Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Geochemistry 20 (2005) 1106–1119
Applied Geochemistry
www.elsevier.com/locate/apgeochem

Radon in soil gas: How to identify anomalies caused by earthquakes

B. Zmazek [a,*], M. Živčić [a,b], L. Todorovski [a], S. Džeroski [a], J. Vaupotič [a], I. Kobal [a]

[a] *Jožef Stefan Institute, 1000 Ljubljana, Slovenia*
[b] *Office of Seismology, Environmental Agency of the Republic of Slovenia, 1000 Ljubljana, Slovenia*

Received 1 March 2004; accepted 26 January 2005
Editorial handling by Å. Danielsson
Available online 21 April 2005

Abstract

Anomalies have been observed in Rn content in soil gas from 3 boreholes at the Orlica fault in the Krško basin, Slovenia. To distinguish the anomalies caused by environmental parameters (air and soil temperature, barometric pressure, rainfall) from those resulting solely from seismic activity, the following approaches have been used: (i) deviation of Rn concentration from the seasonal average, (ii) correlation between time gradients of Rn concentration and barometric pressure, and (iii) regression trees within a machine learning program. Approach (i) is much less successful in predicting anomalies caused by seismic events than approaches (ii) and (iii) if $\pm 2\sigma$ criterion is used and is equally successful if $\pm 1\sigma$ is used. Approaches (ii) and (iii) did not fail to observe an anomaly preceding an earthquake, but show false seismic anomalies, the number of which is much lower with (iii) than with (ii). Model trees are shown to outperform other approaches. A model has been built which, in the seismically non-active periods when Rn is presumably influenced only by environmental parameters, predicts the concentration with a correlation of 0.8. This correlation is reduced significantly in the seismically active periods.

1. Introduction

Since the advent of the nuclear era in Slovenia in the sixties, ^{222}Rn and other radionuclides have been systematically monitored in ground and surface waters (Kobal et al., 1978, 1990; Kobal, 1979; Kobal and Renier, 1987; Kobal and Fedina, 1987; Vaupotič and Kobal, 2001; Vaupotič, 2002; Popit et al., 2002, 2004). The first Rn analyses with the objective of forecasting earthquakes (Ulomov and Mavashev, 1971; Scholz et al., 1973; Mjachkin et al., 1975; King, 1978, 1986; Ui et al., 1988; Ohno and Wakita, 1996; Pulinets et al., 1997; Toutain and Baubron, 1999; Planinić et al., 2000, 2001; Belyaev, 2001; Virk et al., 2001) were carried out in Slovenia in 1982 (Zmazek et al., 2000a). Radon concentrations were determined weekly in 4 thermal water springs, while Cl^-, SO_4^{2-}, hardness and pH, were determined monthly. In 1998, this study was extended to other thermal water springs (Zmazek et al., 2000b, 2002a,b) and also to soil gas (Zmazek et al., 2000c, 2002c) at selected, seismically relevant sites, and sampling frequency was increased from once a week to once an hour. In this

* Corresponding author. Tel.: +386 1 477 35 80; fax: +386 1 477 38 11.
E-mail address: boris.zmazek@ijs.si (B. Zmazek).

0883-2927/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apgeochem.2005.01.014

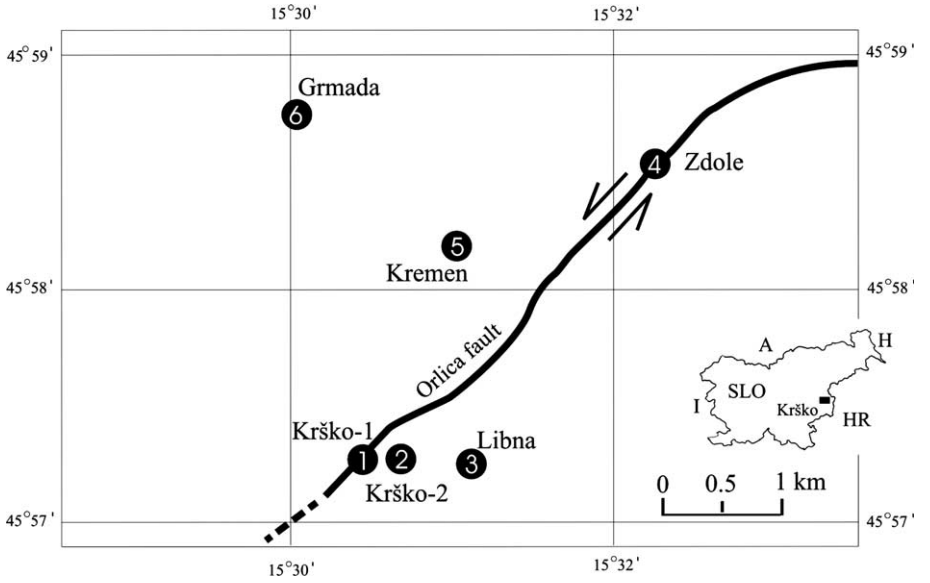

Fig. 1. Map of the Krško basin with locations of Rn monitoring stations at the Orlica fault with strike-slip displacement. The inset shows the position of Krško (SLO – Slovenia, I – Italy, A – Austria, H – Hungary, HR – Croatia).

paper the authors will focus on the Rn concentration in soil gas.

Since April 1999, in 60–90 cm deep boreholes at 6 locations in the Krško basin, Rn concentration in soil gas, barometric pressure and soil temperature have been measured and recorded once an hour, using barasol probes (MC-450, ALGADE, France). Other meteorological data, such as air temperature and rainfall, have been provided by the Office of Meteorology of the Environmental Agency of the Republic of Slovenia, and seismic data by the Office of Seismology of the same agency. Boreholes 1 and 4 are located in the Orlica fault zone, at a distance about 4000 m from each other, while the other boreholes are at distances from 150 to 2500 m on either side of the fault zone (Fig. 1). Air temperature and rainfall were measured at the meteorological station Bizeljsko, approximately 14 km from the boreholes. This paper interprets data only from stations 1 (Krško-1), 5 (Kremen) and 6 (Grmada), because measurements were often interrupted and disturbed at the others. The experimental procedure is reported elsewhere (Zmazek et al., 2002c). As often utilised, for earthquakes Dobrovolsky's equation (Dobrovolsky et al., 1979) was used to calculate R_D i.e., $R_D = 10^{0.43M}$, where M is the earthquake magnitude and R_D the radius of the zone within which precursory phenomena may be manifested (so-called Dobrovolsky's radius in km). Earthquakes for which the distance R_E between the epicentre and the measuring site was equal or less than $2R_D$ have been used in the interpretation.

Following general practice in this field (Yasuoka and Shinogi, 1997; Singh et al., 1999; Virk et al., 2001), an anomaly in Rn concentration is defined as a significant deviation from the mean value and is then related to seismic activity (Zmazek et al., 2002c). It is often impossible to distinguish an anomaly caused solely by a seismic event, from one resulting from meteorological or hydrological parameters. For this reason, the implementation of more advanced statistical methods in data evaluation (Di Bello et al., 1998; Cuomo et al., 2000; Biagi et al., 2001; Belyaev, 2001; Negarestani et al., 2001; Planinić et al., 2003; Steinitz et al., 2003) is important. In this paper, in addition to Rn anomalies expressed as deviations of Rn concentration from the average seasonal value by more than a multiple of standard deviation, and those based on a positive correlation between the time gradients of barometric pressure and of Rn concentration, the authors have applied regression trees for the first time to forecast earthquakes. Data mining and machine learning methods used for that purpose have been already successfully applied to many environmental problems, as reviewed by Džeroski (2002). Here, Rn concentrations are predicted on the basis of environmental data (air and soil temperature, barometric pressure, rainfall) during seismically non-active periods, and then is tested the hypothesis that the prediction is significantly worsened during seismically active periods is tested.

2. Methodology of data analysis

Experimental data have been analysed by searching for Rn anomalies (i) defined as deviations in Rn concen-

tration of more than $\pm\times\sigma$ (multiple $\times$ of standard deviation σ), from the average seasonal value (ii) expressed as those time intervals when time gradients of barometric pressure and Rn concentration have the same sign, and (iii) by using regression trees, which have been confirmed as outperforming other regression methods in this area (Zmazek et al., 2003).

