Evaluation of impact of solar activity on occurrences of earthquakes and selection of effective variables

Nishii, Ryuei
Institute of Mathematics for Industry, Kyushu University | International Center for Space Weather Science and Education, Kyushu University

Qin, Pan
Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology

Kikuyama, Ryosuke
Graduate School of Mathematics, Kyushu University

Evaluation of impact of solar activity on occurrences of earthquakes and selection of effective variables

Ryuei Nishii[1,2*], Pan Qin[3], and Ryosuke Kikuyama[4]

[1]Institute of Mathematics for Industry, Kyushu University, Motooka 744, Nishiku, Fukuoka 819-0395, Japan
[2]International Center for Space Weather Science and Education, Kyushu University, Motooka 744, Nishiku, Fukuoka 819-0395, Japan
[3]Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China
[4]Graduate School of Mathematics, Kyushu University, Motooka 744, Nishiku, Fukuoka 819-0395, Japan
[*]corresponding author nishii@imi.kyushu-u.ac.jp

ABSTRACT

Possible triggers of earthquakes such as tidal stress, rainfall, the building of artificial water reservoirs, mining, and extraction of fossil fuels have been discussed in the literature. Furthermore, solar activity has been also speculated to be a trigger of earthquakes. In this article, we tackle this problem by support vector regression in machine learning approach, wherein we used physical measurements of solar activity and related variables for predicting frequencies of earthquakes globally. We conclude that solar activity is a trigger of earthquakes with magnitudes from 3.0 to 5.9, whereas it has little effect on triggering larger earthquakes. In addition, we show that the Earth's electric field is the most effective variable related to earthquakes.

Introduction

Forecasting and prediction are two branches of seismology. Forecasting concerns the probabilistic assessment of earthquake hazards, including the frequency and magnitude of damaging earthquakes in a given area over the course of years[1]. Prediction mainly focuses on the specification of the time, location, and magnitude of a future earthquake within stated limits[2]. Prediction is often based on various precursors, such as animal behavior[3], radon emissions[4], and variations in the travel time ratio of S (secondary) to P (primary) waves[5]. Seismo-electromagnetics might also be a basis for short-term earthquake prediction[6].

Earthquake prediction and forecasting are far from mature. However, knowledge can be accumulated from seismicity-related data by using statistical methods. The development of observation technologies, including GPS and InSAR[7], has given rise to new theories on and approaches to studying earthquakes. Therefore, a thorough understanding of the data and consequent improvements to data quality are crucial[8–10]. Here, several factors associated with the distributions of earthquake size and frequency have been discovered[11–13]. Another important role of statistics is to construct models, such as the point process model[14,15] and hidden Markov model[16], that can be used to estimate the probability of earthquakes occurring within a specific spatio-temporal window. Furthermore, statistical data facilitate a discussion of possible triggers of earthquakes, for example, tidal stress[17,18], rainfall-triggered earthquakes[19], and human activity such as the building of artificial water reservoirs[20], mining[21], and extraction of fossil fuels[22].

"Does solar activity cause earthquakes?" is a question that dates back to the early Twentieth century[23]. At that time, very few measurements on solar activity (SA) like sun spot numbers (SSN) were available. Recently, a lot of novel space data related to SA has become available, e.g., real-time proton density and temperature[24]. Nowadays, the various SA data have been available, but still one is negative[25] to the question and the other is partly positive[26,27].

We tackled the problem by using support vector regression (SVR), which is a regression model dealing with non-linear phenomena. Daily earthquake frequencies are regressed by nine physical measurements related to SA as well as earthquake frequencies observed earlier. From the analysis of seven years worth of data, we concluded that SA is a trigger of earthquakes with magnitudes 3 to 5.9 ($3 \leq M \leq 5.9$), but it has little effect on triggering earthquakes of magnitudes $M \geq 6$. We also find that, according to the model selection, the Earth's electric field is the most effective variable in statistical relation to earthquakes.

Results

Table 1 lists nine physical measurements related to the solar activity and the magnetosphere. We call these measurements SA variables for simplicity and give their abbreviations in the table. The daily data of these variables are available from the

Table 1. External variables and abbreviations

Abb.	SA variables	units
IMF	Magnitude of IMF (Interplanetary Magnetic Field)	nT
Temp	Proton Temperature	K
D	Proton Density	n/cc
V	Solar Wind Velocity	km/s
P	Dynamic Pressure	nPa
Ey	Earth's Electric Field (east-west component)	mV/m
SSN	Sun Spot Number	n
DST	Disturbance Storm Time index	nT
PCI	Polar Cap Index	

OMNIWeb database supported by NASA[28]. The website provides information about magnetic fields, plasma, and energetic particles relevant to the heliosphere. The value IMF at date t is denoted by IMF(t) and other eight values are similarly defined. We analyzed the data from Jan. 1, 2006 ($t = 1$) to Dec. 31, 2012 ($t = 2527$).

The daily earthquake data were downloaded from the Advanced National Seismic System (ANSS) database of the northern California earthquake data center[29], which provides accurate and timely data. The database also provides information related to earthquake events, including damage to buildings and structures. We used the earthquake data observed in the same period.

Let EQ3(t) be earthquake frequencies with $3 \leq M < 4$ (EQ3) for $t = 1,...,2527$. Similarly EQ4(t), ..., EQ7(t) are defined, and EQ8(t) for $M \geq 8$. SVR was employed for the target variable $\log(\text{EQ3}(t+\nu)+1)$ with explanatory variables: EQ3(s), ... , EQ8(s) for $s = t, t-1, ..., t-d_y+1$ and

$$\log(\text{IMF}(u)),\ \log(\text{Temp}(u)),\ \log(\text{D}(u)),\ \log(\text{V}(u)),\ \log(\text{P}(u)),\ \text{Ey}(u),\ \log(\text{SSN}(u)+1),\ \text{DST}(u),\ \text{PCI}(u)$$

for $u = t, t-1, ..., t-d_x+1$, where d_y and d_x denote the maximum time delays of earthquakes and SA respectively. Here, $\nu (= 1,2,...,5)$ denotes ν-day-ahead prediction. By the numerical study, the transformation of non-negative variables improved prediction accuracy. In this article, SVR model is estimated by the first 2000 samples up to June 23, 2011 (training data), and remaining 557 samples are used for validation (test data). We calculated test correlation between actual and predicted target data. Note that the case $d_x = 0$ means that SA variables are not used for prediction.

