Nat. Hazards Earth Syst. Sci., 8, 1207–1216, 2008
www.nat-hazards-earth-syst-sci.net/8/1207/2008/

Natural Hazards and Earth System Sciences

Detection of hydrogeochemical seismic precursors by a statistical learning model

L. Castellana and P. F. Biagi

Department of Physiscs, University of Bari, Via Amendola 173 – 70126 Bari, Italy

Received: 19 May 2008 – Revised: 16 September 2008 – Accepted: 16 September 2008 – Published: 10 November 2008

Abstract. The problem of detecting the occurrence of an earthquake precursor is faced in the general framework of the statistical learning theory. The aim of this work is both to build models able to detect seismic precursors from time series of different geochemical signals and to provide an estimate of number of false positives. The model we used is k-Nearest-Neighbor classifier for discriminating "no-disturbed signal", "seismic precursor" and "co-post seismic precursor" in time series relative to thirteen different hydrogeochemical parameters collected in water samples from a natural spring in Kamchachta (Russia) peninsula. The measurements collected are ion content (Na, Cl, Ca, HCO_3, H_3BO_3), parameters (pH, Q, T) and gases (N_2, CO_2, CH_4, O_2, Ag). The classification error is measured by Leave-K-Out-Cross-Validation procedure. Our study shows that the most discriminative ions for detecting seismic precursors are Cl and Na having an error rates of 15%. Moreover, the most discriminative parameters and gases are Q and CH_4 respectively, with error rate of 21%. The ions result the most informative hydrogeochemicals for detecting seismic precursors due to the peculiarities of the mechanisms involved in earthquake preparation. Finally we show that the information collected some month before the event under analysis are necessary to improve the classification accuracy.

1 Introduction

One of the most challenging problems in earth science is the simultaneous prediction of spatial and temporal localization as well as magnitude of earthquakes in order to mitigate damages to things and people. Although forecasting of seismic events remains an open issue and the focus of current debates (Geller et al., 1997; Wyss, 1997; Nature, 1999), many geophysical parameters have been investigated to face this problem.

Scientific efforts are directed toward the monitoring of changes in seismicity (Vorobieva et al., 1993), ground electrical resistivity and conductivity (Telesca et al., 2005), crustal deformation rates (Stephenson et al., 2003), ground water chemistry (Biagi et al., 2000; Guangcai et al., 2005), geomagnetic measurements (Rozhnoi et al., 2004; Hattori et al., 2004). Traditionally, the detection of these changes is performed by visual inspection of the filtered and smoothed time series (Kingsley et al., 2001) or of the spectral content of the signal (Biagi et al., 2006). Once the anomalous signal shapes satisfy the IASPEI validation criteria for precursor candidates (Biagi et al., 2000, 2001), the relation between signal anomalies and earthquakes is looked for. To this end, all possible phenomenon (i.e. meteorological conditions, volcanic activity, etc.) that could produce anomalies in the time series are investigated to be sure of associating the anomalies univocally to seismicity.
Although such approaches have successfully detected the association between anomalies in the signals analyzed and the occurrence of seismic events, they provide only qualitative answers and do not address the problem of measuring *quantitatively* the ability of a model to predict an event by analyzing geophysical signals.

In this paper we face two problems: a) to build models which are able to detect seismic precursors from time series of different geochemical signals, b) to assess the accuracy of such models on never seen before cases. We have applied well founded models and principles developed in the field of statistical learning theory (Vapnik, 1998), in which classifiers of events are built starting from the knowledge of *examples* of events that we are interested to discriminate. In this setting, an example is an (input, output) pair $(\boldsymbol{x}_t^m, y_t^m)$ connecting m observations of the independent variable $\boldsymbol{x}_t^m=(x_{t+1}, x_{t+2}, ..., x_{t+m})$ with the observation of the

Correspondence to: P. F. Biagi
(biagi@fisica.uniba.it)

Published by Copernicus Publications on behalf of the European Geosciences Union.

dependent variable $y_t^m = y_{t+m}$. Here $\{x_t\}_{t=1}^{\infty}$ is a real valued time series composed of daily observations of a given hydrogeochemical measurement and $\{y_t\}_{t=1}^{\infty}$ is time series of categorical variables with $y_t \in \{-1, 0, +1\}$, indicating the occurrence of "no-disturbed signal", "seismic precursor", "co-post seismic precursor" respectively. This association was established both on the basis of k_s index proposed by Molchanov et al. (2003) and the of ε index proposed by Dobrovolsky et al. (1979). These indexes take into account the relation between earthquake magnitude and distance from the epicenter to capital city Petropavlovsk. Then the problem is to build classifiers which are able to discriminate different types of events starting from a set $S=\{(\boldsymbol{x}_1^m, y_1^m), (\boldsymbol{x}_2^m, y_2^m), ..., (\boldsymbol{x}_\ell^m, y_\ell^m)\}$ composed of a finite and limited amount of examples which embodies the information or knowledge of the phenomena that we are interested to detect.

Under this perspective, the problem of seismic precursor detection or classification can be seen as *a supervised learning* problem, or *a learning from examples* problem in which the goal is to discriminate seismic precursors from no seismic ones, or to distinguish among different types of events, starting from the knowledge of a finite and limited number of examples. This approach has been successfully applied in several different application domains for solving actual problems such as object detection in images (Ancona et al., 2003), odor classification (Distane et al., 2003), electroencephalographic signal analysis on patients with hemicrania (Ancona et al., 2005), systolic pressure signal analysis (Ancona et al., 2005), statistical assessment of cancer predictors (Ancona et al., 2006). Independently of the specific applicative domain, it is worth to point out that the ultimate goal of this model is to determine the correct output relative to a never seen before input pattern, by using a training set composed of a finite number of examples.

Here we applied k-Nearest-Neighbor classifier to thirteen hydrogeochemical time series and we computed the classification accuracy by the Leave-K-Out-Cross-Validation procedure (Ambroise et al., 2002). We found that ions are more effective than other hydrogeochemicals in discovering seismic precursors. When m=100 observations are used, the estimated error rate is 15% for Cl and Na, 29% for H_3BO_3, 21% for HCO_3 and 35% for Ca; pH, Q and T report a prediction error of 50%, 21%, 34% respectively, when m=110 is set up; finally, the error on gases is 33% on average, choosing m=110.

The details about the models we applied and the experiments we carried out are in the following sections.