2.1. Deviations in radon concentration

Averages of Rn concentration were calculated for spring, summer, autumn and winter. The periods when Rn concentration deviates by more than $\pm1\sigma$, $\pm1.5\sigma$ and $\pm2\sigma$ from the related seasonal value are considered as Rn anomalies caused possibly by earthquake events and not by meteorological parameters (Yasuoka and Shinogi, 1997; Singh et al., 1999; Virk et al., 2001).

2.2. Dependence of radon concentration on barometric pressure

There is an inverse relationship between Rn exhalation and barometric pressure (Klusman and Webster, 1981). Not so much the absolute value of barometric pressure, but its time gradient plays an important role. A decrease in barometric pressure, with values of other environmental parameters remaining constant, generally causes an increase in Rn exhalation from the ground (Hubbard and Hagberg, 1996; Robinson et al., 1997a,b). Therefore, the periods when the time gradient of barometric pressure, $\mathrm{d}P/\mathrm{d}t$, and the time gradient of Rn concentration, $\mathrm{d}C_{\mathrm{Rn}}/\mathrm{d}t$, in soil gas have the same sign were considered as Rn anomalies attributed to seismic activity and not to environmental parameters.

2.3. Regression methods

Since Rn concentration is a numerical variable, the authors have approached the task of predicting Rn concentration from meteorological data using regression (or function approximation) methods. Regression trees (Breiman et al., 1984) were used, as implemented with the WEKA data mining suite (Witten and Frank, 1999).

Regression trees are a representation for piece-wise constants or piece-wise linear functions. Like classical regression equations, they predict the value of a dependent variable (called class) from the values of a set of independent variables (called attributes). Data presented in the form of a table can be used to learn or automatically construct a regression tree. In that table, each row (example) has the form $(x_1, x_2, \ldots, x_N, y)$, where x_i are values of the N attributes (e.g., air temperature, barometric pressure, etc.) and y is the value of the class (e.g., Rn concentration in soil gas). Unlike classical regression approaches, which find a single equation for a given set of data, regression trees partition the space of examples into axis-parallel rectangles and fit a model to each of these partitions. A regression tree has a test in each inner node that tests the value of a certain attribute and, in each leaf, a model for predicting the class. The model can be a linear equation or just a constant. Trees having linear equations in the leaves are also called model trees (MT).

Given a new example for which the value of the class should be predicted, the tree is interpreted from the root. In each inner node, the prescribed test is performed and, according to the result of the test, the corresponding left or right sub-tree is selected. When the selected node is a leaf then the value of the class for the new example is predicted according to the model in the leaf.

Tree construction proceeds recursively, starting with the entire set of training examples (entire table). At each step, the most discriminating attribute is selected as the root of the sub-tree and the current training set is split into subsets according to the values of the selected attribute. Technically speaking, the most discriminating discrete attribute or continuous attribute test is the one that most reduces the variance of the values of the class variable. For discrete attributes, a branch of the tree is typically created for each possible value of the attribute. For continuous attributes, a threshold is selected and two branches are created, based on that threshold. The attributes that appear in the training set are considered as thresholds. For the subsets of training examples in each branch, the tree construction algorithm is called recursively. Tree construction stops when the variance of the class values of all examples in a node is small enough (or if some other stopping criterion is satisfied). These nodes are called leaves and are labelled with a model (constant or linear equation) for predicting the class value.

An important mechanism used to prevent trees from over-fitting data is tree pruning. Pruning can be employed during tree construction (pre-pruning) or after the tree has been constructed (post-pruning). Typically, a minimum number of examples in branches can be prescribed for pre-pruning and a confidence level in error estimates expressed in leaves for post-pruning.

A number of systems exist for inducing regression trees from examples, such as CART (Breiman et al., 1984) and M5 (Quinlan, 1992). M5 is one of the best known programs for regression tree induction. The authors used the system M5′ (Wang and Witten, 1997), a reimplementation of M5 within the WEKA data mining suite (Witten and Frank, 1999). The parameters of M5′ were set to their default values, unless stated otherwise.

Both types of regression trees mentioned above were used. Ordinary regression trees predict a constant value in each leaf node. Model trees use a linear regression for prediction in each leaf node.

The predictive performance of the regression methods was determined using two different measures, i.e., correlation coefficient and root mean squared error. The correlation coefficient (r) expresses the level of correlation between the measured and predicted values of Rn concentration. The root mean squared error (RMSE) measures the discrepancy between measured and predicted values of Rn concentration. In the authors previous study (Zmazek et al., 2003), using different regression methods, model trees, MT, were confirmed to outperform all the other methods and have been therefore used in this paper.

In order to estimate the performance of predictors on measurements that were not used for training the predictor, a standard 10-fold cross validation method was applied.

3. Results and discussion

Experimental results are shown in Fig. 2. In the following sections, these raw data will be analysed by applying the approaches mentioned above. It will be seen that some earthquakes are preceded and accompanied by Rn anomalies (denoted as CA case: *correct anomaly* related to seismic events), some are not (denoted as NA case: *no anomaly* observed for an earthquake), and, also, that there are anomalies during seismically non-active periods (denoted as FA case: *false anomaly* appearing without a seismic event). Sometimes a single, short anomaly appears, but more often swarms of anomalies are observed over longer periods. The duration period of a swarm, also called *total time of anomalies*, is defined as the time from the beginning of the first to the ending of the last anomaly in the swarm. On the other hand *net time of anomalies* in a swarm is called the sum of duration times of all anomalies in the swarm.

3.1. Approach 1: deviations from seasonal average radon concentration

Daily averages of Rn concentrations, C_{Rn} at the Krško-1, Kremen and Grmada stations are shown in Figs. 3 and 4. Due to operational failures of barasols, much data at Kremen and Grmada is missing. Because the Rn level in soil gas is mostly influenced by soil temperature (Klusman and Jaacks, 1987), average Rn concentrations have been calculated and are drawn by solid lines for 4 seasons, defined from seasonal variation of the soil temperature and not taken from the calendar. They were as follows: 16.04.99–09.07.99, 10.07.99–08.08.99, 09.08.99–05.01.00, 06.01.00–10.02.00, 11.02.00–25.06.00, 26.06.00–07.08.00, 08.08.00–07.01.01, 08.01.01–14.02.01, 15.02.01–02.07.01, 03.07.01–16.08.01, 17.08.01–25.01.02. The $\pm 2\sigma$ region is bounded by dashed lines. Concentrations outside this region are considered as anomalies, not caused by environmental parameters, but possibly by seismic activity. In the figures earthquakes (local magnitude, M_L) are shown for which distances between the epicentre and measuring site (R_E) were equal or less than $2R_D$ with R_D being Dobrovolsky's radius, defined above (Dobrovolsky et al., 1979).

All these anomalies are collected in Table 1 in which, in addition to seismic information, the following data are also given: duration period of anomalies in the swarm, anomaly type (CA, FA or NA), numbers of days an anomaly started before the earthquake occurred, duration of anomalies (net time of anomalies/total time of the swarm), number of anomalies in a swarm and surface area of the swarm (the area between the $\pm 2\sigma$ line and the C_{Rn}–t curve). Several earthquakes occurring within a few days (such as 14.04.00, 16.04.00 and 17.04.00, 24.08.00 and 31.08.00, 29.10.00 and 31.10.00) are considered as one gross seismic event. Some earthquakes are preceded and accompanied by Rn anomalies (CA case), some are not (NA case), and, also, there are anomalies during seismically non-active periods (FA case). Sometimes a single short anomaly appears, like CA on 11.04.–12.04.99 and 18.05.–19.05.99 at Krško-1, but more often swarms of anomalies over a longer time period have been observed, like FA on 14.03.–10.04.01 (composed of 3 short anomalies) and CA on 14.08.–14.09.01 (composed of 8 short anomalies) at Krško-1. The same earthquake did not always cause an anomaly at all stations. Thus, seismic activity on 19.02.01 caused anomalies at Krško-1 but not at Kremen and Grmada. It is surprising that at Krško-1, on one hand, two weak ($M_L = 0.8$ and 0.7) earthquakes on 13.04.99 and 22.05.99 with $R_E/R_D = 1.4$ and 1.6, respectively, were accompanied by Rn anomalies, while on the other hand, the earthquake on 16.04.00 with $M_L = 3.2$ and $R_E/R_D = 0.5$ was not. The FA anomalies appearing during seismically non-active periods are undesirable, such as those during 21.07.–22.07.99, 28.05.–30.05.00, 06.01.00–07.01.01, 14.03.–10.04.01 and 15.11.–16.11.01 at Krško-1, during 06.03.–21.03.01 and 15.11.–16.11.01 at Kremen, and during 23.07.–24.07.01 and 03.10.–04.10.01 at Grmada. NA cases, such as those for earthquakes on 14.04.–16.04.00, 24.08.–31.08.00, 29.10.–31.10.00 and 29.11.00 at Krško-1, when no anomaly was observed even though $R_E/R_D < 1$, are problems from the point of view of forecasting earthquakes.