Prediction of earthquake frequencies with/without SA variables

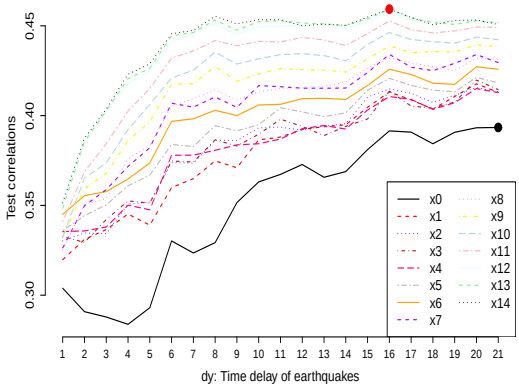
Figure 1 depicts the SA effect on the frequencies of earthquakes by magnitude. Figure 1 (a) shows the test correlations between $\log(\text{EQ3}(t)+1)$ and values predicted by the SVR, where the horizontal and vertical axes correspond to time delay d_y and correlation respectively. Test correlations were calculated for the test data with $t = 2001,..,2557$ by using the optimal model based on the training data with $t = 1,...,2000$.

The solid curve of Figure 1 (a) shows test correlations based on the SVR model without SA variables ($d_x = 0$). The red dotted curve denoted as x1 was derived by SVR model with SA variables observed at a date t ($d_x = 1$). The red dotted curve is above the solid line, which means an effect of SA was detected. The largest correlation was attained by the curve x14 using SA variables from one day to two weeks ago ($d_x = 14$). The maximum correlation without SA was about 0.36, whereas it was over 0.45 with the SA variables. Figure 1 (b)-(f) give similar plots for EQ4 to EQ8. It is seen that SA variables have much effect on EQ4 and a little effect on EQ5. However, SA effect cannot be detected for earthquakes with $M \geq 6$.

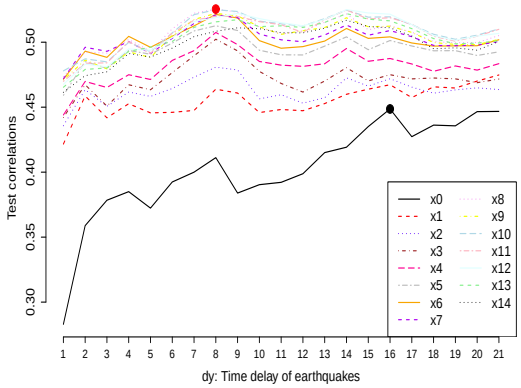
Figure 2 illustrates SA effect on two-day-ahead prediction of frequencies of EQ3 to EQ6. Figure 2 (b) shows the biggest improvement by using the SA variables for EQ4; the correlation exceeded 0.51. By comparison, Figure 2 (c) and (d) for EQ5 and EQ6 do not show any improvement by additional use of SA variables. This implies that SA has little effect on triggering large earthquakes more than magnitude 6. Figure 3 shows the effect of SA for three- to five-day-ahead predictions of frequencies of EQ3 and EQ4. An increase in correlation due to SA is clearly seen in these cases. The increase for EQ3 is especially significant.

Evaluation of impact of SA on seismic activity

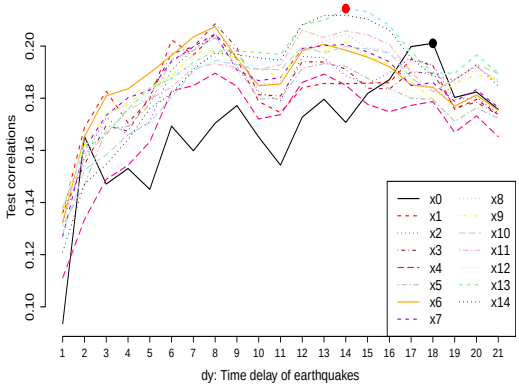
Table 2 tabulates the maximum test correlations with or without SA measurements corresponding to Figures 1-3. The ν, d_y, d_x, and increment columns respectively show the prediction steps, time delays of EQ and SA, and increments had by adding SA measurements as explanatory variables. The sixth and ninth columns denoted by "corr." correspond to the black and red points respectively in the subfigures of Figures 1 to 3.



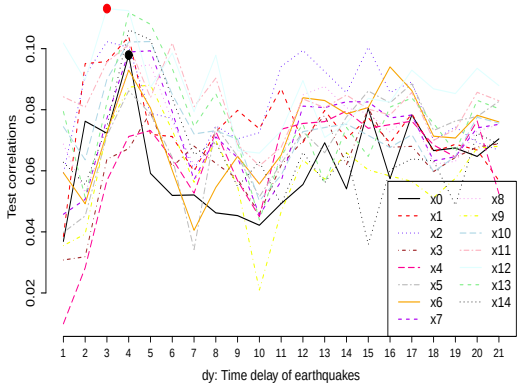
(a) Prediction of $\log(\mathrm{EQ3}(t+1)+1)$



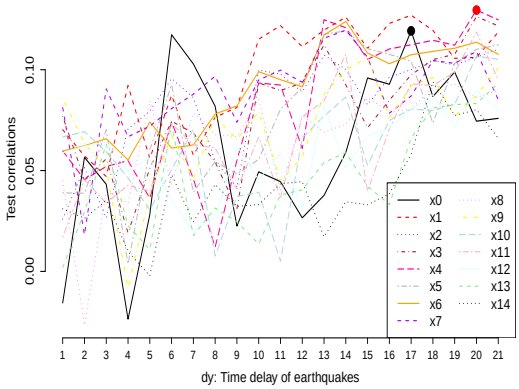
(b) Prediction of $\log(\mathrm{EQ4}(t+1)+1)$