2 Theoretical framework

2.1 Seismic precursor detection as a machine learning problem

The general formulation of the learning problem involves three components (Vapnik, 1998): a *set* of input vectors $\boldsymbol{x} \in \mathbb{R}^m$ drawn independently from a fixed but unknown probability $p(\boldsymbol{x})$; a *set* of output values y for every input vector $\boldsymbol{x}$, according to the fixed condition density $p(y|\boldsymbol{x})$ that is also unknown; the *learning machine* capable of implementing a set of function $f(\boldsymbol{x}, \alpha), \alpha \in \Omega$, where Ω is a set of parameters used only to index the set of functions. The problem of learning is that of choosing from the given set of functions $f(\boldsymbol{x}, \alpha)$, the one which predicts the supervisor's response in the best possible way. The selection is based on a training set $S=\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$ of ℓ independent and identically distributed (i.i.d.) observations drawn according to $p(\boldsymbol{x}, y)=p(\boldsymbol{x})p(y|\boldsymbol{x})$. To this end, the *loss* or discrepancy $\mathcal{L}(y, f(\boldsymbol{x}))$ between the response y of the supervisor to a given input $\boldsymbol{x}$ and the response $f(\boldsymbol{x})$ provided by the leaning machine is measured. By definition, the *risk functional* or *generalization error*, is given by the expected value of the loss:

$$L[f]=\mathbb{E}\{\mathcal{L}(y, f(\boldsymbol{x}))\}=\int\int \mathcal{L}(y, f(\boldsymbol{x}))p(\boldsymbol{x}, y)d\boldsymbol{x}dy \quad (1)$$

so the ultimate goal of learning problem is to find the function $f(\boldsymbol{x})$ which minimize the risk functional $L[f]$. In the cases when $p(\boldsymbol{x}, y)$ is known and $\mathcal{L}$ is the form squared loss function, i.e.

$$\mathcal{L}(y, f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2$$

then the function minimizing the risk functional (1) is the regression function (Vapnik, 1999):

$$f^*(\boldsymbol{x}) = \int y\ p(y|\boldsymbol{x})\ dy$$

is the function minimizing the risk functional (1). Since in the real cases of data analysis problems, the probability distribution is unknown, the only available information is contained in the training set S. The best $f^*(\boldsymbol{x})$ will be the *machine* most capable to generalize, that is to predict the correct output y relative to a never seen before input pattern $\boldsymbol{x}$, by using a training set composed of a finite number of examples (Tibshirani et al., 2001). Thus the central problem is not classifying the training data in S, because any sufficiently complex learning machine could separate S without errors. The crucial problem is to design machines having low error rate on new data.

In the problem at hand, an example is an input-output pair $(\boldsymbol{x}_t^m, y_t^m)$ where the input variable $\boldsymbol{x}_t^m$ is composed of m consecutive observations, $\boldsymbol{x}_t^m=(x_{t+1}, x_{t+2}, ..., x_{t+m})$, of a real valued time series $\{x_t\}_{t=1}^{\infty}$ of daily observations of a

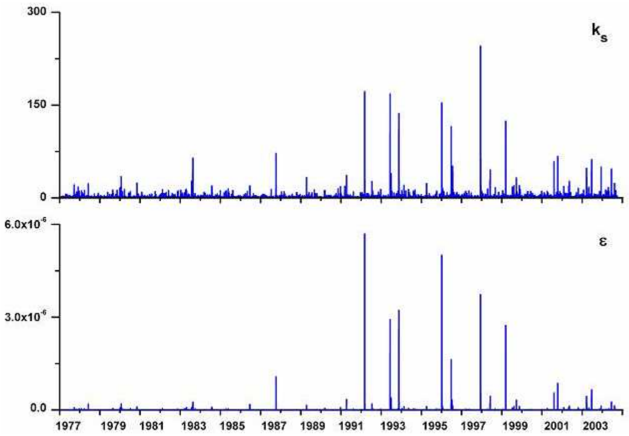


Fig. 1. Energy indexes of the earthquakes occurred from January 1977 to December 2004, related to the location of the capital city Petropavlovsk: at the top the trend of the k_s index and at the bottom the trend of the ε index.

given hydrogeochemical measurements and the output variable $y_t^m = y_{t+m}$ is a value of the time series of categorical variable $\{y_t\}_{t=1}^{\infty}$, obtained by exploiting k_s and ε indexes, with $y_t \in \{-1, 0, +1\}$, indicating the occurrence of "no-disturbed signal", "seismic precursor", "co-post seismic precursor" events respectively.

In the following two sections we will describe the class of predictors used in this study, their properties and a well founded statistical procedure for estimating the prediction accuracy of models by using the data in S.

2.2 kNN classifier

For simplicity, let us consider a two-class classification problem in which the label y_t^m can take values on $\{0, +1\}$. Let $\boldsymbol{x} \in \mathbb{R}^m$ be a sample to be classified by and reorder the ℓ i.i.d. examples in $S = \{(\boldsymbol{x}_1^m, y_1^m), (\boldsymbol{x}_2^m, y_2^m), ..., (\boldsymbol{x}_\ell^m, y_\ell^m)\}$ according to increasing values of $\|\boldsymbol{x}_i^m - \boldsymbol{x}\|$. The reordered data sequence is denoted by

$$(\boldsymbol{x}_{(1)}^m, y_{(1)}^m), (\boldsymbol{x}_{(2)}^m, y_{(2)}^m), ..., (\boldsymbol{x}_{(\ell)}^m, y_{(\ell)}^m)$$

and it is such that

$$\|\boldsymbol{x}_{(1)}^m - \boldsymbol{x}\| < \|\boldsymbol{x}_{(2)}^m - \boldsymbol{x}\| < ... < \|\boldsymbol{x}_{(\ell)}^m - \boldsymbol{x}\|$$

By this notation, $\boldsymbol{x}_{(1)}^m$ is the nearest neighbor of $\boldsymbol{x}$ and $\boldsymbol{x}_{(k)}^m$ is the $k-$th nearest neighbor of $\boldsymbol{x}$ in terms of the Euclidean distance defined on $\mathbb{R}^m$.
The *Nearest Neighbor* (NN) rule is defined as:

$$g_1(\boldsymbol{x}) = \begin{cases} 1 \text{ if } I_{\{y_{(1)}^m = 1\}} > I_{\{y_{(1)}^m = 0\}} \\ 0 \text{ if otherwise} \end{cases}$$

