection* U which is equal to the share of correctly detected epicenters of target earthquakes Q^* from all Q target events, $U = Q^*/Q$ and (2) *alarm volume* which is equal to the share of the number of grid nodes, falling in the alarm area L^*, from the number of all grid nodes L of the analyzed area $V = L^*/L$. Often the quality of the forecast is determined by the dependence $U(V)$, which practically coincides with the error curve represented by the ROC curve [5].

Algorithm [9] allows for any given alarm volume V to find optimal potential ESZ zones in which the maximum number of epicenters of target earthquake fall. We are considering a version of the algorithm that provides a solution close to optimal.

To begin to solve the problem, one should determine the magnitude of the target earthquakes and the alarm volume V_0 equal to the fraction of the analysis zone that the potential ESZ may occupy. We are not aware of the mathematically reasoned answers to both of these questions. Our model assumes that potential ESZ refer only to areas with anomalous geological and geophysical properties. As for the alarm volume V_0, here we can proceed from the fact that it's equal to the probability of detecting the epicenters of target earthquakes in random zones consisting of $L^* = VL$ grid nodes. Given this, the alarm volume V_0 can be determined empirically from the ratio $U(V_0) \geq (1.5 - 2.0)V_0$, which shows that the source zones of the target earthquakes were not randomly selected and are well represented by fields of features.

The algorithm.

1. Generate a training sample $\{\mathbf{f}^{(q)}, v^{(q)}\}$, in which the precedents $\mathbf{f}^{(q)}$ represent all the epicenters of the target eartquakes $q = 1, \ldots, Q$.
2. Arrange the precedents $\mathbf{f}^{(q)}$ in accordance with the increase in alarm volumes: $v^{(1)} \leq v^{(2)} \leq, \ldots, \leq v^{(q)} \leq, \ldots, \leq v^{(Q)}$.
3. Set all the grid nodes of the potential ESZ field the value 1. Change the value 1 of potential ESZ field to $V(1) = v^{(1)}$ at the grid nodes corresponding to the base points of the precedent $\mathbf{f}^{(1)}$; replace the value 1 with $V(2) = |W^{(1)} \cup W^{(2)}|/L$ in the grid nodes belonging to the base points of the precedent $\mathbf{f}^{(2)}$ and then, successively to replace the values of 1 with the corresponding sizes of the union of alarm volumes. We obtain the sequence $V(1), \ldots, V(q), \ldots, V(Q)$. This sequence uniquely determines the relationship $U(V)$, and V_0 determines a set of the target earthquakes $q = 1, 2, \ldots, Q_0 \leq Q$, for which the values of the feature fields distinguish the places of their source zones from the regions where target earthquakes are impossible. The precedents $\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \ldots, \mathbf{f}^{(Q_0)}$ define a domain A in the feature space consisting of the union of Q_0 half-intervals with vertices at the points $\mathbf{f}^{(q)}$.

4. Calculate for the feature fields $\mathbf{F}_i$ the standard deviations in the analysis area: $\sigma_1, \sigma_2, \ldots, \sigma_i, \ldots, \sigma_I$.
5. Calculate the domain B that covers the domain A. The domain B consists of the union Q_0 of the half-intervals with vertices at the points $\mathbf{g}^{(q)} = \{g_i^{(q)}\} = \{f_i^{(q)} - \mu\sigma_i\}$, where μ is a parameter that determines the accuracy of the approximation of the area A.
6. Build the membership matrix $\left(P_{qr}\right)$. The row of the matrix denotes the vertex of the half-interval $\mathbf{g}^{(q)}$, and the column denotes the target event r. An element of the matrix $P_{qr} = 1$ if the point r belongs to the half-interval with the vertex $\mathbf{g}^{(q)}$, that is, if $g_i^{(r)} \leq f_i^{(q)}$ for all $i = 1, 2, \ldots, I$.
7. Find the subset of the minimum number of half-intervals that contain all the use case points. For this, it is necessary to find in the membership matrix $\left(P_{qr}\right)$ a subset of the minimum number of rows for which at least one unit remains in each column of the matrix. For this, standard methods of minimizing the disjunctive normal forms of Boolean functions are used [19].