Table 2 summarizes all CA, FA and NA anomalies and swarms for the 3 stations. At Krško-1 and Grmada, the number of CA swarms outweighs the number of FA swarms, while the opposite is true for Kremen. Sizes of anomalies, expressed by the average surface area per anomaly, are similar at Krško-1 and Kremen, but smaller at Grmada. A positive anomaly (+) is defined as an increase of C_{Rn} with respect to the seasonal average

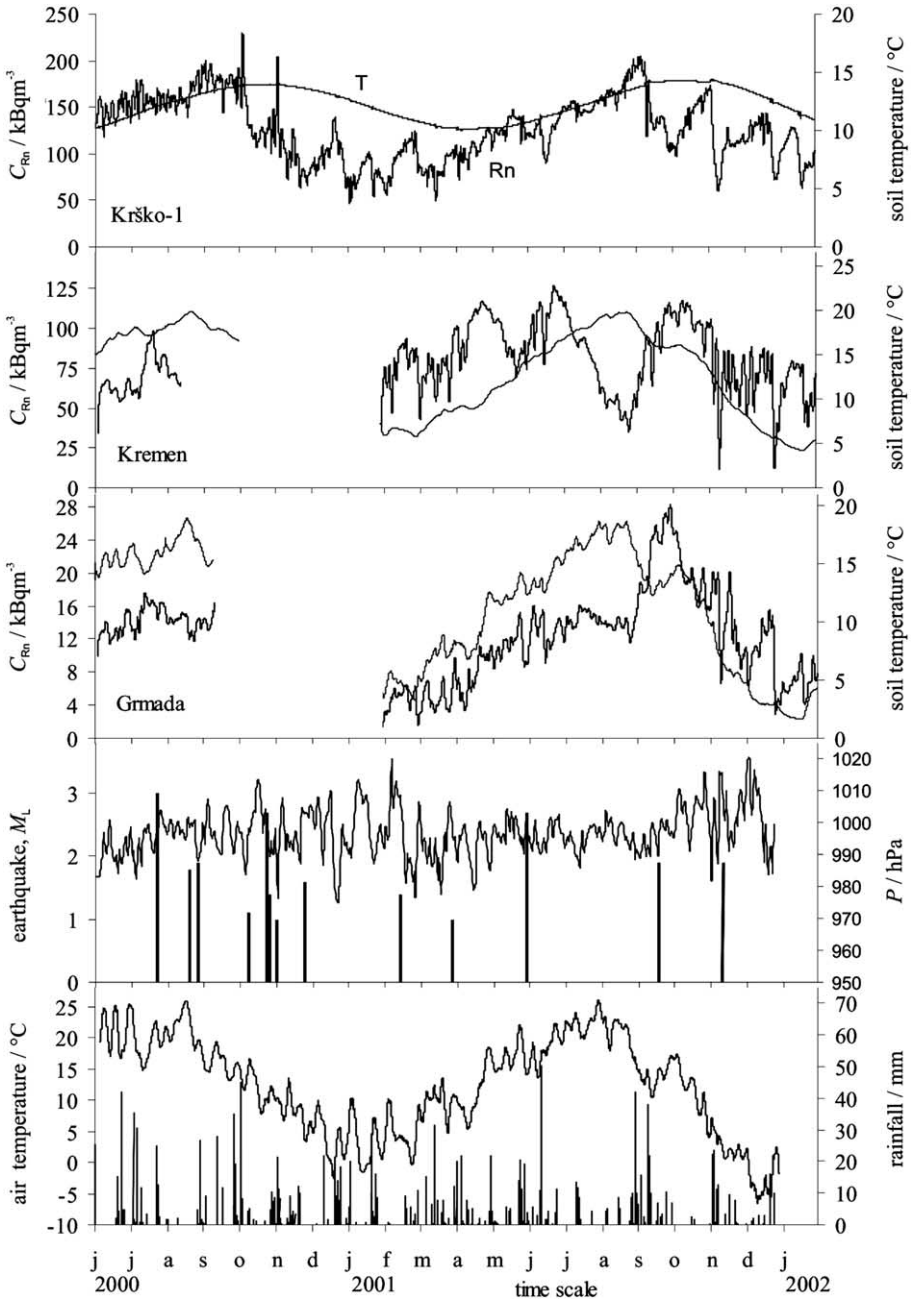

Fig. 2. Time run of daily average Rn concentration in soil gas and of soil temperature recorded with barasol probes in 60 cm deep boreholes at the Krško-1, Kremen and Grmada stations at the Orlica fault in the Krško basin during the period from June 2000 to January 2002. Local earthquakes with R_E/R_D equal to or less than 2 (Dobrovolsky et al., 1979), barometric pressure, air temperature and rainfall at the nearby meteorological station Bizeljsko are also shown.

value, and negative (−) as a decrease. Only at Krško-1, the number of CA+ always outweighs the number of CA−.

Table 3 shows for the Krško-1 station how the number of CA, FA and NA cases changes if different values of threshold are taken in defining an anomaly. When moving from $\pm1\sigma$ to $\pm2\sigma$ the number of FA cases is decreased by 3, but at the same time, the number of CA cases is decreased by 3 and the number of NA cases increased by 3. Thus, $\pm1\sigma$ threshold seems to be optimal, although with a number of undesirable FA cases.

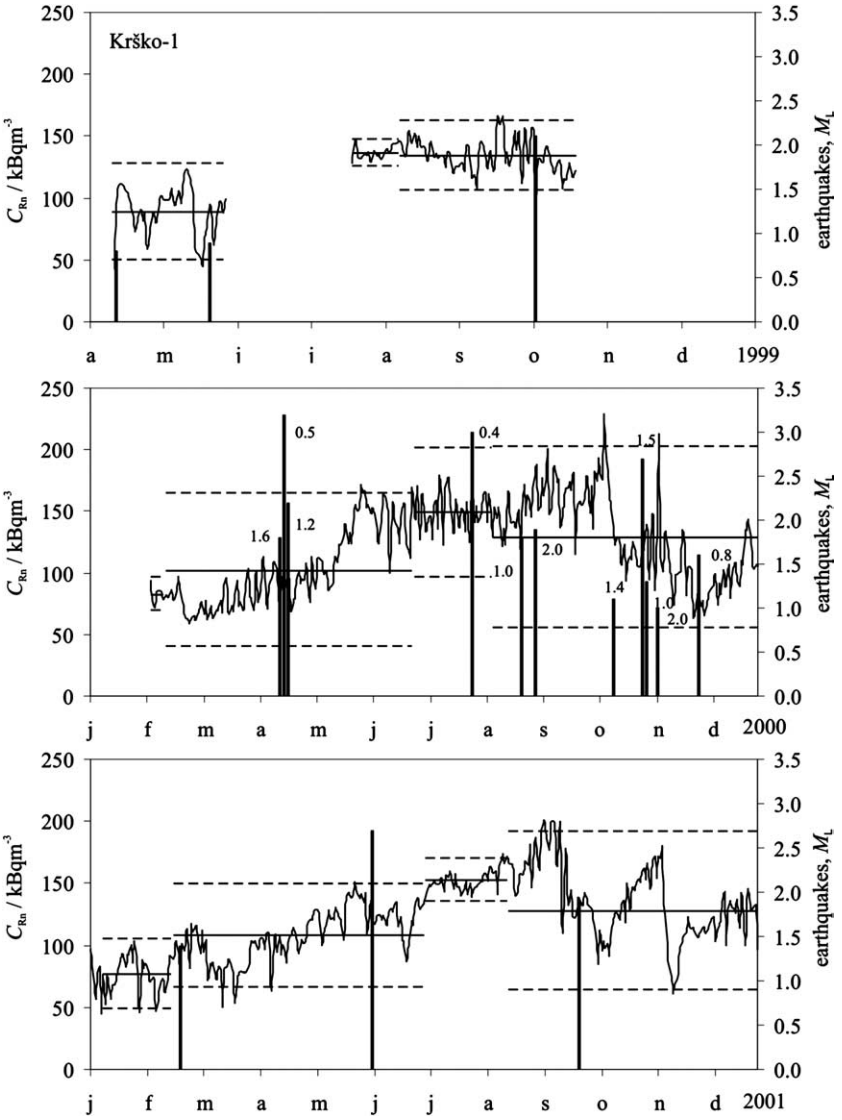

Fig. 3. Time run of daily average Rn concentration in soil gas, C_{Rn} at the Krško-1 station. Numbers attached to the earthquake bars are R_E/R_D values (Dobrovolsky et al., 1979). Full lines represent average radon concentrations for 4 seasons (for definition of seasons see the text), while dashed lines indicate $\pm 2\sigma$ deviation from the seasonal average value. Radon anomalies are C_{Rn} values outside the $\pm 2\sigma$ region.