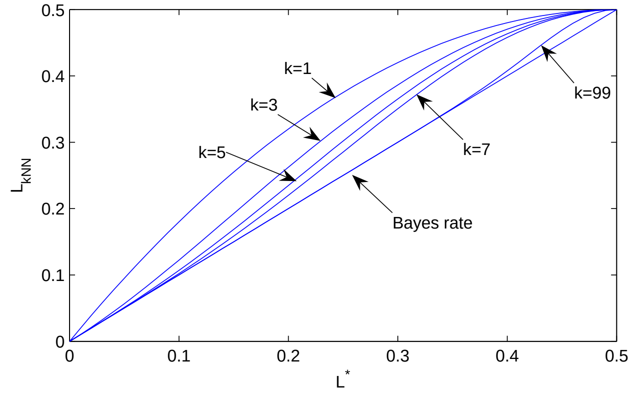


Fig. 2. Bounds on kNN error rates (L_{kNN}): as k increases, L_{kNN} gets progressively closer to the lower bound, the Bayes rate (L^*).

where I_A is the indicator function that takes value 1 when the event A occurs. This rule assigns to $\boldsymbol{x}$ the label $y_{(1)}^m$ of its nearest neighbor $\boldsymbol{x}_{(1)}^m$. In order to analyze the classification error performed by the NN rule, let us consider the loss function:

$$\mathcal{L}(y_{(1)}^m, y) = I_{\{g_1(\boldsymbol{x}) \neq y\}}$$

where y is the "true" label of $\boldsymbol{x}$. Then, by applying (1), the $\ell-$sample NN error is:

$$L(\ell) = \mathbb{E}\{\mathcal{L}(y_{(1)}^m, y)\} = \mathbb{P}\{y_{(1)}^m \neq y\}. \tag{2}$$

and the "large sample" NN risk becomes:

$$L_{\text{NN}} = \lim_{\ell \to \infty} L(\ell) \tag{3}$$

It is to be pointed out that NN rule utilizes only the classification of the nearest neighbor and the $\ell-1$ remaining examples $\boldsymbol{x}_{(i)}^m$ are ignored. If the number of samples is large, it makes sense to use, instead of the single nearest neighbor, the majority vote of the nearest k neighbors, and we have the *k-Nearest Neighbor* (kNN) rule:

$$g_{\text{kNN}}(\boldsymbol{x}) = \begin{cases} 1 \text{ if } \sum_{i=1}^{k} I_{\{y_{(i)}^m = 1\}} > \sum_{i=1}^{k} I_{\{y_{(i)}^m = 0\}} \\ 0 \text{ if otherwise} \end{cases}$$

In this case the loss function is:

$$\mathcal{L}(g_{\text{kNN}}(\boldsymbol{x}), y) = I_{\{g_{\text{kNN}}(\boldsymbol{x}) \neq y\}}$$

and the $\ell-$sample kNN error becomes:

$$L_k(\ell) = \mathbb{E}\{\mathcal{L}(g_{\text{kNN}}(\boldsymbol{x}), y)\} = \mathbb{P}\{g_{\text{kNN}}(\boldsymbol{x}) \neq y\}. \tag{4}$$

Analogously to (3), in the case of "large sample", the kNN risk is:

$$L_{\text{kNN}} = \lim_{\ell \to \infty} L_k(\ell) \tag{5}$$

It is to be noted that in the case of multiple-class classification problem, the label of the sample is assigned according the majority vote rule.

Although simple, kNN classifiers enjoy very interesting asymptotic properties (Cover and Hart, 1967; Devroye et al., 1996) which make it a model suitable for facing real prediction problems. The most important being that for large values of sample size ℓ, the algorithm is guaranteed to yield an error rate no worse than twice the Bayes error which is minimum achievable error rate. If we indicate with L^* the Bayes error (Duda et al., 1996) defined as:

$$L^* = \begin{cases} 0 \text{ if } \mathbb{P}\{y=0|\boldsymbol{x}\} \geq \mathbb{P}\{y=1|\boldsymbol{x}\} \\ 1 \text{ if otherwise} \end{cases}$$

then, for large ℓ, the risk L_{kNN} is related to the minimum achievable risk L^* of the Bayes classifier through the two sided inequalities (Devroye et al., 1996):

$$L^* \leq ... \leq L_{(2k+1)\mathrm{NN}} \leq L_{(2k-1)\mathrm{NN}} \leq ... \leq L_{3\mathrm{NN}} \leq L_{\mathrm{NN}} \leq 2L^*(1-L^*) \leq 2L^*.$$

Two notes have to be remarked: first the upper and lower bounds are in general as tight as possible, since the Bayes classifier is the best one (Duda et al., 1996); second, the previous inequalities show that as k increases, the upper bounds get progressively closer to the lower bound (see Fig. 2) and, as k goes to infinity, the two bounds meet, making the kNN rule *optimal*.

2.3 Assessment of the prediction accuracy

As we have already pointed out in Sect. (2.1), for measuring the performances of a learning machine we have to estimate the generalization error or risk $L[f]$ in (1) which is the probability to correctly classify new input pattern.
A common procedure used for estimating the risk is the Leave-One-Out (LOO) error (Luntz and Brailovsky, 1969). This procedure provides an almost unbiased estimate of the risk (1) and it allows of assessing the performances of a supervised learning machine from a finite number of data. The computation of LOO is very simple: for every i=1, ..., ℓ, let f_{S^i} be the machine trained on the set S^i={($\boldsymbol{x}_1$, y_1), ..., ($\boldsymbol{x}_{i-1}$, y_{i-1}), ($\boldsymbol{x}_{i+1}$, y_{i+1}), ...($\boldsymbol{x}_\ell$, y_ℓ)} obtained from S removing the $i-$th sample. Then the function f_{S^i} is tested on the left out example $(\boldsymbol{x}_i, y_i)$ and the value of the loss function $\mathcal{L}(y_i, f_{S^i}(\boldsymbol{x}_i, \alpha))$ is measured. Finally, this procedure is repeated on each of the ℓ examples of S and the LOO error is given by the sum of the errors, i.e.:

$$\mathcal{L}(S) = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathcal{L}(y_i, f_{S^i}(\boldsymbol{x}_i))$$

Although the bias of LOO is low, it is highly variable and has a little statistical significance (Ancona et al., 2006). On the contrary, the Leave-K-Out cross validation (LKOCV) error is more significant and so it makes sense using it to measure the accuracy of the classifier (Ancona et al., 2006; Mukherjee et al., 2006). Its computation is similar to LOO error, with the exception that k examples are randomly removed from S. It is very expensive, because we should repeat the procedure for $\ell-$choose$-k$ possible trials. To make it more feasible, a sampling procedure is adopted. In order to explain it, let p be the training set size, with p=1, 2, ..., $\ell-1$ and let $q=\ell-p$ be the resulting test set size. We build s pairs (D_p, T_q) of training and test sets with p and q examples, respectively, by random sampling without replacement the data set S. For each of these s random splits, we evaluate the error rate e_{p_i} of the classifier trained on D_p examples and test it on T_q. So, the LKOCV error e_p is given by:

$$e_p = \frac{1}{s}\sum_{i=1}^{s} e_{p_i}.$$

The same procedure has been applied for estimating the free parameter (i.e. the number of neighbors) in kNN classifier.