It's seen that the conditions $P_{qr} = 1$ if $g_i^{(r)} \leq f_i^{(q)}$ for all $i = 1, 2, \ldots, I$ are conjunctions. All rows of the matrix $\left(P_{qr}\right)$ correspond to a disjunction from Q_0 conjunctions. Covering the domain A by the domain B simplifies this expression. Obviously, the larger value μ, the less accurate the approximation of the domain A by the domain B and the smaller the number of conjunctions needed to represent B. Further, the resulting logical expression is easily represented as a text expression. For this, you can use pre-prepared linguistic variables and templates.

At the last stage, we added to the potential ESZ the sites with target earthquake epicenters that were not detected using the decision rule.

3 Testing

Dynamic processes in the Caucasus region are determined by the convergence of the Arabian and Eurasian plates [16]. The seismicity of the region is largely determined by the zones of development of thrusts, shifts, and faults. Work [20] shows on models that strain to cause shifts in overthrusts, shifts, and faults are related approximately as 15:5:1. These zones can display the distribution of tectonic strains, and one can use them as a geological interpretation of the cause-effect regional model of strong earthquakes [8]. According to the model, it can be assumed that strong earthquakes are localized in the regions of the intersection of the inhomogeneities of the earth's crust. Let us consider an example of finding potential ESZ with target magnitudes $m \geq 6.0$. There are 20 epicenters with $m \geq 6.0$ in the area of analysis. Following to formula of Yu.V. Riznichenko [18], we assume that the projection of the target earthquake source zone onto the earth's surface is a circle with a radius of 16 km. We get that the area occupied by the centers of 20 earthquakes amounts to 15.5% of the analysis area. We will look for potential ESZ, which make up 25% of the analysis area.

The following primary data were used to solve the problem [11]: heights of the surfaces of the day relief, the consolidated foundation and Mohorovicic, the amplitude of neotectonics movements, isostatic, and deep gravity anomalies in the Bouguer reduction, the velocity gradient of vertical tectonic movements in the post-Sarmatian time (T = 17 million years), deviation of the travel times of the first primary waves and secondary waves in the upper mantle, heat flow, magnetic field anomalies, faults ranked by age, by the age of the last activation, by the type of movements and by the tectonic significance, as well as catalogs of earthquakes [1,14].

We analyzed the effectiveness of the solution using the primary fields listed above and a series of secondary fields. During the analysis, we compared the quality indicators of the solution obtained by the training sample and obtained through cross-validation. The best match of the indicators was shown by two calculated from primary fields of features: $\mathbf{F}_1$ is the field of absolute values of the velocity gradient of vertical tectonic movements in the post-Sarmatian time [1/year], and $\mathbf{F}_4(\mathbf{F}_2, \mathbf{F}_3)$, where $\mathbf{F}_2$ is the field of absolute values of the gradient of gravity anomalies in the Bouguer reduction, $\mathbf{F}_3$ is the field of absolute values of the gradient of the digital elevation model, the values of the field $\mathbf{F}_4$ grid nodes are $f_4 = \min\{(f_2 - \overline{f_2})/\sigma_2, (f_3 - \overline{f_3})/\sigma_3\}$, where f_2, f_3 are the values of the fields $\mathbf{F}_2$ and $\mathbf{F}_3$ in grid nodes, $\overline{f_2}$, $\overline{f_3}$ and σ_2, σ_3 mean and standard deviations within the area of analysis. The fields $\mathbf{F}_1$ and $\mathbf{F}_4$ are shown in Fig. 1.

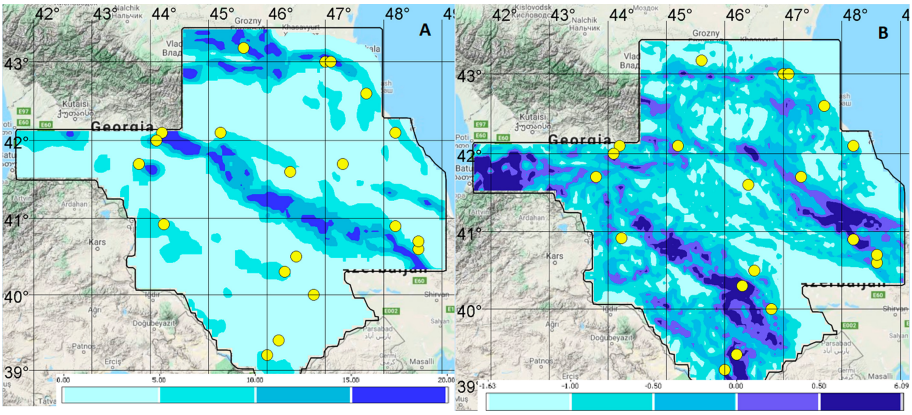

Fig. 1. Fields of features selected for the solution. (A) $\mathbf{F}_1$ is the field of absolute values of the velocity gradient of vertical tectonic movements in the post-Sarmatian time [1/year], (B) $\mathbf{F}_4$ is the field of a minimum of absolute values of the gradient of gravity anomalies in the Bouguer reduction and absolute values of the gradient of the digital elevation model.