3.2. Approach 2: the same sign of $\Delta C_{Rn}/\Delta t$ and $\Delta P/\Delta t$

From the measured Rn concentrations in soil gas, C_{Rn} and barometric pressure, P, time gradients $\Delta C_{Rn}/\Delta t$ and $\Delta P/\Delta t$ were calculated and plotted versus time. Tests with values of 1, 2, 6, 12 and 24 h for Δt have been performed. Because of Rn data fluctuations, results for $\Delta t < 6$ h were discarded. $\Delta t = 24$ h appeares to give better results than $\Delta t = 12$ h, and was used in this paper. Although these plots have been drawn for all stations for the whole duration of measurements, Fig. 5 shows as examples only plots for 3 typical periods at Krško-1. Time intervals on the abscissa during which both gradients have the same sign, with an additional criterion that $\Delta P/\Delta t > 2$ hPa d^{-1}, indicate Rn anomalies, possibly related to seismic activity. In order to maintain clarity in showing Rn anomalies, the time scale of these plots cannot be more condensed, and therefore all of them cannot be shown.

All the Rn anomalies observed on the plots described above are collected and classified as CA, FA and NA cases, as done for the previous approach in Table 1.

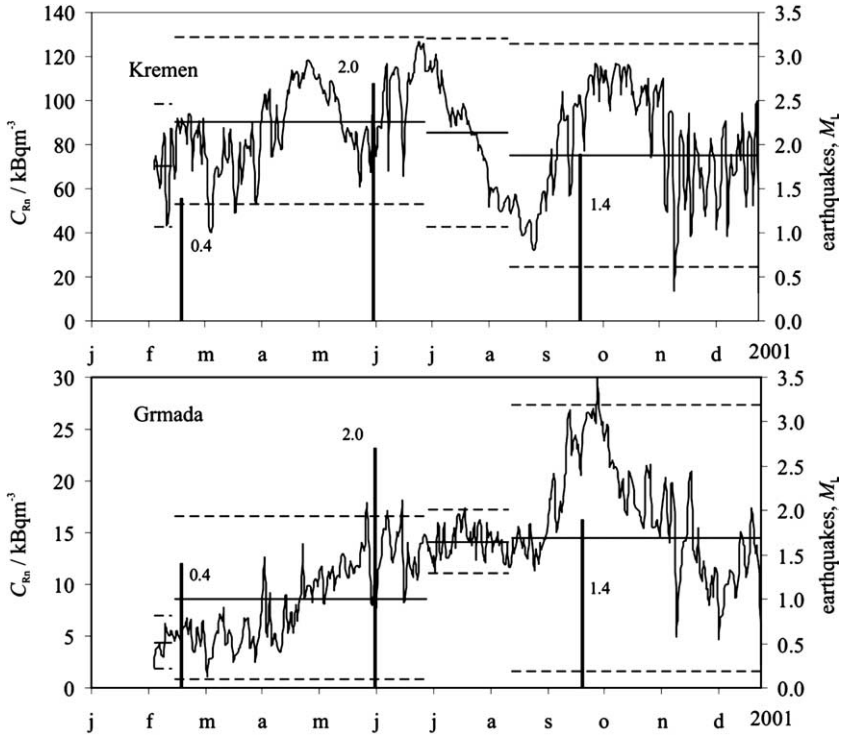

Fig. 4. Time run of daily average Rn concentration in soil gas, C_{Rn} at the (a) Kremen and (b) Grmada stations. Numbers attached to the earthquake bars are R_E/R_D values (Dobrovolsky et al., 1979). Full lines represent average Rn concentrations (for definition of seasons see the text), while dashed lines indicate $\pm 2\sigma$ deviation from the seasonal average value. Radon anomalies are C_{Rn} values outside the $\pm 2\sigma$ region.

Results are given only in a summarized form in Table 4. The number of anomalies in a swarm is here larger than for the case of the $\pm\times\sigma$ definition of anomaly. The number of CA cases is always larger than the number of FA cases, this being especially true for Krško-1. The average size of CA and FA anomalies, expressed by time of duration of the anomaly, is similar. If both gradients are positive the anomaly is denoted as positive (+) and if both gradients are negative, the anomalies are denoted as negative (−). Number of the former cases is by far larger than the number of the latter ones.

Table 5 compares numbers of CA, FA and NA cases at Krško-1 for different values of threshold $\Delta P/\Delta t$. If it is increased from 2 to 10 Pa d^{-1}, the number of FA cases is reduced but, concomitantly, the number of CA cases is undesirably decreased largely and the number of NA cases largely increased. Therefore, $\Delta P/\Delta t > 2$ hPa d^{-1} would be recommended for earthquake forecasting.

3.3. Approach 3: decision trees

To test the hypothesis that the predictability of Rn concentration in periods with seismic activities is worse than in periods without seismic activities, the following procedure was applied. Firstly, the value of the class (i.e., daily Rn concentration) and the values of attributes (i.e., daily average barometric pressure, daily average air temperature, daily average soil temperature, difference between daily soil and daily air temperatures, daily amount of rainfall, and difference in daily barometric pressure) were selected. The difference between the pressures on day $i + 1$ and day i is related to day i. Secondly, this dataset was split into two parts. In the first part (labelled SA and amounting to 23% of the data), data for the periods with seismic activity were included. As a first estimate, periods of 7 days before and after an earthquake were taken. Data for the remaining days were included in the second part, belonging to the seismically non-active periods (labelled non-SA and amounting to 77%). Then, to evaluate the predictability of Rn concentration in the non-SA periods, the performance of MT on the non-SA data were estimated with cross-validation. Furthermore, a model tree is induced on the total non-SA data and its performance measured on the SA data in order to evaluate the practicability of predicting Rn concentration in the SA periods. If the authors hypothesis is true, the first prediction should be better than the second.

Table 1
Earthquakes listed with (1) the date of occurrence, (2) M_L magnitude, and (3) R_E/R_D value (R_E distance of the measuring site from the epicentre; R_D Dobrovolsky's radius (Dobrovolsky et al., 1979)), together with Rn anomalies defined as $\pm 2\sigma$ (σ, standard deviation) deviations of Rn concentration from the seasonal average value and characterised by: (4) duration period of the anomaly, (5) type (CA – *correct anomaly*, FA – *false anomaly*, NA – *no anomaly*), (6) how many days the anomaly appeared before the seismic event, (7) duration time of the anomaly in days (net time of anomalies/total time from the start of the first to the end of the last anomaly in the swarm), (8) number of anomalies in a swarm, (9) the anomaly's area in kBq m^{-3} d, i.e., the area between the $\pm 2\sigma$ line and the C_{Rn}–t curve in Fig. 4