3 Dataset description

In this study we have analyzed thirteen hydrogeochemical time series relative to daily measurements of ion content (Na, Cl, Ca, HCO_3, H_3BO_3), parameters (pH, Q, T) and gases (N_2, CO_2, CH_4, O_2, Ag) in water samples collected from a “natural” spring (S1) 50 km far away Petropavlovsk, Kamchachta peninsula (Russia).

The Kamchatka peninsula is an active margin where the Pacific plate subducts beneath the North American and Eurasia plate. More than 80 volcanoes exist and many of them are active. The relative plate motion changes from underthrusting of the Pacific plate at the Kuril-Kamchatka arc to strike slip motion along the Aleutin arc at the junction of the Kamchatka and Aleutian trench.

The majority of earthquakes occur in a zone located offshore 60–100 km south-east of the Pacific coast of the peninsula with focal depths up to 650 km; the direction of the maximum extension of their isoseismals is parallel to the east coast of the peninsula. In this zone, earthquakes with magnitudes up to 9.0 could occur. Earthquakes also happen in the continental part of Kamchatka, but with a frequency much less than in the subduction zone. Basically these continental earthquakes are related to volcanic activity, their magnitude rarely exceeds 6.0 and their focal depth is not more than 50 km.

The hydrogeochemical time series we used in this study have been collected by the Geophysical Service of Kamchatka since 1977 to 2004. In particular, the measurements of Na, Cl, HCO_3 and HBO_3 ions started in 1977 and, up to 29 December, 2004 these time series are composed of 10 224 data; Ca measurements started in 1987 and the available data are 6297 (see Fig. 3).

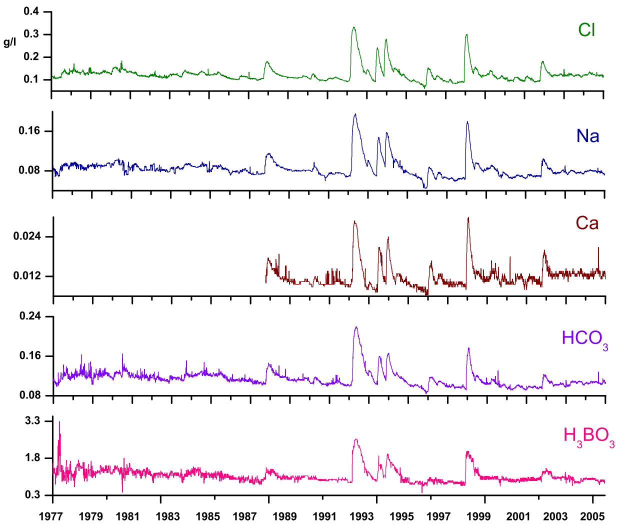


Fig. 3. Time series of Na, Cl, HCO_3 and HBO_3 ions, collected from 2 January 1977 to 29 December 2004. The Ca time series were collected from 4 October 1987 to 29 December 2004.

The time series of pH, Q and T, acquired in the period 1977–2004, are composed of 10 224 data (see Fig. 4).

Finally, the gases have been collected since July 1984 up to 1998 and the total number of data is 5282 (see Fig. 5).

The Na and Ca concentration was measured by flame emission spectrometry; the Cl, HCO_3 and SO_4 concentration by titration methods; the pH value by pH-meter; the content of dissolved gases after thermo-vacuum degassing was measured by means of gas chromatography. The accuracy of the measurements ranges from 2% to 10%. Generally a sampling frequency of three days was used for the hydrogeochemical measurements although sometimes a frequency of six days and rarely of one day was used. Finally, the dissolved ions and gases listed above were recorded at S1 only if the value of the content was over the sensitivity of the measurement method.

The composition of each time series with respect the number of events for each class is reported in Table 1.

Each time series is used for building a data set $\mathcal{D}_n^m=\{(\boldsymbol{x}_1^m, y_1^m), (\boldsymbol{x}_2^m, y_2^m), ..., (\boldsymbol{x}_n^m, y_n^m)\}$ composed of n examples, where n depends both on the measurements in the time series and on the model order m (see Sect. 2.1). The set of labels is obtained on the basis of k_s index proposed by Molchanov et al. (2003) and the of ε index proposed by Dobrovolsky et al. (1979), shown in Fig. 1. In order to associate the correct label to each input vector, the Seismic Bulletin of Kamchachta concerning the localization, magnitude and time of earthquakes was looked up too.

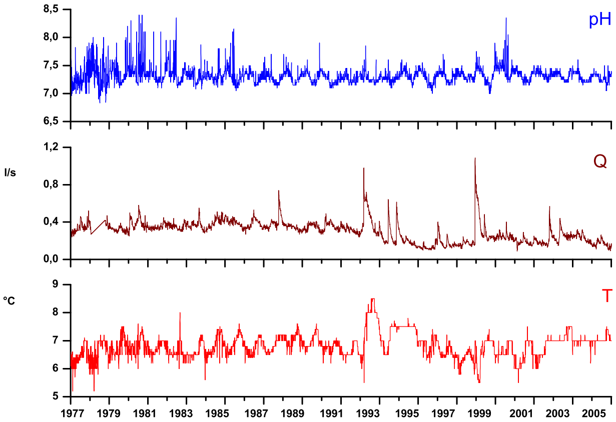


Fig. 4. Time series of pH, Q and T parameters, collected from 2 January 1977 to 29 December 2004.