Figure 2 shows the dependence $U(V)$ obtained by the training sample. It can be seen that with the alarm volume of 25%, the probability of hit of the target earthquakes epicenters in the potential ESZ is 80%. The following logical rule corresponds to the resulting solution:

IF

$$(f_1 \geq 3.37 \wedge f_4 \geq -0.36) \vee (f_1 \geq 10.07 \wedge f_4 \geq -1.11), \quad (1)$$

THEN

the earthquake sources with the magnitudes $m \geq 6.0$ are possible.

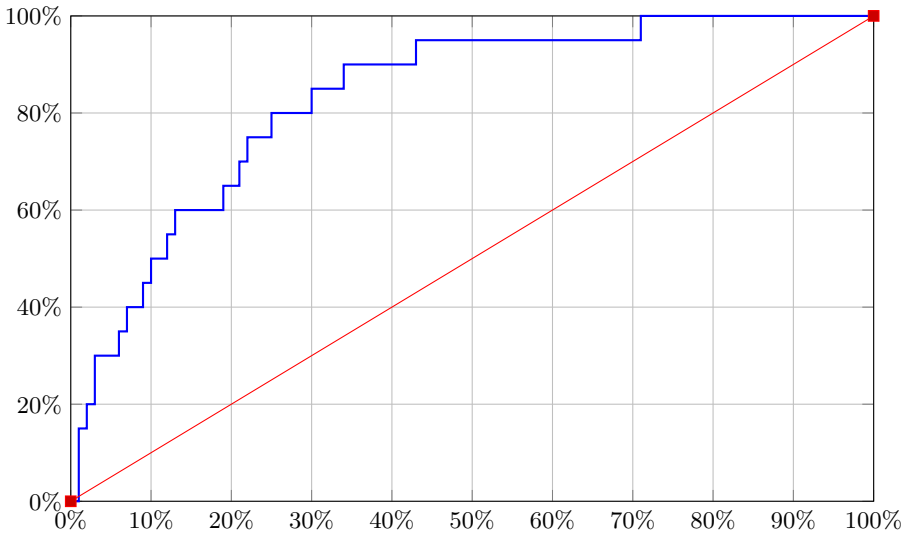

Fig. 2. The dependence $U(V)$ obtained by the training sample.

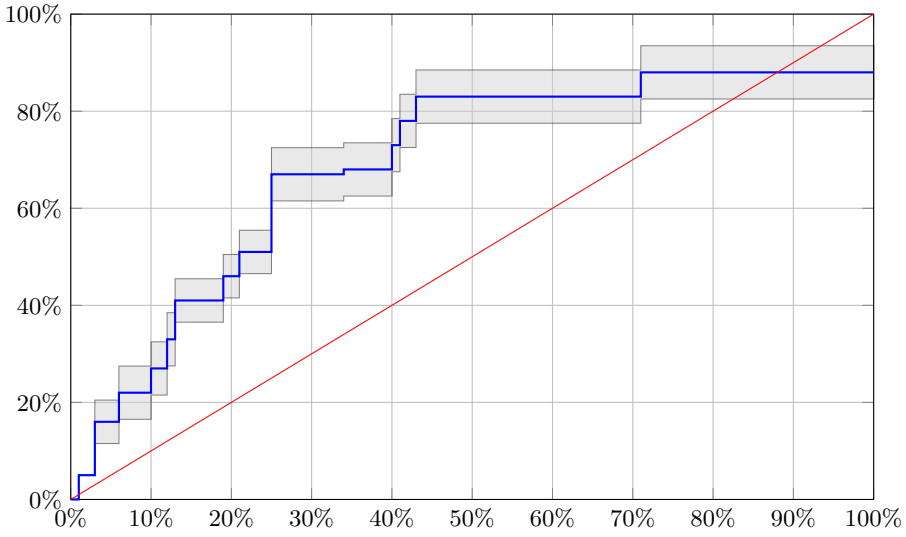

Fig. 3. The results of 5-fold cross-validation tests to verify the result with $k = 2$: mean and standard deviations of the $U(V)$ dependence.