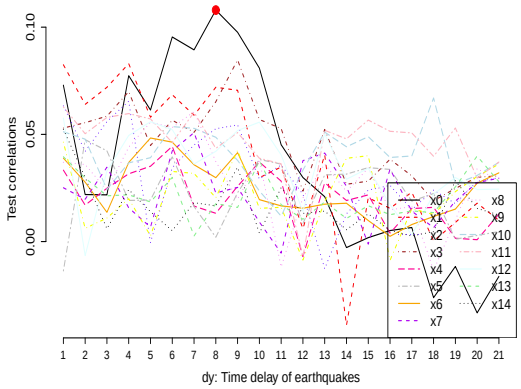
(c) Prediction of $\log(\mathrm{EQ5}(t+1)+1)$



(d) Prediction of $\log(\mathrm{EQ6}(t+1)+1)$

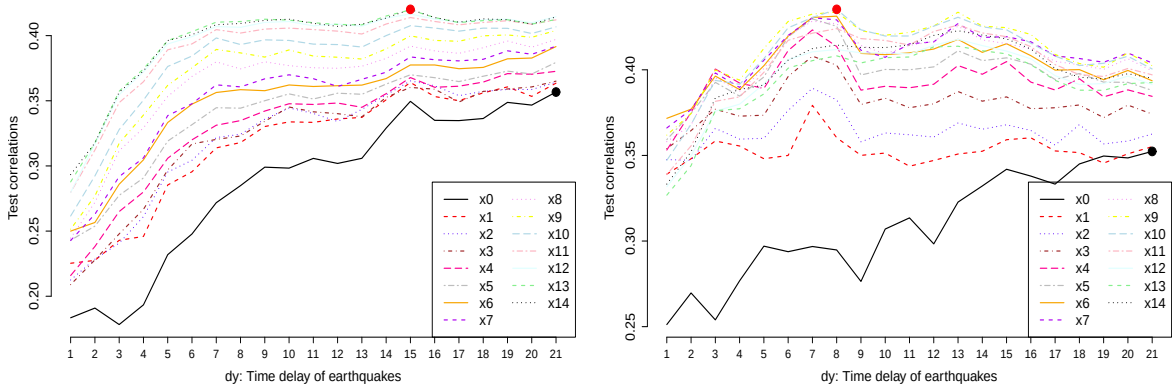


(e) Prediction of $\log(\mathrm{EQ7}(t+1)+1)$



(f) Prediction of $\log(\mathrm{EQ8}(t+1)+1)$

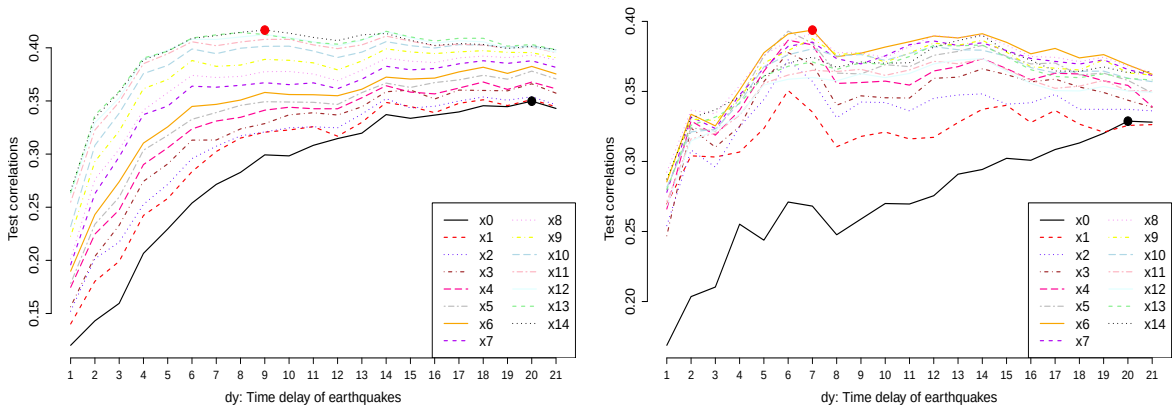
Figure 1. Test correlations for one-day-ahead prediction



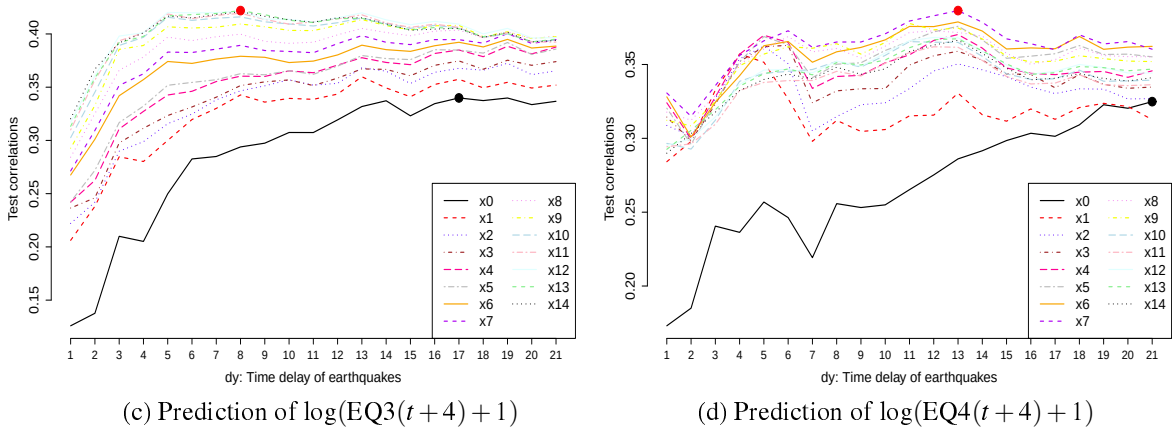
(a) Prediction of $\log(\mathrm{EQ3}(t+2)+1)$ (b) Prediction of $\log(\mathrm{EQ4}(t+2)+1)$