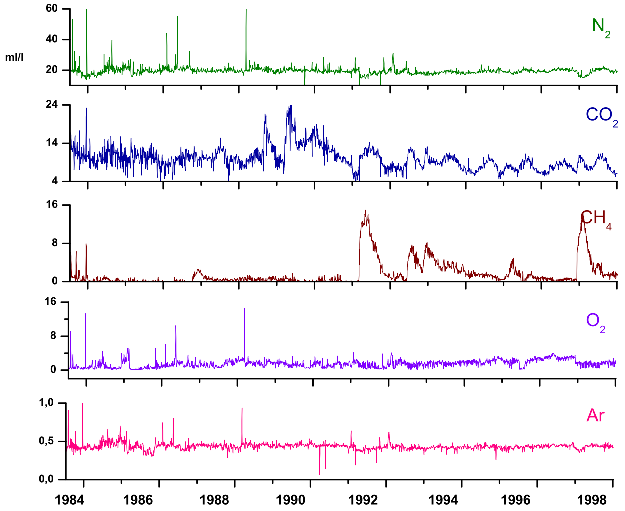


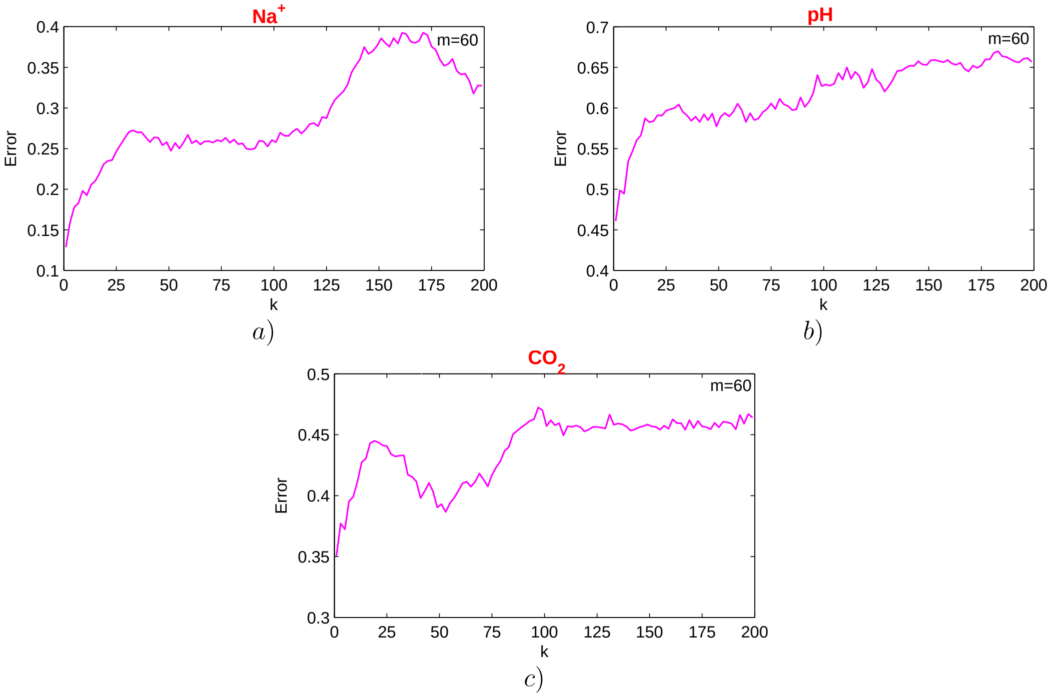
Fig. 5. Time series of N_2, CO_2, CH_4, O_2 and Ar gases, collected from 13 July 1984 to 29 December 1998.

4 Experimental results

We estimated the classification accuracy of kNN model for each dataset $\mathcal{D}_n^m$ associated to each time series for different model order m. To this end we split $\mathcal{D}_n^m$ into a training and validation set. The first one, composed of ℓ=300 examples (100 examples for each class) randomly chosen from $\mathcal{D}_n^m$, was used for selecting the best classifier from the available data. The validation set, composed of data not belonging to the training set, was used for assessing the performances of the selected classifier on never seen before data. It is worth noting that the data in the validation set are not used for selecting the free parameter k in kNN classifier. This ensures that the estimate of the generalization error obtained is unbiased.

Table 1. Number of examples belonging to “no-disturbed signal”, “seismic precursor” and “co-post seismic precursor” class in the thirteen time series analyzed.

Time series	Number of examples			Total
	no-disturbed signal	seismic precursor	co-post seismic precursor	
Na, Cl, Ca, HCO_3, H_3BO_3, pH, Q, T	8497	1277	450	10 224
Ca	4570	1277	450	6297
N_2, CO_2, CH_4, O_2, Ag	3756	1076	450	5282



a)

b)

c)

Fig. 6. Test errors, obtained by averadging the error rates on 500 cross-validations on a classifier trained on D_p and tested on T_q, are plotted versus k, the parameter of the kNN classifier. The trend is reported in **(a)** for Na, **(b)** for pH and **(c)** for CO_2, at m=60 and using p=240 examples in training and q=60 in test. For each k, a kNN classifier is obtained. The best is the one minimizing the test error.

The best model was determined as the kNN classifier minimizing the LKOCV error measured on the training set. To this end, for each value of the k parameter in a suitable range, 500 cross validations of the examples in the training set were carried out. In each cross validation, the training set was split in two sets D_p and T_q composed of p and $q=\ell-p$ examples respectively. The error rate of kNN classifiers trained on D_p and tested on T_q was evaluated. The best parameter k was selected as the one minimizing the average error rate. To elucidate the role of the k in kNN classifiers, we show the behavior of the error rate as a function of k in the case of Na ion, pH parameters and CO_2 gas (see Fig. 6a, b and c respectively). The test error exhibits a minimum always at small values of k pointing out that good generalization capability can be reached by considering the labels of few neighbors of the test pattern.

In order to understand the dependence of the error rate on the number p of the examples used in the learning phase, we fixed the model order m and we carried out experiments in which p and q varied. Then we repeated all the experiments for different values of m in the range $[10, \; 120]$. The main reason for varying the value of m was to establish how many observations have to be analyzed for detecting precursors and co-post seismic events in hydrogeochemical time series. For safe of clearness, we underline that in each experiment where m is fixed, we selected the best kNN classifier as explained before and we associated the value of p to the minimum test error. In Figs. 7, 8 and 9, error rates on ions, parameters and gases respectively have been plotted versus different values of p, expressed in percentage, in the case of m=60. As the pictures show, increasing the number p of examples used for training, the error rate decreases of more than 10%. The best

Table 2. Classification error on ions, evaluated both on balanced dataset (*Test* #1) and on unbalanced one (*Test* #2), by varying the length m of model order. The last row of the table reports the percentage difference error rate obtained on each time series of ions.