Earthquakes (EQ)			Radon anomalies					
1 Date	2 M_L	3 R_E/R_D	4 Duration period	5 Type	6 Start time before EQ (d)	7 Duration time (d)	8 Number of anomalies	9 Surface area (kBq m^{-3} d)
(a) Krško-1								
13.04.99	0.8	1.4	11.04.–12.04.99	CA	2	1/1	1	7.3
22.05.99	0.7	1.6	18.05.–19.05.99	CA	3	1/1	1	5.7
–	–	–	21.07.–22.07.99	FA	–	1/1	1	0.4
06.10.99	2.1	2.0	19.09.–05.10.99	CA	17	3/17	3	18
14.04.00	1.8	1.6	–	NA	–	–	–	–
16.04.00	3.2	0.5						
17.04.00	2.2	1.2						
–	–	–	28.05.–30.05.00	FA	–	3/3	1	12.2
28.07.00	3.0	0.4	25.06.–26.06.00	CA	33	1/1	1	5.2
24.08.00	1.8	1.0	–	NA	–	–	–	–
31.08.00	1.9	2.0						
13.10.00	1.1	1.4	08.10.–09.10.00	CA	5	2/2	1	29.4
29.10.00	2.7	1.5	–	NA	–	–	–	–
31.10.00	1.3	1,0						
06.11.00	1.0	2.0	05.11.–06.11.00	CA	1	1/1	1	11.8
29.11.00	1.6	0.8	–	NA	–	–	–	–
–	–	–	06.01.–07.01.01	FA	–	1/1	1	10.5
19.02.01	1.4	0.4	28.01.–06.02.01	CA	22	2/10	2	3.8
–	–	–	14.03.–10.04.01	FA	–	3/28	3	5.6
04.06.01	2.7	2.0	25.05.–26.06.01	CA	10	1/1	1	2.5
25.09.01	1.9	1.4	14.08.–14.09.01	CA	33	8/32	5	53.1
–	–	–	15.11.–16.11.01	FA	–	1/1	1	3.2
(b) Kremen								
28.07.00	3.0	0.4	24.07.–25.07.00	CA	3	2/2	1	10.8
19.02.01	1.4	0.4	–	NA	–	–	–	–
–	–	–	06.03.–21.03.01	FA	–	5/16	2	31.2
04.06.01	2.7	2.0	–	NA	–	–	–	–
25.09.01	1.9	1.4	–	NA	–	–	–	–
–	–	–	15.11.–16.11.01	FA	–	1/1	1	9.2
(c) Grmada								
28.07.00	3.0	0.4	10.07.–19.07.00	CA	18	3/10	3	1.0
24.08.00	1.8	1.0	17.08.–03.09.00	CA	7	4/18	4	2.0
31.08.00	1.9	2.0						
19.02.01	1.4	0.4	–	NA	–	–	–	–
04.06.01	2.7	2.0	31.05.–19.06.01	CA	4	3/20	3	3.5
–	–	–	23.07.–24.07.01	FA	–	1/1	1	0.2
25.09.01	1.9	1.4	–	NA	–	–	–	–
–	–	–	03.10.–04.10.01	FA	–	2/2	1	2.9

The period of 7 days was estimated after inspecting correlation changes between Rn concentration and barometric pressure (Zmazek et al., 2002c), and was here confirmed by an analysis in which the length of the SA periods was varied from 1 to 7 days. Table 6 summarises the correlation coefficients and RMSEs obtained with the MT method for different lengths of the SA period (assuming that Rn changes appear 1–7 days before an

Table 2
Summary of characteristics of anomalies from Table 1

	Krško-1			Kremen			Grmada		
	CA	FA	NA	CA	FA	NA	CA	FA	NA
Total number of anomalies[a]	16/9	7/5	4	1/1	3/2	2	10/3	2/2	2
Total duration of anomalies[b] (d)	18/44	11/61	–	2/2	6/17	–	13/48	3/3	–
Average duration time (d)	1.13	1.57	–	2	2	–	1.3	1.5	–
Total surface area of anomalies (kBq $^{-3}$ d)	136.8	31.9	–	10.8	40.4	–	6.5	3.1	–
Average surface area per anomaly (kBq m^{-3} d)	8.6	6.6	–	10.8	13.5	–	0.7	1.6	–
Number of '+' anomalies	11	2	–	1	0	–	4	2	–
Number of '–' anomalies	5	5	–	0	3	–	6	0	–

[a] Number of anomalies/number of swarms.
[b] Net time of duration of anomalies/total time of duration of swarms.

Table 3
Krško-1: $\pm\sigma$ sigma approach: different threshold for anomaly

Anomaly	$\pm1\sigma$	$\pm1.5\sigma$	$\pm2\sigma$
CA	12	10	9
FA	8	8	5
NA	1	3	4

earthquake). The largest drop in the correlation coefficient was observed between 6 and 7 days before an earthquake.

The data for 7 days in Table 6 (bold) clearly confirm the hypothesis: the correlation in the SA periods is much lower than in the non-SA periods, for all 3 stations. The drop in the correlation coefficient ranged from 0.17 to

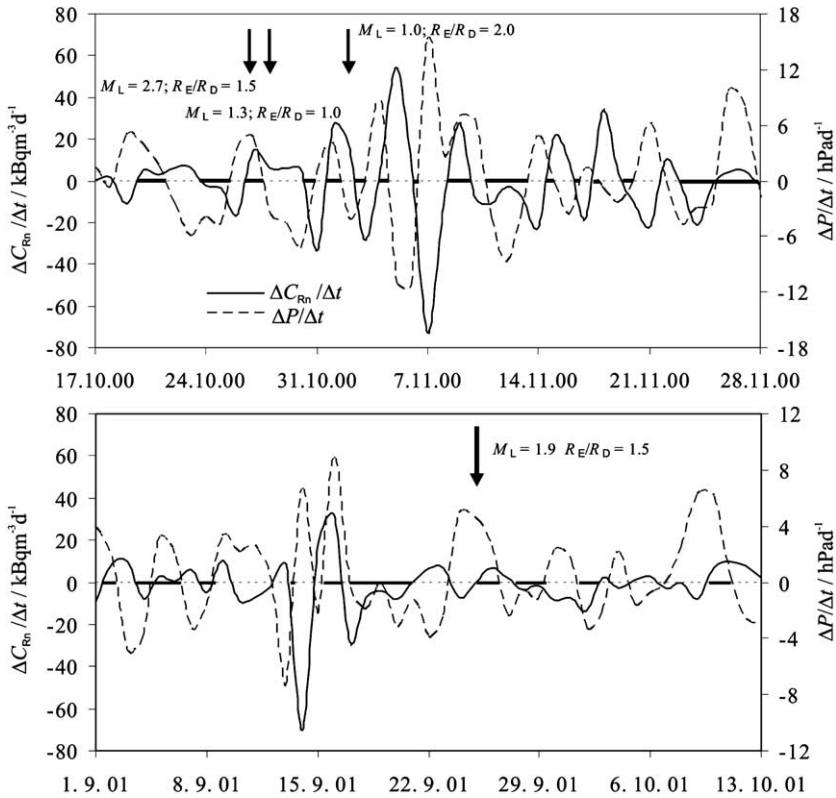

Fig. 5. Time run of the gradients of daily average Rn concentration in soil gas, C_{Rn} and of barometric pressure, P, i.e., $\Delta C_{Rn}/\Delta t$ and $\Delta P/\Delta t$ (for $\Delta P/\Delta t > 2$ hPa d^{-1}), with $\Delta t = 24$ h, for selected periods at the Krško-1 station. Times of earthquake occurrence are indicated by arrows; numbers attached are M_L and R_E/R_D values (Dobrovolsky et al., 1979). The black parts of the abscissa represent Rn anomalies, i.e., intervals when $\Delta C_{Rn}/\Delta t$ and $\Delta P/\Delta t$ have simultaneously the same sign.

Table 4
Summary of characteristics of anomalies defined as time intervals in which $\Delta C_{Rn}/\Delta t$ and $\Delta P/\Delta t$ ($\Delta P/\Delta t > 2$ hPa d^{-1}) have simultaneously the same sign

	Krško-1			Kremen			Grmada		
	CA	FA	NA	CA	FA	NA	CA	FA	NA
Total number of anomalies[a]	73/11	50/6	1	49/7	34/5	1	53/7	48/5	0
Total duration of anomalies[b] (d)	90/337	64/239	–	59/228	39/182	–	77/233	57/273	–
Average duration time (d)	1.23	1.28	–	1.20	1.15	–	1.45	1.19	–
Number of '+' anomalies	44	33	–	33	18	–	40	34	–
Number of '−' anomalies	29	22	–	16	16	–	13	14	–

[a] Number of anomalies/number of swarms.
[b] Net time of duration of anomalies/total time of duration of swarms.