(c) Prediction of $\log(\mathrm{EQ5}(t+2)+1)$ (d) Prediction of $\log(\mathrm{EQ6}(t+2)+1)$

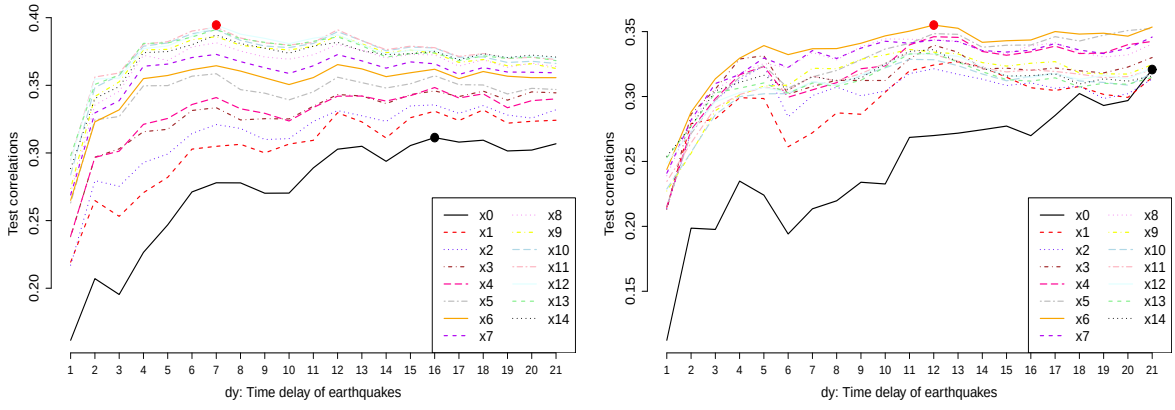
Figure 2. Test correlations for two-day-ahead prediction



(a) Prediction of $\log(\mathrm{EQ3}(t+3)+1)$ (b) Prediction of $\log(\mathrm{EQ4}(t+3)+1)$



(c) Prediction of $\log(\mathrm{EQ3}(t+4)+1)$ (d) Prediction of $\log(\mathrm{EQ4}(t+4)+1)$



(e) Prediction of $\log(\mathrm{EQ3}(t+5)+1)$ (f) Prediction of $\log(\mathrm{EQ4}(t+5)+1)$

Figure 3. Test correlations for three- to five-day-ahead prediction

Table 2. Maximum test correlations with/without nine SA measurements

ν-day-ahead	Magnitudes	Figures	d_y	d_x	corr.	d_y	d_x	corr.	increment
1	3	Fig. 1 (a)	21	0	0.3934	16	14	0.4593	0.0659
	4	Fig. 1 (b)	16	0	0.4486	8	8	0.5255	0.0768*
	5	Fig. 1 (c)	18	0	0.2011	14	13	0.2144	0.0133
	6	Fig. 1 (d)	4	0	0.0979	3	12	0.1131	0.0152
	7	Fig. 1 (e)	17	0	0.1192	20	4	0.1296	0.0103
	8	Fig. 1 (f)	8	0	0.1080	8	0	0.1080	0
2	3	Fig. 2 (a)	21	0	0.3567	15	14	0.4202	0.0635
	4	Fig. 2 (b)	21	0	0.3523	8	10	0.4353	0.0830**
	5	Fig. 2 (c)	17	0	0.1724	17	0	0.1724	0
	6	Fig. 2 (d)	3	0	0.0868	3	1	0.1209	0.0341
3	3	Fig. 3 (a)	20	0	0.3498	9	14	0.4166	0.0668
	4	Fig. 3 (b)	20	0	0.3288	7	6	0.3938	0.0650
	5	———	16	0	0.1460	18	10	0.1510	0.0049
4	3	Fig. 3 (c)	17	0	0.3399	8	13	0.4219	0.0820**
	4	Fig. 3 (d)	21	0	0.3248	13	7	0.3865	0.0616
	5	———	15	0	0.1410	17	10	0.1523	0.0113
5	3	Fig. 3 (e)	16	0	0.3114	7	12	0.3945	0.0832**
	4	Fig. 3 (f)	21	0	0.3208	12	6	0.3550	0.0342
	5	———	14	0	0.1528	16	8	0.1781	0.0253

Note: Numerals with "*" and "**" mean that the increments are more than 0.07 and 0.08 respectively.

For example, the second row of Table 2 gives maximum correlations (0.3934) with the time delay $d_y = 21$ without SA measurements ($d_x = 0$) for one-day-ahead prediction $\log(\mathrm{EQ3}(t+1)+1)$. It also shows that the correlation increased to 0.4593 with SA measurements with time delays $(d_y, d_x) = (16, 14)$, and the increment was 0.0659, see Fig. 1 (a). The increment of 0 in the last column means that additional SA variables failed to increase the test correlation.

From this table, we can conclude the following.

- The increments in the last column are mostly positive for ν-day-ahead prediction for $\nu = 1, 2, ..., 5$. The increase in test correlation as a result of using SA measurements for EQ3 and EQ4 is about 0.07.
- The optimal time delay d_y is about 20 days for EQ3, EQ4 and EQ5. This implies that aftershocks occur for three weeks.
- The contribution from SA to EQ4 is outstanding for one- or two-day-ahead prediction. For more than two-day-ahead predictions, however, the contribution from SA to EQ3 is highest.
- The contribution from SA to EQ5 can be detected, but it is not significant.
- The contribution from SA to earthquakes of $M \geq 6$ is insignificant.

Variable selection of SA measurements

The results show that SA variables impact EQ3 to EQ5, but they do not give impact equally. We tested two variable selection methods as follows.

The first approach focuses on SA variables useful for one-day-ahead prediction. If an effective variable for earthquake prediction is eliminated from the regression function, the test correlation would become worse. Given this consideration, we can find effective variables for EQ3 prediction in the following steps.

- Obtain the optimal test correlations by eliminating one of the nine physical measurements for one-day-ahead prediction with time delays $d_y = 1, ..., 21$ and $d_x = 0, 1, ..., 14$.
- Find an effective order of SA measurements according to the small order of the test correlations.