m	Error on *Test* #1					Error on *Test* #2				
	Cl	Na	Ca	H_3BO_3	HCO_3	Cl	Na	Ca	H_3BO_3	HCO_3
10	0.30	0.29	0.51	0.32	0.31	0.25	0.35	0.47	0.42	0.47
20	0.25	0.25	0.49	0.25	0.26	0.24	0.38	0.40	0.36	0.43
30	0.21	0.18	0.38	0.28	0.23	0.22	0.28	0.44	0.34	0.39
40	0.20	0.15	0.34	0.23	0.21	0.24	0.27	0.46	0.33	0.32
50	0.14	0.19	0.34	0.26	0.19	0.26	0.29	0.43	0.28	0.36
60	0.11	0.13	0.42	0.19	0.19	0.21	0.22	0.51	0.27	0.30
70	0.16	0.13	0.32	0.17	0.16	0.24	0.24	0.44	0.25	0.28
80	0.10	0.16	0.27	0.19	0.23	0.16	0.33	0.38	0.28	0.33
90	0.14	0.15	0.27	0.22	0.21	0.19	0.24	0.38	0.27	0.31
100	0.06	0.06	0.26	0.16	0.12	0.16	0.15	0.35	0.29	0.21
110	0.06	0.07	0.23	0.15	0.12	0.15	0.14	0.33	0.23	0.20
120	0.14	0.07	0.25	0.14	0.09	0.17	0.14	0.38	0.19	0.52
					$\triangle_{\%}$Err					
	80%	79%	55%	56%	71%	32%	60%	19%	55%	68%

Table 3. Classification error on parameters, evaluated both on balanced dataset (*Test* #1) and on unbalanced one (*Test* #2), by varying the length m of model order. The last row of the table reports the percentage difference error rate obtained on each time series of parameters.

m	Error on *Test* #1			Error on *Test* #2		
	pH	Q	T	pH	Q	T
10	0.60	0.34	0.53	0.64	0.44	0.58
20	0.52	0.33	0.46	0.58	0.40	0.42
30	0.53	0.31	0.42	0.64	0.16	0.45
40	0.48	0.23	0.51	0.56	0.32	0.51
50	0.45	0.19	0.40	0.54	0.28	0.37
60	0.47	0.27	0.34	0.54	0.40	0.34
70	0.45	0.15	0.27	0.54	0.23	0.37
80	0.39	0.17	0.26	0.55	0.30	0.32
90	0.38	0.15	0.22	0.49	0.20	0.36
100	0.35	0.16	0.24	0.52	0.23	0.39
110	0.36	0.16	0.24	0.50	0.21	0.34
120	0.36	0.11	0.23	0.50	0.17	0.38
			$\triangle_{\%}$Err			
	42%	68%	58%	22%	61%	45%

performances were obtained with p=240, that is the 80% of training set size. In fact, although the error rate decreases in the range [80%, 90%], it is constant in [70%, 80%] and the choice of 80% protects us from the problems of overfitting. Since this behavior holds for all the m tested, we selected p=240 as the best number of training examples to use for assessing the performances of the model.

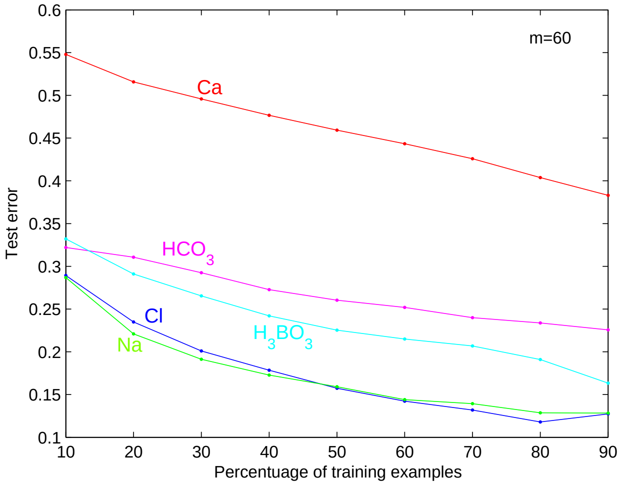


Fig. 7. The Leave-K-Out error in the five ion time series is plotted versus the percentage of training examples, at m=60.

The problem of studying the performances of the selected kNN model on data not used in the training phase was faced by using two validation sets: *Test* #1, a balanced set composed of 900 examples (300 examples for each class), and *Test* #2, an unbalanced set composed of *all* the data belonging to $\mathcal{D}_n^m$ and not used in the training. Having determined the optimal number of training examples (p=240), we evaluated the error rate of the best kNN classifier on *Test* #1 and *Test* #2, for a fixed model order m. Different values of m, ranging from 10 to 120, have been used to establish how many observations have to be used to build x_t^m such that the

Table 4. Classification error on gases, evaluated both on balanced dataset (*Test* #1) and on unbalanced one (*Test* #2), by varying the length m of model order. The last row of the table reports the percentage difference error rate obtained on each time series of gases.

m	Error on *Test* #1					Error on *Test* #2				
	N_2	CO_2	CH_4	O_2	Ar	N_2	CO_2	CH_4	O_2	Ar
10	0.49	0.45	0.31	0.44	0.53	0.50	0.55	0.34	0.52	0.53
20	0.44	0.39	0.31	0.41	0.47	0.50	0.40	0.34	0.51	0.50
30	0.38	0.37	0.28	0.35	0.49	0.45	0.52	0.36	0.50	0.53
40	0.41	0.36	0.20	0.35	0.42	0.48	0.48	0.34	0.48	0.46
50	0.41	0.33	0.24	0.34	0.42	0.45	0.51	0.37	0.45	0.45
60	0.36	0.28	0.20	0.32	0.38	0.45	0.45	0.32	0.42	0.45
70	0.34	0.27	0.18	0.30	0.35	0.43	0.43	0.32	0.38	0.43
80	0.35	0.30	0.27	0.28	0.33	0.39	0.45	0.31	0.37	0.39
90	0.33	0.25	0.15	0.27	0.40	0.41	0.41	0.26	0.38	0.44
100	0.31	0.23	0.12	0.25	0.30	0.40	0.38	0.26	0.38	0.38
110	0.31	0.23	0.12	0.24	0.31	0.36	0.41	0.21	0.36	0.34
120	0.22	0.22	0.12	0.24	0.35	0.34	0.41	0.22	0.34	0.39
					$\triangle_{\%}$Err					
	55%	51%	61%	45%	43%	32%	31%	38%	35%	36%

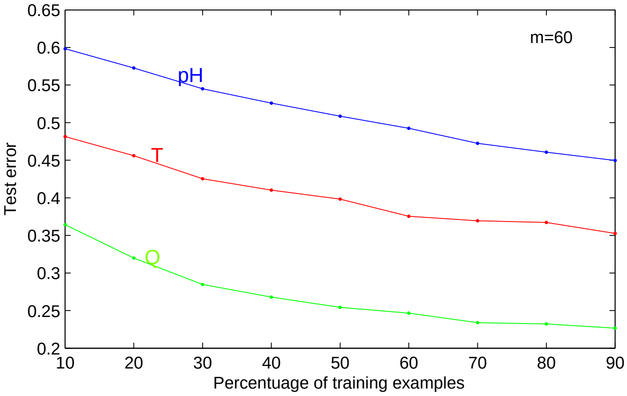


Fig. 8. The Leave-K-Out error in the three parameter time series is plotted versus the percentage of training examples, at m=60.