Table 5
Krško-1: gradient approach: different threshold of $\Delta P/\Delta t$ (hPa d^{-1}) in defining an anomaly

Anomaly	$\Delta P/\Delta t > 2$	$\Delta P/\Delta t > 5$	$\Delta P/\Delta t > 10$
CA	11	8	3
FA	6	4	1
NA	1	4	9

0.73. The RMSE is higher in the SA periods than in the non-SA periods. The RMSE increase ranged from 13% to 72%. This confirmation of the hypothesis allows the prediction of seismic activity in the following manner. A model tree is built that predicts the concentration of Rn in soil gas on the basis of data measured during the non-SA periods. The discrepancy between the measured values of Rn concentration and the values pre-

Table 6
Predictability of Rn concentration obtained by decision trees in terms of correlation coefficients (r) and of root mean squared error (RMSE), and compared for different delay times (from 1 to 7 days) between the start time of a Rn anomaly and occurrence of the related seismic event

Delay time (d)	Non-SA periods		SA periods		Performance change %	
	r	RMSE	r	RMSE	r	RMSE
Krško-1						
7	**0.83**	**18719**	**0.69**	**23536**	**−17.3**	**25.7**
6	0.84	18135	0.73	22551	−13.5	24.4
5	0.82	18913	0.76	21659	−6.9	14.5
4	0.82	19052	0.76	21291	−6.7	11.8
3	0.80	19792	0.79	19975	−1.6	0.9
2	0.80	19972	0.78	19534	−2.5	−2.2
1	0.80	19920	0.78	19197	−2.4	−3.6
Kremen						
7	**0.81**	**13243**	**0.54**	**14910**	**−33.7**	**12.6**
6	0.81	13342	0.53	13223	−34.4	−0.9
5	0.81	13288	0.66	11208	−18.4	−15.7
4	0.82	12720	0.79	8457	−4.1	−33.5
3	0.81	13183	0.71	10295	−11.9	−21.9
2	0.81	12990	0.74	12272	−8.8	−5.5
1	0.82	12667	0.73	12225	−10.8	−3.5
Grmada						
7	**0.80**	**3076**	**0.22**	**5299**	**−72.9**	**72.3**
6	0.79	3185	0.14	5404	−82.6	69.7
5	0.80	3142	0.18	5321	−77.4	69.4
4	0.79	3211	0.36	4982	−54.1	55.1
3	0.77	3415	0.37	4836	−51.3	41.6
2	0.77	3397	0.25	4884	−67.5	43.8
1	0.76	3451	0.35	4593	−53.9	33.1

Performance change is defined as the difference in the values of either r or RMSE in the SA and non-SA periods, expressed as a percentage of the value in the non-SA period.

dicted by MT is then followed. If the discrepancy is low, no seismic activity is anticipated; if it starts to increase, this may be attributed to oncoming seismic events.

In order to facilitate the visualisation of Rn anomalies found by MT, the expression $(C_{Rn})_m/(C_{Rn})_p - 1$ was plotted versus the time elapsed, as shown in Fig. 6 for selected periods. Here, $(C_{Rn})_m$ is the measured Rn concentration and $(C_{Rn})_p$ is the Rn concentration predicted with decision trees. In the plots, in addition to the $(C_{Rn})_m/(C_{Rn})_p - 1 = 0$ line, the ±0.2 region is indicated by dashed lines. Values of the expression $(C_{Rn})_m/(C_{Rn})_p - 1$ falling beyond the dashed lines were considered as anomalies. All the anomalies found over the total period of observation were collected and classified as CA, FA and NA cases. The results obtained are given in a summarized form in Table 7. The number of CA cases largely outweighs the number of FA cases at Krško-1 but the opposite is true at Kremen and Grmada. The

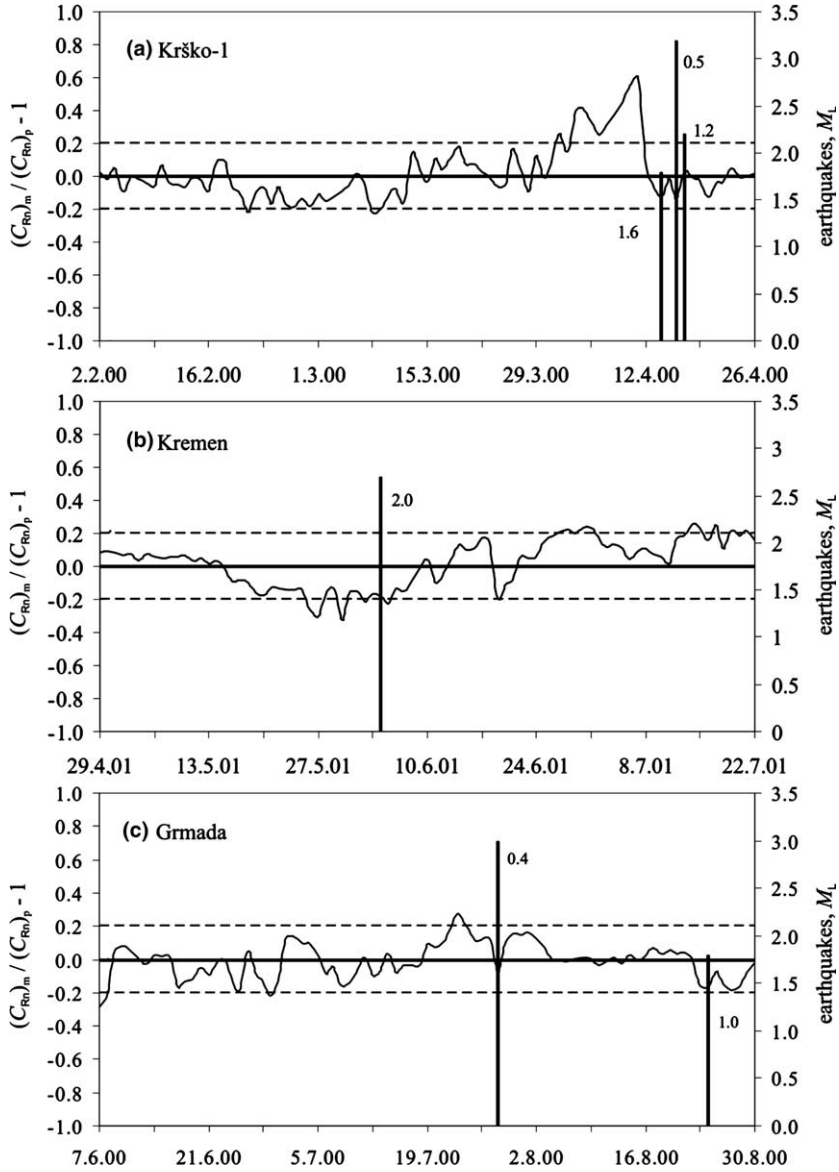

Fig. 6. Time run of the of expression $(C_{Rn})_m/(C_{Rn})_p - 1$ (C_{Rn}, Rn concentration in soil gas, m – measured, p – predicted with decision trees) for selected periods at the (a) Krško-1, (b) Kremen and (c) Grmada stations. The solid line is drawn at $(C_{Rn})_m/(C_{Rn})_p - 1 = 0$, and dashed lines at ±0.2. Numbers attached to the earthquake bars are R_E/R_D values (Dobrovolsky et al., 1979). Radon anomalies are the $(C_{Rn})_m/(C_{Rn})_p - 1$ values outside the ±0.2 region.