The second approach is based on the variable-reduction method. Table 2 shows the best models based on the whole nine SA variables. This approach is to reduce variables sequentially as long as the optimal test correlation is improved.

- Obtain the best test correlation by omitting one of SA variables for $d_y = 1, ..., 21$ and $d_x = 0, 1, ..., 14$.
- Find the maximum value among the nine best correlations derived in the previous step. If the maximum is less than the maximum based on the whole variables, we regard the whole variable set as the best and stop the procedure. Otherwise, we regard the variable redundant and omit from the explanatory variables.

This procedure is conducted repeatedly until the test correlation become worse with respect to the maximum value derived by the previous step. This was conducted for EQ3 to EQ5 and $\nu = 1, 2, ..., 5$.

Table 3 unifies the optimal models derived by the first and second approaches. The third column shows selected SA variables expressed in the variable order of Table 1. For example, a variable set 00000 1111 represents a set of Ey, SSN, DST and PCI (the last four variables in Table 1). Increments denote values in comparison with the maximum values without SA variables. Actually, the second approach succeeded to find the best model except two cases with ($\nu = 3$, EQ5) and ($\nu = 5$, EQ5). Apparently, the selection of SA variables was effective. In addition, these findings are valid for one- to five-day-ahead predictions.

Table 3. Maximum test correlations with selected SA variables

ν-day-ahead	Magnitudes	SA variables	d_y	d_x	corr.	increment
1	3	00000 1111	20	12	0.4777	0.0843**
	4	11011 1101	8	8	0.5298	0.0812**
	5	01000 1111	6	6	0.2404	0.0394
2	3	10010 1111	15	14	0.4342	0.0775*
	4	01010 1111	7	9	0.4438	0.0914***
	5	00010 1100	6	10	0.1993	0.0269
3	3	10010 1110	9	14	0.4319	0.0821**
	4	00010 1111	14	14	0.4134	0.0847**
	5	00000 1110	16	6	0.1650	0.0190
4	3	11010 1110	8	13	0.4357	0.0958***
	4	00010 1100	13	13	0.4059	0.0811**
	5	01000 1111	17	7	0.1752	0.0341
5	3	10010 1110	7	12	0.4143	0.1030***
	4	01000 1011	21	6	0.3833	0.0625
	5	00000 1110	4	6	0.1884	0.0356

Note 1: The column "SA variables" denotes the selected variables in the order of Table 1.
Note 2: Numerals with "*", "**" and "***" mean that the increments are more than 0.07, 0.08, and 0.09+ respectively. .

Let us count the frequencies of each SA variable in the 12-best models shown in Table 3. The sixth variable Ey was selected 12 times; i.e., all the best models need Ey. This means that Ey is very effective for predicting earthquakes. This is understandable because it is the east-west component of the electric field induced on the ground by geomagnetic activity. SSN, Velocity, DST and PCI selected by 10, 9, 9 and 8, respectively, are also effective variables, whereas D (density) was not selected.

Discussion

We investigated the relation between solar activity and earthquakes. We predicted daily global earthquake frequencies at different magnitude levels for the whole earth by using SVR. Here, we used earthquake frequencies up to 21 days earlier and nine SA variables up to 14 days earlier. The parameters of SVR were tuned by minimizing cross validation errors on the basis of 2000 training time-series data collected from Jan.1, 2006 to June 23, 2011. The model accuracy was evaluated by its correlation with 557 test data collected from June 24, 2011 to Dec. 31, 2012. We obtained the following results.

1. SA undoubtedly affected EQ3 and EQ4. SA affected EQ5 as well, but not so strongly. The time delays were about three weeks for earthquakes and two weeks for SA.

2. Our approach did not detect any SA effect on EQ6 and higher.

3. The increase in test correlation as a result of considering SA measurements was significant for EQ3 and EQ4. The increase was about 0.08 in each case. It was about 0.03 in the case of EQ5.

4. Variable selection of SA measurements was examined. The selected SA variables slightly increased the test correlations obtained by all the SA variables about 0.005 to 0.01.

5. The Ey, SSN, Velocity, DST and PCI physical measurements are effective SA variables for predicting EQ3, EQ4, and EQ5.

We still have several problems to tackle. Our purpose here was to determine whether SA affects earthquakes, and this was positively shown by analyzing seven years worth of daily data. The next problem is to examine our approach for periods longer than one solar cycle. The second problem regards development of a less computationally intensive variable selection method. This is very important because running the variable-reduction method on five computers took two months for it to make a selection.

In this study, we used SVR as given by equation (7), which treats all the explanatory variables equally. As we have seen, however, some variables are more important than others. Another problem is to evaluate the effectiveness of the explanatory variables in SVR modeling. This problem will be discussed elsewhere.

Data Sets and Methods

This section describes the time series data on earthquakes, solar activity, and magnetosphere observed from 01/01/2006 to 12/31/2012.

Daily frequencies of earthquakes

Table 4 illustrates the frequencies of earthquakes whose Richter magnitude scales are 3 or larger ($M \geq 3$). The bottom row of the table gives the sum of each column. It can be seen that the sum of the EQ3 frequencies is less than that of the EQ4 frequencies. This may come from difficulties with the observing system of EQ3. The data contain measurements of the earthquake $M = 7.2$ (04/05/2010) that occurred in Estado de Baja California, Mexico, and the Touhoku earthquake $M = 9.0$ (03/11/2011) in north-east Japan. Large earthquakes always caused aftershocks, whereas small earthquakes may be foreshocks of larger one. Therefore, frequencies of the earthquakes are also taken as explanatory variables as well as SA variables for predicting future occurrences of earthquakes.