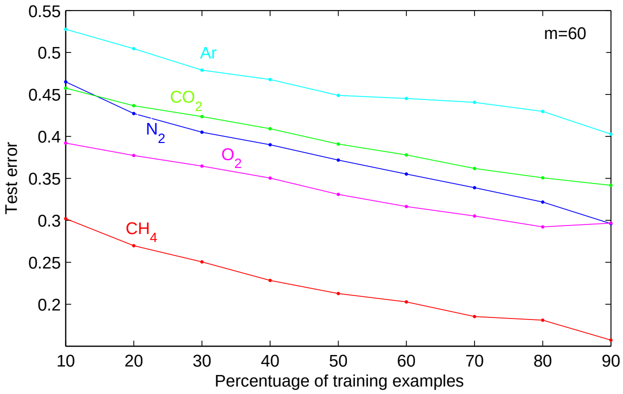


Fig. 9. The Leave-K-Out error in the five gas time series is plotted versus the percentage of training examples, at m=60.

classifier accuracy should be as high as possible. Test errors on ions, parameters and gases are reported in Tables 2, 3 and 4 respectively. The results show that, under the same m, the test errors obtained on *Test* #1 are comparable to the ones obtained in the training phase. This behavior indicates that the error rates estimated during the learning phase of a kNN classifier are a good estimates of the errors on never seen before examples, in the case of balanced data. In fact, as the results on *Test* #2 show, the error rates in testing is noticeably grater than the ones in training. Moreover, we observe that the classification error decreases by increasing m, both on *Test* #1 and *Test* #2 such as during the training phase. This confirms the behavior of the error rate as a function of m estimated during training (data not shown). The minimum error is obtained at m=100 for ions and at m=110 for parameters and gases. In particular, when the balanced validation set is used, the error is lesser than 15% for ions (Ca excepted); lesser than 25% for parameters (pH excepted) and lesser than 30% for gases. In the case of unbalanced validation set, the highest error (50%) is reached by pH and the lowest one (15%) by Na.

By looking at the behavior of the error rate on each time series, the results show high variability in terms of percentage difference, defined as $\triangle_{\%}\text{Err}=\frac{\text{Err}_{\max}-\text{Err}_{\min}}{\text{Err}_{\max}}\times 100$ (see last row of Tables 2, 3 and 4). In details, on *Test* #1 we annotate it varies from 80% (Cl) to 55% (Ca) for ions, from 68% (Q) to 42% (pH) for parameters, from 61% (CH_4) to 43% (Ar) for gases; on *Test* #2 we annotate it varies from 68%

(HCO_3) to 19% (Ca) for ions, from 61% (Q) to 22% (pH) for parameters, from 38% (CH_4) to 31% (CO_2) for gases.

By analyzing the difference between the maximum and minimum error rate of the classifiers by varying m (see the last row of Tables 2, 3 and 4) we note that kNN classifiers are poorly influenced by the model order m if they are trained on Ca, among ions; on pH, among parameters; on Ar (in the case of balanced data) and on CO_2 (in the case of unbalanced data) among gases.

5 Discussion and conclusions

In this paper we address the problem of detecting hydrogeochemical seismic precursors by a machine learning approach, in which short temporal windows of the original signal are classified. Under this perspective the classification error estimated by LKOCV provides a quantitative measure of the capability of the kNN classifier in detecting no-disturbed signal/seismic precursor/co-post seismic signal. It mainly depends on two factors: the training set size and the model order m. As experimental results show, for a fixed m, the generalization error decreases increasing the number of training data. It reaches a plateau with ℓ=240. This shows that the size of our data set is more than enough for training accurate classifiers of this type of signals.

Concerning the number m of consecutive measurements to build x_t^m, the results show that information collected some months before the event under analysis are necessary to improve the classification accuracy. In other words, it means that to detect precursors in the ion, parameters and gas time series we need to analyze m=100, m=110 and m=110 observations respectively.

In the light of the previous considerations, it follows that ions are the most informative hydrogeochemicals for detecting seismic precursors. This behavior can be justified by taking into account the peculiarities of the mechanisms involved in earthquake preparation. In fact, the fractures and cracks due to earthquakes produced the mixing of spring groundwater with new waters characterized by larger salinity (Biagi et al., 2006). As a consequence, deep changes in Na and Cl water content are produced and these ions appear very useful in detecting seismic precursors, since they have the lower classification error. Also Q is strongly influenced by this process and, among the hydrogeochemical parameters, it can be considered a good time series to train kNN classifiers with low error rates. Regarding gases, in active tectonic environments, methane is released close to major crustal discontinuities. In particular, CH_4 bearing spring gases are often present in volcanic environments (Toutain and Baubron, 1999). It could justify why CH_4 is the most indicative hydrogeochemical gas in discriminating these types of signal patterns. Finally, the gases have higher classification errors than parameters and ions. This is due to the fact that the time series analyzed are shorter in time. It means that the availability of a large amount of data improves the classification accuracy. The same justification is true for Ca time series that exhibits, at m=100, an error rate equal to 26% on balanced data and 35% on unbalanced one.

In conclusion, we have proposed a quantitative approach to detect different seismic events, building classifiers by using a finite number of observations and estimating their classification accuracy. The analysis was carried out on each time series related to hydrogeochemical measures in groundwater content and the results suggest that, starting from the knowledge of a finite and limited number of examples, it is possible to detect the occurrence of seismic precursors. In order to better investigate how to improve the classification accuracy, we plan to study the mutual correlations among these time series as future developments of this work.

Edited by: M. Contadakis
Reviewed by: two anonymous referees

References

Ancona, N., Cicirelli, G., Stella, E., and Distante, A.: Ball detection in static images with support vector machines for classification, Image and Vision Computing, 21(8), 675–692, Elsevier, 2003.

Ancona, N., Angelini, L., De Tommaso, M., Marinazzo, D., Nitti, L., Pellicoro, M., and Stramaglia, S.: Measuring randomness by leave-one-out prediciton error analysis of EEG after painful stimulation, Physica A, 365, 491–498, 2006.