Table 7
Summary of characteristics of Rn anomalies defined with the value of the expression $(C_{Rn})_m/(C_{Rn})_p - 1$ outside the ±0.2 region; $(C_{Rn})_m$ and $(C_{Rn})_p$ is Rn concentration, measured and predicted by model trees, respectively

	Krško-1			Kremen			Grmada		
	CA	FA	NA	CA	FA	NA	CA	FA	NA
Total number of anomalies[a]	36/12	19/6	0	9/3	27/4	0	14/4	37/3	0
Total duration of anomalies[b] (d)	102/179	42/135	–	20/40	48/202	–	50/102	93/179	–
Average duration time (d)	2.83	2.21	–	2.22	1.78	–	3.57	2.51	–
Total surface area of anomalies (kBq m^{-3} d)	14.58	3.12	–	1.01	3.63	–	11.70	20.20	–
Average surface area per anomaly (kBq m^{-3} d)	0.41	0.16	–	0.11	0.13	–	0.84	0.55	–
Number of '+' anomalies	22	11	–	3	14	–	7	17	–
Number of '–' anomalies	14	8	–	6	13	–	7	20	–

[a] Number of anomalies/number of swarms.
[b] Net time of duration of anomalies/total time of duration of swarms.

average surface area per anomaly (defined as the area between the line at $(C_{Rn})_m/(C_{Rn})_p - 1 = \pm 0.2$ and the curve $(C_{Rn})_m/(C_{Rn})_p - 1$ versus time) is more than 2-fold greater for CA than for FA at Krško-1, but at Kremen and Grmada values for CA and FA are practically the same. A positive anomaly (+) is one with $(C_{Rn})_m/(C_{Rn})_p - 1 > 0$, and negative (–), with $(C_{Rn})_m/(C_{Rn})_p - 1 < 0$. For CA, the number of '+' cases is higher than the number of '–' cases at Krško-1, but the opposite is true at Kremen and at Grmada.

Table 8 reveals that the choice of threshold value for $(C_{Rn})_m/(C_{Rn})_p - 1$ in the range from ±0.15 to ±0.25 does not much effect the number of CA, FA and NA cases. Therefore, a value below ±0.20 seems to be appropriate for earthquake forecasting.

Table 8
Krško-1: model trees: different threshold of expression $(C_{Rn})_m/(C_{Rn})_p - 1$ in defining an anomaly

Anomaly	$(C_{Rn})_m/(C_{Rn})_p - 1$, ±0.15	$(C_{Rn})_m/(C_{Rn})_p - 1$, ±0.20	$(C_{Rn})_m/(C_{Rn})_p - 1$, ±0.25
CA	12	12	11
FA	7	6	4
NA	0	0	1

3.4. Comparison of approaches

Using the ±×σ approach (Table 2), for a total of 13 seismic events at Krško-1, the method failed for 4 events if a ±2σ thresholds is used and failed only for 1 event if ±1σ is used (Table 3). The gradient approach (Table 4) and model trees (Table 7) have never failed for earthquakes with $R_E/R_D < 1$. With the gradient approach, the number of anomalies in a swarm is higher than that with model trees and the number of CA cases does not always outweigh the number of FA cases, thus resulting in misleading earthquake forecasts. Increasing the

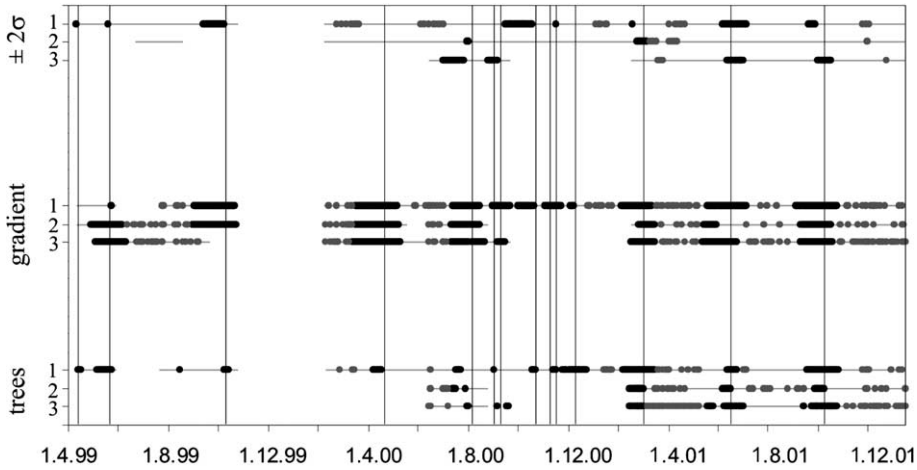

Fig. 7. Appearance of CA (black point) and FA (gray point) anomalies as observed by applying ±2σ, C_{Rn} and P gradients, and model trees to the data from 1 – Krško-1, 2 – Kremen and 3 – Grmada stations. Vertical lines represent earthquakes with $R_E/R_D < 2$. Horizontal dotted lines represent the periods when the instrument was in operation and no anomalies observed.

threshold for $\Delta P/\Delta t$ above 2 further worsens earthquake forecasting. The model trees approach appears not to depend much on the choice for the threshold of $(C_{Rn})_m/(C_{Rn})_p - 1$ (Table 8) and is therefore used with less hesitation. The comparison of the 3 approaches is summarized in Fig. 7 in which, for each method, the appearance of CA and FA anomalies at the 3 stations are shown.

4. Conclusions

The analysis has shown that Rn anomalies (i) based on deviations of Rn concentration from the seasonal average concentration have been less successful in predicting earthquakes as those (ii) based on the time gradient of Rn concentration and barometric pressure having simultaneously the same sign, and (iii) obtained with model trees. Approaches (i) and (ii) strongly depend on the values of $\pm\times\sigma$ and $\Delta P/\Delta t$ thresholds, respectively, while the dependence of approach (iii) on the threshold of $(C_{Rn})_m/(C_{Rn})_p - 1$ is very weak. Approaches (ii) and (iii) have shown a number of false anomalies (FA). The number cannot be reduced in approach (ii) which is based on the assumption that the effect of pressure fluctuation on Rn exhalation is not disturbed by other environmental parameters, an assumption which is often not the case (Hubbard and Hagberg, 1996; Robinson et al., 1997a,b). This assumption is further argued by different Rn transport at compression and dilatation zones (Ui, 1986). The situation is more promising with model trees for which it is expected that, by including additional environmental parameters such as humidity of soil (Ioannides et al., 1996; Fujiyoshi et al., 2002), direction and velocity of wind (Riley et al., 1996), and snow coverage (Fujiyoshi et al., 2002), and by extending the time with further measurements, may improve the machine learning and hence reduce the number of FA cases. Therefore, a great deal of further effort will be devoted to this approach.

Acknowledgements

The study was funded by the Slovenian Ministry of Education, Science and Sport. The authors thank Prof. H. Ui from the Toyama University, Japan, for fruitful discussions and his constructive suggestions.

References

Belyaev, A.A., 2001. Specific features of radon earthquake precursors. Geochem. Int. 12, 1245–1250.

Biagi, P.F., Ermini, A., Kingsley, S.P., Khatkevich, Y.M., Gordeev, E.I., 2001. Difficulties with interpreting changes in groundwater gas content as earthquake precursors in Kamchatka, Russia. J. Seismol. 5, 487–497.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont.

Cuomo, V., Di Bello, G., Lapenna, V., Piscitelli, S., Telesca, L., Macchiato, M., Serio, C., 2000. Robust statistical methods to discriminate extreme events in geoelectrical precursory signals: implications with earthquake prediction. Nat. Hazard 21, 247–261.

Di Bello, G., Ragosta, M., Heinicke, J., Koch, U., Lapenna, V., Piscitelli, S., Macchiato, M., Martinelli, G., 1998. Time dynamics of background noise in geoelectrical and geochemical signals: an application in a seismic area of Southern Italy. Il Nuovo Cimento 6, 609–629.

Dobrovolsky, I.P., Zubkov, S.I., Miachkin, V.I., 1979. Estimation of the size of earthquake preparation zones. Pure Appl. Geophys. 117, 1025–1044.

Džeroski, S., 2002. Applications of KDD methods in environmental sciences. In: Kloesgen, W., Zytkow, J. (Eds.), Handbook of Data Mining and Knowledge Discovery. Oxford University Press.

Fujiyoshi, R., Morimoto, H., Sawamura, S., 2002. Investigation of soil radon variation during the winter months in Sapporo, Japan. Chemosphere 47, 369–373.

Hubbard, L.M., Hagberg, N., 1996. Time-variation of the soil gas radon concentration under and near Swedish house. Environ. Int. 22, S477–S482.

Ioannides, K.G., Papachristodoulou, C., Karamanis, D.T., Stamoulis, K.C., Mertzimekis, T.J., 1996. Measurements of ^{222}Rn migration in soil. J. Radioanal. Nucl. Chem. 208, 541–547.

King, C.Y., 1978. Radon emanation on San Andreas fault. Nature 271, 516–519.

King, C.Y., 1986. Gas geochemistry applied to earthquake prediction: an overview. J. Geophys. Res. 91, 12269–12281.