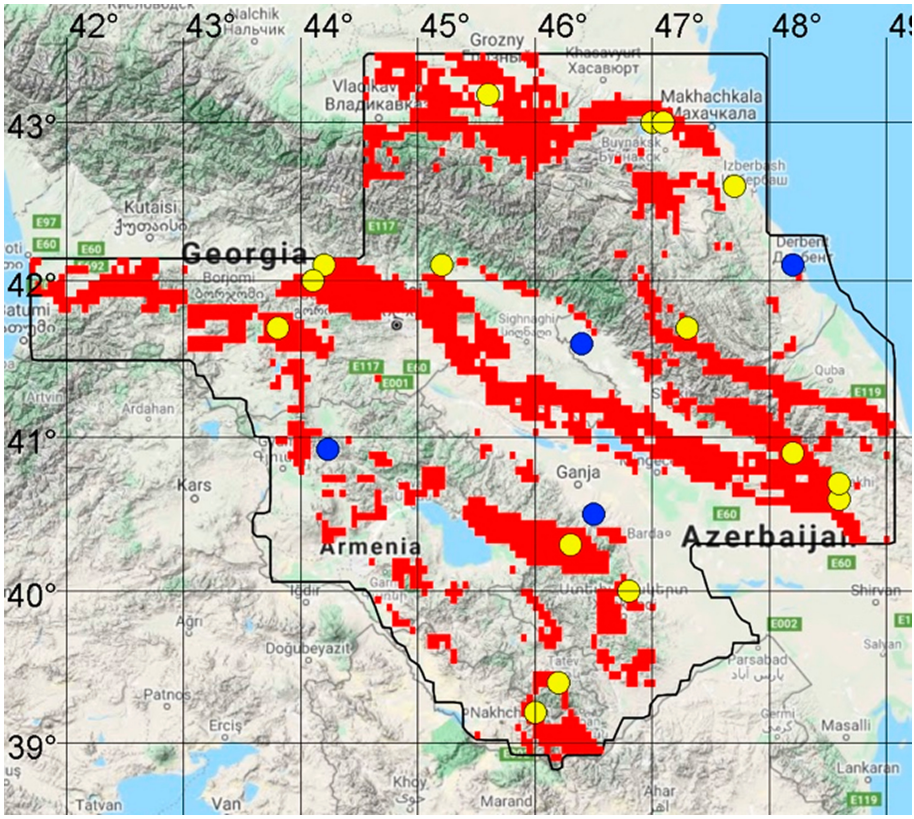

Fig. 4. The potential ESZ obtained by the decision rule 1. The yellow circles are the epicenters of earthquakes in the potential ESZ. The blue circles indicate the epicenters of earthquakes that were not detected by the resulting rule. (Color figure online)

We use k-fold cross-validation tests to verify the result with $k = 2$ [6]. In Fig. 3 shows the result for five tests: mean and standard deviations of the $U(V)$ dependence.

Figure 4 shows the potential ESZ obtained by the decision rule 1. The epicenters of the target earthquakes that hit the potential ESZ are shown in yellow. The blue color indicates the epicenters of earthquakes that were not detected by the resulting rule. The source areas corresponding to these epicenters should be complementary to those discovered by potential ESZ.

4 Conclusion

The data for solving the problem of identifying zones of potential sources of earthquakes consists of earthquake epicenters with target magnitudes and a set of grid-based spatial geophysical and geological fields. In our solution we use the

assumptions about the repeatability of strong earthquakes and the existence of a relationship between the properties of the geological environment and the locations of the earthquake epicenters. It follows that potential ESZ include places where earthquakes with target magnitudes were recorded, and areas identified by the found dependence. Errors arise due to the small number of examples of target earthquakes and due to the incomplete representation of the seismotectonic properties of the geological environment by the available feature fields.