Table 4. Frequencies of earthquakes by magnitude and their notation (n = 2,557)

Date	t	EQ3	EQ4	EQ5	EQ6	EQ7	EQ8
Jan. 1, 2006	1	28	24	4	0	0	0
Jan. 2, 2006	2	22	53	5	0	2	0
Jan. 3, 2006	3	30	33	8	0	0	0
Jan. 4, 2006	4	25	35	2	1	0	0
Jan. 5, 2006	5	26	24	4	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Dec. 27, 2012	2552	6	11	4	0	0	0
Dec. 28, 2012	2553	3	14	4	0	0	0
Dec. 29, 2012	2554	2	17	6	1	0	0
Dec. 30, 2012	2555	7	5	5	0	0	0
Dec. 31, 2012	2556	6	11	3	0	0	0
—-	sum	60,474	76,770	12,023	1,052	89	11

Daily solar activity and magnetosphere data

The daily data of SA variables are available from the OMNIWeb database supported by NASA[28], where the website provides information about magnetic fields, plasma, and energetic particles relevant to the heliosphere. Table 5 shows the first and last five samples of the nine variables related to the solar activity and the magnetosphere.

Table 5. Measurements of solar activity and magnetosphere

Date	t	IMF	Temp	D	V	P	Ey	SSN	DST	PCI
Jan. 1, 2006	1	10.6	69,616	9.3	465	4.38	-2.7	25	3	0.5
Jan. 2, 2006	2	5.7	93,551	5.8	442	2.28	-1.11	24	0	0.3
Jan. 3, 2006	3	4.2	42,310	2.6	423	0.93	-0.32	19	-3	0.5
Jan. 4, 2006	4	4.6	43,776	2.9	380	0.81	-0.4	17	-2	0.3
Jan. 5, 2006	5	5.1	36,086	5.5	339	1.27	0.08	15	2	0.7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Dec. 27, 2012	2552	3.3	18,453	8	292	1.36	-0.07	39	-1	0.2
Dec. 28, 2012	2553	4.6	20,183	13.5	288	2.24	0.16	38	6	0.4
Dec. 29, 2012	2554	5.2	39,178	6.4	296	1.1	0.06	40	-4	0.5
Dec. 30, 2012	2555	8.1	91,979	7.8	372	1.99	-0.08	40	1	0.5
Dec. 31, 2012	2556	3.3	59,632	2	410	0.65	-0.01	64	-5	0.2
—-	mean	4.7	86,224	6.1	417.2	1.71	-0.03	22.7	-7.5	1.5

A short review on support vector regression

Each column of the earthquake frequencies in Table 4 was modeled by using machine learning on the SA variables in Table 1. This section starts by reviewing the general SVR. Then, it introduces an SVR for time-series data with a time delay. Consider a set of training data $\{(y_i,x_i) \mid i = 1,2,...,n\}$ for a general regression problem, where y_i and x_i respectively denote a target variable and a vector of p explanatory variables. Regression problems aim to predict a target variable y_i by using a weighted sum of base functions $\phi_1(x_i),...,\phi_d(x_i)$ of explanatory vectors x_i:

$$y_i \quad = \quad w_1(\phi_1(x_i)+\cdots+w_d\phi_d(x_i)+b+\text{error}_i \equiv w^T\phi(x_i)+b+\text{error}_i. \tag{1}$$

Usually, the unknown vector w is estimated by minimizing the squared sum of errors $Q(w) \equiv \sum_{i=1}^{n}\{y_i - w^T\phi(x_i) - b\}^2$, or its penalized formula given by $CQ(w)+\frac{1}{2}w^Tw$, where $C > 0$ is called a trade-off parameter between the data fitting and smoothness of the regression function. Support vector regression takes an ε-insensitive error function defined by

$$E_\varepsilon(u) \quad = \quad (|u|-\varepsilon)I(|u|>\varepsilon)$$

for the error evaluation, where $\varepsilon > 0$ is a constant and $I(\cdot)$ denotes the indicator function. The target function of SVR to be minimized is given by the following regularized error function:

$$C\sum_{i=1}^{n}E_\varepsilon\left(y_i - w^T\phi(x_i) - b\right)+\frac{1}{2}w^Tw \tag{2}$$

where $C > 0$ is also called a trade-off parameter. The first term of (2) takes zero if and only if $w^T\phi(x_i)+b-\varepsilon \le y_i \le w^T\phi(x_i)+b+\varepsilon$ (ε-tube of y_i) for all i. We relax the condition by introducing slack variables $\xi_i \ge 0$ and $\xi_i^* \ge 0$ satisfying

$$w^T\phi(x_i)+b-\varepsilon-\xi_i^* \le y_i \le w^T\phi(x_i)+b+\varepsilon+\xi_i \quad \text{for} \quad i=1,...,n. \tag{3}$$

Then, formula (2) can be rewritten as

$$C\sum_{i=1}^{n}(\xi_i+\xi_i^*)+\frac{1}{2}w^Tw \tag{4}$$

which must be minimized for nonnegative ξ_i and ξ_i^* subject to the constraints (3). The minimization problem can be solved by using the Lagrangian,

$$\begin{aligned} L(w,b,\{(\xi_i,\ \xi_i^*)\}) \quad &= \quad C\sum_{i=1}^{n}(\xi_i+\xi_i^*)+\frac{1}{2}w^Tw-\sum_{i=1}^{n}(\mu_i\xi_i+\mu_i^*\xi_i^*) \\ &- \quad \sum_{i=1}^{n}a_i\left(w^T\phi(x_i)+b+\varepsilon+\xi_i-y_i\right)-\sum_{i=1}^{n}a_i^*\left(y_i-w^T\phi(x_i)-b+\varepsilon+\xi_i^*\right) \end{aligned} \tag{5}$$

where a_i, a_i^*, μ_i, μ_i^* are the Lagrange multipliers corresponding to the non-negativity restrictions. We can then obtain relations which hold for the optimum values by setting all partial derivatives of L with respect to w, b, ξ_i, ξ_i^* to zero.