Klusman, R.W., Webster, J.D., 1981. Preliminary analysis of meteorological and seasonal influences on crustal gas emission relevant to earthquake prediction. Bull. Seismol. Soc. Am. 71, 211–222.

Klusman, R.W., Jaacks, J.A., 1987. Environmental influences upon mercury, radon and helium concentrations in soil gases at a site near Denver, Colorado. J. Gephys. Res. 27, 259–280.

Kobal, I., Kristan, J., Škofljanec, M., Jerančič, S., Ančik, M., 1978. Radioactivity of spring and surface waters in the region of the uranium ore deposit at Žirovski vrh. J. Radioanal. Chem. 44, 307–315.

Kobal, I., 1979. Radioactivity of thermal and mineral springs in Slovenia. Health Phys. 37, 239–242.

Kobal, I., Renier, A., 1987. Radioactivity of the atomic spa at Podčetrtek, Slovenia, Yugoslavia. Health Phys. 53, 307–310.

Kobal, I., Fedina, Š, 1987. Radiation doses at the Radenci health resort. Radiat. Prot. Dosim. 20, 257–259.

Kobal, I., Vaupotič, J., Mitić, D., Kristan, J., Ančik, M., Jerančič, S., Škofljanec, M., 1990. Natural radioactivity of fresh waters in Slovenia, Yugoslavia. Environ. Int. 16, 141–154.

Mjachkin, V.I., Brace, W.E., Sobolev, G.A., Dieterich, J.H., 1975. Two models for earthquake forerunners. Pure Appl. Geophys. 113, 169–181.

Negarestani, A., Setayeshi, S., Ghannadi-Maragheh, M., Akashe, B., 2001. Layered neural networks based analysis of radon concentration and environmental parameters in earthquake prediction. J. Environ. Radioact. 62, 225–233.

Ohno, M., Wakita, H., 1996. Coseismic radon changes of the 1995 Hyogo-ken Nanbu earthquake. J. Phys. Earth 44, 391–395.

Planinić, J., Radolić, V., Èulo, D., 2000. Searching for an earthquake precursors: temporal variations of radon in soil and water. Fizika B (Zagreb) 9, 75–82.

Planinić, J., Radolić, V., Lazanin, Ż, 2001. Temporal variations of radon in soil related to earthquakes. Appl. Radiat. Isot. 55, 267–272.

Planinić, J., Vuković, B., Radolić, V., Faj, Z., Stanić D., 2003. Deterministic chaos in radon time variations. In: Proceedings 5th Symposium of the Croatian Radiation Protection Association, HDZZ-CRPA, Zagreb, pp. 349–354.

Popit, A., Urbanc, J., Vaupotič, J., Kobal, I., 2002. Radioactivity survey of waters in the Slovenian Karst region. RMZ – Mater. Geoenviron. 49, 487–496.

Popit, A., Vaupotič, J., Kukar, N., 2004. Systematic radium survey in spring waters of Slovenia. J. Environ. Radioact. 76, 337–347.

Pulinets, S.A., Aleseev, V.A., Legenka, A.D., Khegai, V.V., 1997. Radon and metallic aerosolos emanation before strong earthquakes and their role in atmosphere and ionosphere modification. Adv. Space Res. 20, 2173–2176.

Quinlan, J.R., 1992. Learning with continuous classes. In: Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence. World Scientific, Singapore, pp. 343–348.

Riley, W.J., Gadgil, A.J., Nazaroff, W.W., 1996. Wind-induced ground-surface pressures around a single-family house. J. Wind Eng. Ind. Aerodyn. 61, 153–167.

Robinson, A.L., Sextro, R.G., Fisk, W.J., 1997a. Soil-gas entry into an experimental basement driven by atmospheric pressure fluctuations—measurements, spectral analysis, and model comparison. Atmos. Environ. 31, 1477–1485.

Robinson, A.L., Sextro, R.G., Riley, W.J., 1997b. Soil-gas entry in to houses driven by atmospheric pressure fluctuations-the influence of soil properties. Atmos. Environ. 31, 1487–1495.

Scholz, C.H., Sykes, L.R., Agrawal, Y.P., 1973. Earthquake prediction: a physical basis. Science 181, 803–810.

Singh, M., Kumar, M., Jain, R.K., Chatrath, R.P., 1999. Radon in ground water related to seismic events. Radiat. Meas. 30, 465–469.

Steinitz, G., Begin, Z.B., Gazit-Yaari, N., 2003. Statistically significant relation between radon flux and weak earthquakes in Dead Sea rift valley. Geology 31, 505–508.

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

Ui, H., 1986. Fault shear zone as a strain meter, Tokyo University. Crust Chemistry Laboratory Rep. 4, 85–96.

Ui, H., Moriuchi, H., Takemura, Y., Tsuchida, H., Fujii, I., Nakamura, M., 1988. Anomalously high radon discharge from the Atotsugawa fault prior to the western Nagano Prefecture earthquake (M 6.8) of September 14, 1984. Tectonophysics 152, 147–152.

Ulomov, V.I., Mavashev, B.Z., 1971. Forerunners of the Tashkent earthquake. Izv. Akad. Nauk Uzb. SSR, 188–200.

Vaupotič, J., 2002. Radon exposure at drinking water supply plants in Slovenia. Health Phys. 83, 901–906.

Vaupotič, J., Kobal, I., 2001. Radon exposure in Slovenian spas. Radiat Prot. Dosim. 97, 265–270.

Virk, H.S., Walia, V., Kumar, N., 2001. Helium/radon precursory anomalies of Chamoli earthquake, Garhwal Himalaya, India. J. Geodyn. 31, 201–210.

Wang, Y., Witten, I.H., 1997. Induction of model trees for predicting continuous lasses. In: Proceedings of the Poster Papers of the European Conference on Machine Learning, University of Economics, Faculty of Informatics and Statistics, Prague.

Witten, I.H., Frank, E., 1999. Data Mining: Practical Machine Learning Tools and Techniqueswith Java Implementations. Morgan Kaufmann, San Francisco.

Yasuoka, Y., Shinogi, M., 1997. Anomaly in atmospheric radon concentration: a possible precursor of the 1995 Kobe, Japan, earthquake. Health Phys. 72, 759–761.

Zmazek, B., Vaupotič, J., Živčić, M., Premru, U., Kobal, I., 2000a. Radon monitoring for earthquake prediction in Slovenia. Fizika B (Zagreb) 9, 111–118.

Zmazek, B., Vaupotič, J., Živčić, M., Martinelli, G., Italiano, F., Kobal, I., 2000b. Radon, temperature, electrical conductivity and $^{3}He/^{4}He$ measutements in three thermal springs in Slovenia. In: Book of Abstracts: New Aspects of Radiation Measurements, Dosimetry and Spectrometry, 2nd Dresden Symposium on Radiation Protection, September 10–14.

Zmazek, B., Vaupotič, J., Bidovec, M., Poljak, M., Živčić, M., Pineau, J.F., Kobal, I., 2000c. Radon monitoring in soil gas tectonic faults in the Krško basin. In: Book of Abstracts: New Aspects of Radiation Measurements, Dosimetry and Spectrometry, 2nd Dresden Symposium on Radiation Protection, September 10–14.

Zmazek, B., Italiano, F., Živčić, M., Vaupotič, J., Kobal, I., Martinelli, G., 2002a. Geochemical monitoring of thermal waters in Slovenia: relationships to seismic activity. Appl. Radiat. Isot. 57, 919–930.

Zmazek, B., Vaupotič, J., Kobal, I., 2002b. Radon, temperature and electric conductivity in Slovenian thermal waters as potential earthquake precursors. In Book of Abstracts: 1st Workshop Natural Radionuclides in Hydrology and Hydrogeology, Centre Universitaire de Luxembourg, September 4–7.

Zmazek, B., Živčić, M., Vaupotič, J., Bidovec, M., Poljak, M., Kobal, I., 2002c. Soil radon monitoring in the Krško basin, Slovenia. Appl. Radiat. Isot. 56, 649–657.

Zmazek, B., Todorovski, L., Džeroski, S., Vaupotič, J., Kobal, I., 2003. Application of decision trees to the analysis of soil radon data for earthquake prediction. Appl. Radiat. Isot. 58, 697–706.