Ancona, N., Maestri, R., Marinazzo, D., Nitti, L., Pellicoro, M., Pinna, G. D., and Stramaglia, S.: Leave-one-out prediciton error of systolic arterial pressure time series under paced breathing, Physiol. Meas., 26, 363–372, 2005.

Ancona, N., Maglietta, R., Piepoli, A., D'Addabbo, A., Cotugno, R., Savino, M., Liuni, S., Carella, M., Pesole, G., and Perri, F.: On the statistical assessment of classifiers using dna microarray data, BMC Bioinformatics, 7, 387–395, 2006.

Ambroise C. and McLachlan, G.: Selection bias in gene extraction on the basis of microarray gene-expression data, PNAS, 99, 6562–6566, 2002.

Biagi, P. F., Ermini, A., Kingsley, S. P., Khatkevich, Y. M., and Gordeev, E. I.: Groundwater ion content precursors of strong earthquakes in Kamchatka (Russia), Pure and applied Geophysiscs, 157, 1359–1377, 2000.

Biagi, P. F., Ermini, A., Cozzi, E., Khatkevich, Y. M., and Gordeev, E. I.: Hydrogeochemical precursors in Kamchatka (Russia) related to the strongest earthquake in 1988–1997, Natural Hazard, 21, 263–276, 2000.

Biagi, P. F., Piccolo, R., Ermini, A., Fujinawa, Y., Kingsley, S. P., Khatkevich, Y. M., and Gordeev, E. I.: Hydrogeochemical precursors of strong earthquakes in Kamchatka: further analysis, Nat. Hazards Earth Syst. Sci., 1, 9–14, 2001, http://www.nat-hazards-earth-syst-sci.net/1/9/2001/.

Biagi, P. F., Castellana, L., Minafra, A., Maggipinto, G., Maggipinto, T., Ermini, A., Molchanov, O., Khatkevich, Y. M., and Gordeev, E. I.: Groundwater chemical anomalies connected with the Kamchatka earthquake (M=7.1) on March 1992, Nat. Hazards Earth Syst. Sci., 6, 853–859, 2006, http://www.nat-hazards-earth-syst-sci.net/6/853/2006/.

Cover T. M. and Hart, P. E.: Nearest neighbor pattern classification, IEEE Transactions on information theory, 13(1), 21–27, 1967.

Devroye, L., Gyorfi, L., and Lugosi, G.: A probabilistic theory of pattern recognition, Springer Verlag, New York, 1996.

Distante, C., Ancona, N., and Siciliano, P.: Support vector machines for olfactory signals recognition, Sensors and Actuators B, 88(1), 30–39, 2003.

Dobrovolsky, I. P., Zubkov, S. I., and Miachkin, V. I.: Estimation of the size of earthquake preparation zone, Pure and applied Geophysiscs, 117, 1025–1044, 1979.

Duda, R. O., Hart, P. E., and Stork, D. G.: Pattern classification (2nd ed), John Wiley and Sons, INC, 2000.

Geller, R. J., Jackson, D. D., Kagan, Y. Y., and Mulargia, F.: Earthquakes cannot be predicted, Science, 275, 1616–1617, 1997.

Guangcai, W., Zuochen, Z., Min, W., Cravotta, C. A., and Chenglong, L.: Implications of ground water chemistry and flow patterns for earthquake studies, Ground Water, 43(4), 478–484, 2005.

Hattori, K., Takahashi, I., Yoshino, C., Isezaki, N., Iwasaki, H., Harada, M., Kawabata, K., Kopytenko, E., Kopytenko, Y., Maltsev, P., Korepanov, V., Molchanov, O. A., Hayakawa, M., Noda, Y., Nagao, T., and Uyeda, S.: ULF geomagnetic field measurements in Japan and some recent results associated with Iwateken Nairiku Hokubu earthquake in 1998, Phys. Chem. Earth, 29, 481–494, 2004.

http://www.nature.com/nature/debates/earthquake/equake_frameset.html, Nature, Macmillan Publishers Ltd 1998 Registered No. 785998 England, 1999.

Kingsley, S. P., Biagi, P. F., Piccolo, R., Capozzi, V., Ermini, A., Khatkevich, Y. M., and Gordeev, E. I.: Hydrogeochemical precursors of strong earthquakes: a realistic possibility in Kamchatka, Phys. Chem. Earth (C), 26(10–12), 769–774, 2001.

Luntz, A. and Brailovsky, V.: On estimation of characters obtained in statistical procedure of recognition, Technicheskaya Kibernetica, 3, 563–575, 1969.

Molchanov, O., Schekotov, A., Fedorov, E., Belyaev, G., and Gordeev, E.: Preseismic ULF electromagnetic effect from observation at Kamchatka, Nat. Hazards Earth Syst. Sci., 3, 203–209, 2003,
http://www.nat-hazards-earth-syst-sci.net/3/203/2003/.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T., and Mesirov, J.: Estimating dataset size requirements for classifying dna microarray data, J. Comp. Biol., 10, 119–142, 2003.

Rozhnoi, A., Solovieva, M. S., Molchanov, O. A., and Hayakawa, M.: Middle latitude LF(40 kHz) phase variations associated with earthquakes for quiet and disturbed geomagnetic conditions, Phys. Chem. Earth, 29, 589–598, 2004.

Stephenson, R. A., Narkiewicz, M., Dadlez, R., van Wees, J. D., and Andriessen, P.: Tectonic subsidence modelling of the polish basin in the light of new data on crustal structure and magnitude of inversion, Sediment. Geol., 156, 59–70, 2003.

Telesca, L., Balasco, M., Colangelo, G., and Lapenna, V.: Analysis of extreme events in geoelectrical time series measured in a seismic area of southern appenine chain Italy, Natural Hazard and Earth System, 34, 177–198, 2005.

Tibshirani, R., Hastie, T., and Friedman, J.: The elements of Statistical Learning, Springer, 2001.

Toutain, J. P. and Baubron, J. C.: Gas geochemistry and seismotectonics: a review, Tecnophisics, 304, 1–27, 1999.

Vapnik, V.: Statistical learning theory, John Wiley and Sons, INC, 1998.

Vapnik, V.: An overview of statistical learning theory, IEEE Transactions on neural networks, 10(5), 988–999, 1999.

Vorobieva, I. A. and Panza, G. F.: Prediction of the occurrence of related strong earthquakes in italy, PAGEOPH, 141(1), 25–41, 1993.

Wyss, M.: Cannot earthquakes be predicted?, Science, 278, 487–488, 1997.