In the article, we propose the method as a tool that can help a specialist to solve this important practical problem. The decision is divided into three stages. First, using the machine learning method proposed in the article, examples of target earthquakes that occurred in zones that differ in geological and geophysical data from other places in the region are selected. Then for these examples, a logical decision rule is calculated, which determines potential ESZ by the attribute fields. After that, the resulting map of potential ESZ are supplemented by the sources of the remaining target earthquakes. The decision rule allows specialists to give a geological interpretation of the solution.

Examples of the application of the method are performed for constructing zones with magnitudes $m \geq 6.0$ in the Caucasus region. We used the cross-validation procedure to select feature fields. The article provides two examples for which the quality of the forecast for cross-validation and training sample are the most successful.

We were not able to obtain convincing results when identifying potential ESZ with magnitudes $m \geq 5.5$. Possible reasons are the high magnitude of the background earthquakes in the Caucasus, which are possible anywhere in the analyzed area, and the lack of data on the properties of the geological environment.

References

1. Ananin, I.: European part of the USSR, Ural, West Siberia. In: A New Catalog of Strong Earthquakes in the USSR from Ancient Times to 1975, pp. 465–470 (1975). (in Russian)
2. Bishop, C.M.: Machine Learning and Pattern Recognition. Information Science and Statistics. Springer, Heidelberg (2006)
3. Bune, V., Gorshkov, G.: Seismic zonation of USSR, p. 307. Nauka, Moscow (1980)
4. Burton, P.W.: Seismic risk in southern Europe through to India examined using Gumbel's third distribution of extreme values. Geophys. J. Int. **59**(2), 249–280 (1979)
5. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
6. Geisser, S.: Predictive Inference, vol. 55. CRC Press, Boca Raton (1993)
7. Gelfand, I., et al.: Pattern recognition applied to earthquake epicenters in California. Phys. Earth Planet. Inter. **11**(3), 227–283 (1976)
8. Gitis, V., Ermakov, B.: Fundamentals of spatiotemporal forecasting in geoinformatics, p. 256. FIZMATGIS, Moscow (2004)
9. Gitis, V.G., Derendyaev, A.B.: Machine learning methods for seismic hazards forecast. Geosciences **9**(7), 308 (2019)

10. Gitis, V., Andrienko, G., Andrienko, N.: Exploration of seismological information in analytical web GIS. Izvestiya. Phys. Solid Earth **40**(3), 216–225 (2004)
11. Gitis, V., Ermakov, B., Ivanovskaya, L., Osher, B., Trofimov, D., Shchukin, Y.K.: The GEO expert system: application for seismic hazard analysis of the caucasus region. In: Cahiers du Centre Europeen de Geodynamique et de Seismologie. Centre européen de géodynamique et de séismologie (1992)
12. Khan, S.S., Madden, M.G.: A survey of recent trends in one class classification. In: Coyle, L., Freyne, J. (eds.) AICS 2009. LNCS (LNAI), vol. 6206, pp. 188–197. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-17080-5_21
13. Kijko, A.: Estimation of the maximum earthquake magnitude, m_{max}. Pure Appl. Geophys. **161**(8), 1655–1681 (2004)
14. Kondorskaia, N., Shebalin, N.: New Catalog of Strong Earthquakes in the USSR from Ancient Times Through 1977, vol. 31. World Data Center A for Solid Earth Geophysics (1982)
15. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: a review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. **160**, 3–24 (2007)
16. Philip, H., Cisternas, A., Gvishiani, A., Gorshkov, A.: The Caucasus: an actual example of the initial stages of continental collision. Tectonophysics **161**(1–2), 1–21 (1989)
17. Pisarenko, V.: Statistical evaluation of maximum possible earthquakes. Phys. Solid Earth **27**(9), 757–763 (1991)
18. Riznichenko, Y.V.: The source dimensions of the crustal earthquakes and the seismic moment. In: Issledovaniya po fizike zemletryasenii, pp. 9–27 (1976)
19. Sapozhenko, A., Chukhrov, I.: Boolean function minimization in the class of disjunctive normal forms. J. Sov. Math. **46**(4), 2021–2052 (1989)
20. Sibson, R.H.: Frictional constraints on thrust, wrench and normal faults. Nature **249**(5457), 542–544 (1974)
21. Soloviev, A., Novikova, O., Gorshkov, A., Piotrovskaya, E.: Recognition of potential sources of strong earthquakes in the Caucasus region using GIS technologies. Dokl. Earth Sci. **450**(2), 658 (2013). https://doi.org/10.1134/S1028334X13060159