Finally, formula (5) can be rewritten as

$$\tilde{L}(\{(a_i,\, a_i^*)\}) \quad = \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i-a_i^*)(a_j-a_j^*)K(x_i,x_j)-\varepsilon\sum_{i=1}^{n}(a_i+a_i^*)+\sum_{i=1}^{n}(a_i-a_i^*)y_i \tag{6}$$

where $K(x_i,x_j)$ denotes the inner product $\phi(x_i)^T\phi(x_j)$ of the transformed input vectors x_i, x_j. The dual problem is required to maximize the formula (6) subject to the constraints $0 \le a_i \le C$ and $0 \le a_i^* \le C$. Then, the regression function $f(x)$ of an explanatory vector x is given by

$$f(x) \quad = \quad \sum_{i=1}^{n}(a_i-a_i^*)\,K(x,x_i)+b. \tag{7}$$

The explanatory vector x_i with $a_i - a_i^* \neq 0$ is called a support vector because it contributes to regression. Note that there is no need to know the actual formula of $\phi(u)$ in the optimization process. We only need to know the exact formula of the kernel $K(u,v)=\phi(u)^T\phi(v)$ (the kernel trick). For a detailed discussion on SVR, see Chap. 7[30].

Tuning of SVR parameters by cross-validation

In the analysis of the EQ-SA time series, we use a **Gaussian kernel** given by $K(u,v)=\exp\left(-\gamma\|u-v\|^2\right)$ for $\gamma>0$. The trade-off parameter $C>0$ used in formula (2) is tuned by **ten-fold cross validation**. The procedure is as follows.

1. Consider a positive constant C and randomly split the training data into ten subsets $D_1,...,D_{10}$ having approximately the same sample size.
2. For $s=1,...,10$,
   (a) solve the dual problem (6) by using the training data $D_1\cup\cdots\cup D_{s-1}\cup D_{s+1}\cup\cdots\cup D_{10}$, and
   (b) calculate the sum of squared errors (SSE_s) between the observed target variables of D_s and predicted values of the form (7).
3. Calculate the sum of ten SSEs (test error).

This procedure is repeated many times to obtain the average test errors with the given $C>0$. Thus, the trade-off parameter C is optimized such that it minimizes the average test errors.

SVR for earthquake frequencies

As an illustration of our analysis, we will consider SVR for predicting EQ3(t): the frequency of EQ3 at date t. The target variable at the date $t+1$ is predicted by earthquake frequencies as well as the SA measurements at the current date t and earlier. More precisely, let $x(t)$ and $y(t)$ be vectors of explanatory variables at date t defined by

$$y(t)=(\mathrm{EQ3}(t),\mathrm{EQ4}(t),...,\mathrm{EQ8}(t))^T : 6\times 1 \;\text{ and }\; x(t)=(\mathrm{IMF}(t),\mathrm{Temp}(t),...,\mathrm{PCI}(t))^T : 9\times 1 \tag{8}$$

for $t=1,2,...,2000$. Now, the vector $y(t)$ denotes earthquake frequencies by magnitudes and $x(t)$ denotes SA variables. Let $y(t-d_y+1:t)\equiv\left(y(t-d_y+1)^T, y(t-d_y+2)^T,...,y(t)^T\right)^T$ be a $6d_y$-dimensional vector of the earthquake frequencies from the date $t-d_y+1$ to t. The effect of aftershocks with a time delay of d_y days may be evaluated as

$$\mathrm{EQ3}(t+1) \quad = \quad f_{\mathrm{EQ}}\left(\, y(t-d_y+1:t)\,\right)+\mathrm{error}. \tag{9}$$

Note that EQ3(t), ... , EQ3($t-d_y+1$) are explanatory variables of EQ3($t+1$) (auto-regressive model) because the first element of each vector $y(s)$ given by (8) was defined by EQ3(s). The regression function $f_{\mathrm{EQ}}(\cdot)$ can be approximated by the earthquake frequencies of the training data. The actual formula is a weighted sum of Gaussian kernels $K(x,x_i)=\exp(-\gamma\|x-x_i\|^2)$ of the form (7).

If solar activities have no effect on EQ3, the regression function with additional SA variables given by

$$\mathrm{EQ3}(t+1) \quad = \quad f_{\mathrm{EQ\cup SA}}(\cdot)\left(\,\left[y(t-d_y+1:t)^T,\, x(t-d_x+1:t)^T\right]^T\,\right)+\mathrm{error} \tag{10}$$

would show no improvement from (9), where $x(t-d_x+1:t):9d_x\times 1$ is similarly defined as $y(t-d_y+1:t)$ with the maximum time delay d_y of SA. The regression function $f_{\mathrm{EQ\cup SA}}(\cdot)$ is estimated using earthquake and SA time series of training data in the form (7).

Parameter estimation and test correlations

To obtain the optimal model structure, ten-fold cross validation was used to tune the model parameters of SVR on the basis of the training samples. Here, the trade-off parameter C was chosen from among 81 values $\exp(-8.0)$, $\exp(-7.8)$, $\exp(-7.6), ..., \exp(8.0)$, which maximized the test-correlation derived in the cross validation. The optimal model was then applied to the remaining test samples and its prediction performance was evaluated. The SVR was that of the R library `kernlab`[31]. The above estimation procedure was used for the predictions of $EQ3(t+\nu)$, $\log(EQ3(t+\nu)+1)$,..., $EQ8(t+\nu)$, $\log(EQ8(t+\nu)+1)$ for $\nu = 1, 2, ..., 5$. The prediction performance for $\log(EQ3(t+\nu)+1)$ was always better than that for $EQ3(t+\nu)$, and this is true for the larger earthquakes.

References

1. Kanamori, H. Earthquake prediction: An overview. *Int. Geophys.* **81,** 1205-1216 (2003). doi: 10.1016/S0074-6142(03)80186-9
2. Geller, R. J., Jackson, D. D., Kagan, Y. Y. & Mulargia, F. Earthquakes cannot be predicted. *Science*. **275,** 1616-1616 (1997). doi: 10.1126/science.275.5306.1616
3. Bolt, B. A. *Earthquakes and geological discovery.*: A Scientific American Library, (1993).
4. Cicerone, R. D., Ebel, J. E. & Britton, J. A systematic compilation of earthquake precursors. *Tectonophysics* **476,** 371-396 (2009). doi: 10.1016/j.tecto.2009.06.008
5. Aggarwal, Y. P., Sykes, L. R., Simpson, D. W. & Richards, P. G. Spatial and temporal variations in ts/tp and in P wave residuals at Blue Mountain Lake. *J. Geophys. Res.* **80,** 718-732 (1975). doi: 10.1029/JB080i005p00718
6. Campbell, W. H. Natural magnetic disturbance fields, not precursors, preceding the Loma Prieta earthquake. *J. Geophys. Res.* **114**, A5 (2009). doi: 10.1029/2008JA013932
7. Qiao, X., Wang, Q., Yang, S., Li, J., Zou, R. & Ding, K. The 2008 Nura Mw6.7 earthquake: A shallow rupture on the Main Pamir Thrust revealed by GPS and InSAR. *Geod. Geodyn.* **6**, 91-100 (2015). doi: 10.1016/j.geog.2015.01.005
8. Panzera, F., Zechar, J.D., Vogfjörd, K.S. & Eberhard, D.A.J. A revised earthquake catalogue for South Iceland. *Pure Appl. Geophys.* **173,** 97-116 (2016). doi:10.1007/s00024-015-1115-9
9. Zhang, S. & Zhou, S. Spatial and temporal variation of the b-value in Southwest China. *Pure Appl. Geophys.* **173,** 85-96 (2016). doi: 10.1016/0031-9201(94)90092-2
10. Zhang, S., Wu, Z. & Jiang, C. The Central China North- South Seismic Belt: Seismicity, Ergodicity, And Five-Year PI Forecast in Testing. *Pure Appl. Geophys.* **173,** 245-254 (2016). doi: 10.1007/s00024-015-1123-9
11. Schorlemmer, D., Wiemer, S. & Wyss, M. Variations in earthquake-size distribution across different stress regimes. *Nature* **437**, 539-542 (2005). doi:10.1038/nature04094
12. Nishikawa, T. & Ide, S. Earthquake size distribution in subduction zones linked to slab buoyancy. *Nat. Geosci.* **7**, 904-908 (2014). doi:10.1038/ngeo2279
13. Ide, S., Yabe, S. & Tanaka, Y. Earthquake potential revealed by tidal influence on earthquake size-frequency statistics. *Nat. Geosci.* **9**, 834-837 (2016). doi:10.1038/ngeo2796
14. Ogata, Y. Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Stat. Assoc.* **83,** 9-27 (1988). doi: 10.1080/01621459.1988.10478560
15. Schoenberg, F.P. Multidimensional residual analysis of point process models for earthquake occurrences, *J. Am. Statist. Ass.* **98**, 789–795 (2004). doi: 10.1198/016214503000000710
16. Wang, T., Zhuang, J., Obara, K. & Tsuruoka, H. Hidden Markov modelling of sparse time series from non-volcanic tremor observations. *J. Roy Stat. Soc. C-App.*, doi:10.1111/rssc.12194 (2016). doi: 10.1111/rssc.12194
17. Metivier, L., Viron, O.D., Conrad, C.P., Renault, S., Diament, M. & Patau, G. Evidence of earthquake triggering by the solid earth tides. *Earth Planet. Sc. Lett.* **278,** 370-375 (2009). doi: 10.1016/j.epsl.2008.12.024
18. Palumbo, A. Gravitational and geomagnetic tidal source of earthquake triggering. *IL Nuovo Cimento C* **12**, 685–693 (1989). doi: 10.1007/BF02511969
19. Hainzl, S., Kraft, T., Wassermann, J., Igel, H. & Schmedes, E. Evidence for rainfall-triggered earthquake activity. *Geophys. Res. Lett.* **33**(19) (2006). doi: 10.1029/2006GL027642

20. Kerr, R. & Stone, R. A human trigger for the Great Quake of Sichuan? *Science* **323**, 322-322 (2009). doi: 10.1126/science.323.5912.322

21. Redmayne, D. W. Mining induced seismicity in UK coalfields identified on the BGS National Seismograph Network. *Geol. Soc. Spec. Publ.* **5,** 405-413 (1988). doi: 10.1144/GSL.ENG.1988.005.01

22. Van Eijs, R. M. H. E., Mulders, F. M. M., Nepveu, M., Kenter, C. J. & Scheffers, B. C. Correlation between hydrocarbon reservoir properties and induced seismicity in the Netherlands. *Eng. Geol.* **84,** 99-111 (2006). doi: 10.1016/j.enggeo.2006.01.002

23. Simpson, J.F. Solar activity as a triggering mechanism for earthquakes, *Earth Planet. Sc. Lett.* **3**, 417–425 (1967). doi: 10.1016/0012-821X(67)90071-4

24. Space Weather Prediction Center: `http://www.swpc.noaa.gov/`.

25. Love, J.J. & Thosma, J.N. Insignificant solar-terrestrial triggering of earthquakes. *Geophys. Res. Lett.*, **40**, 1165–1170 (2013). doi: 10.1002/grl.50211

26. Odintsov, S.D., Ivanov-Kholodnyi, G.S., and Georgieva, K. Solar activity and global seismicity of the earth, *Bull. Russ. Acad. Sci. Phys.* **71**, 593–595 (2007). doi: 10.3103/S1062873807040466

27. Qin, P., Yamasaki, T. & Nishii, R. Statistical detection of the influence of solar activities to weak earthquakes. *Pac. J. Math. Ind.* **6,** 1–8 (2014). doi:10.1186/s40736-014-0006-9

28. OMNIWeb: http://omniweb.gsfc.nasa.gov.

29. ANNS - Advanced National Seismic System: http://earthquake.usgs.gov/monitoring/anss/.

30. Bishop, C.M. *Pattern Recognition and Machine Learning*. Springer (2006)

31. Karatzoglou, A., Smola, A.& Hornik, K. Package "kernlab". https://cran.r-project.org/web/packages/kernlab/kernlab.pdf (2016)

Acknowledgements

This research was supported by a Grant-in-Aid for Scientific Research (B) 15H02670.

Author contributions statement

R.N. conceived the application of SVR to earthquake data. Q.P. provided data sets, R.N. and R.K. conducted the numerical experiments. All authors reviewed the manuscript.