Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Investigation of the correlation of successive earthquakes preceding main shocks in the Greek territory

D. Chorozoglou, D. Kugiumtzis & E. Papadimitriou

To cite this article: D. Chorozoglou, D. Kugiumtzis & E. Papadimitriou (2021): Investigation of the correlation of successive earthquakes preceding main shocks in the Greek territory, Journal of Applied Statistics, DOI: 10.1080/02664763.2021.1939661

To link to this article: https://doi.org/10.1080/02664763.2021.1939661

Published online: 15 Jun 2021.

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2021.1939661

Taylor & Francis
Taylor & Francis Group

APPLICATION NOTE

Check for updates

Investigation of the correlation of successive earthquakes preceding main shocks in the Greek territory

D. Chorozoglou[a], D. Kugiumtzis[b] and E. Papadimitriou[a]

[a]Department of Geophysics, School of Geology, Aristotle University of Thessaloniki, Thessaloniki, Greece; [b]Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece

ABSTRACT

The Canonical Correlation Analysis (CCA) estimates the correlation between two vector variables by maximizing the correlation of linear combinations of their respective components. Here, the CCA is used to find correlation patterns in the last five successive, per pairs, earthquakes ($M \geq 4.0$) preceding 271 main shocks ($M \geq 5.5$) that occurred in the Greek territory during 1964–2018. The vector variables have two components, the earthquake magnitude and interevent time. The statistical significance of CCA is determined by the standard parametric test along with two proposed randomization tests, one using random shuffling of each paired dataset and one using randomly selected pairs of successive earthquakes. Simulations were designed on synthetic data from vector variables having the statistical characteristics of the real observations. The results on uncorrelated variables showed the correct size for the two randomization tests but larger type I error for the parametric significance test for small sample size. For correlated variables, the test power was equally high for both test types. The application of CCA and the significance tests to the Greek seismicity evidence the significant correlation among the last five successive preshocks, proving to be a promising tool in an a posteriori short-term earthquake forecasting.

ARTICLE HISTORY
Received 12 April 2020
Accepted 1 June 2021

KEYWORDS
Canonical correlation analysis (CCA); randomization significance test; preshock & main shock; Greek seismicity; a posteriori short-term earthquake forecasting

1. Introduction

The investigation of the complex seismicity behavior constitutes a major scientific challenge for improving our knowledge on seismogenesis and for seismic hazard assessment. Statistical analysis is engaged for revealing characteristics of this complexity. Basic among them is the Gutenberg–Richter law [20] that establishes the power law type of the earthquake occurrence frequency as a function of the released seismic energy, and the Omori law [32,41,42] expressing the decaying aftershock rate. The most widely applied, yet simple, model for the seismic hazard assessment is the Poisson model built under the assumption of strong earthquakes occurrence independence in time and space [28,6,8,29]. The hidden Markov [31,23,43] and semi-Markov models [3,37], are considered as suitable tools for earthquake probability estimation, remove the independence assumption and

CONTACT D. Chorozoglou chorozod@geo.auth.gr

let the earthquakes depend on each other in space and time. Alternative approaches include autoregressive seismicity forecast models [11] and equivalent dimension transformation [33] for investigating the spatiotemporal development of seismicity. A recently emerging field for the seismic hazard assessment is based on network (graph) theory, which investigates the topology and dynamics of a complex system. The complex earthquake network analysis was introduced by Abe and Suzuki [1] to study seismicity as a spatiotemporal complex system. More recently, the investigation of earthquake networks topology (structure) has been performed, among others, for the seismic hazard assessment by Daskalaki *et al.* [13] and Chorozoglou *et al.* [12] in Italy and Greece, respectively.

In this work, we use the Canonical Correlation Analysis (CCA) as a statistical tool to investigate whether there are patterns in earthquake occurrence times and magnitudes of successive earthquakes (preshocks) preceding a strong earthquake (main shock). The CCA is a classic and highly versatile statistical approach that gives a measure of linear correlation of two vector variables. It is for the first time applied for seismic hazard assessment, which is the ultimate goal of seismological research aiming to contribute to the seismic risk mitigation caused by strong earthquakes (e.g. $M \geq 5.5$). It is usually considered as the occurrence probability of strong earthquakes within a given time, space, and magnitude ranges. The CCA may account for all these three components, i.e. time, space and magnitude. Here, the space is fixed, and the analysis is performed on well-defined seismic subareas, and we address the CCA to the time and magnitude for seismic hazard assessment.

The concept of correlation introduced by Sir Francis Galton [18] and formulated mathematically by Pearson [35,36] has been extended to other formulations, including partial correlation, all concerning two scalar variables. Canonical Correlation, developed by Hotelling [25], extends the correlation to two sets of random variables (vector variables), defined as the maximum correlation between any two linear combinations of the respective components of each set. The CCA has been applied in various research fields. In health science, e.g. strong correlation between health and demographic variables was reported in Dunn and Doeksen [16] and Masters and Wallston [30]. The CCA was further used in climatology, e.g. for the detection of multivariate patterns in magnetic data sets in Walden *et al.* [44], for the investigation of the interdependence between primary and secondary pollutants in Hsu [26], for the determination of simultaneous relationships between the sample variograms of the pollutants and the sample variograms of the atmospheric variables in De Iaco *et al.* [15] and for the investigation of the relation among winter precipitation indices with extreme events in the Emilia-Romagna region in Busuioc *et al.* [9]. CCA has also been used in biomedicine, e.g. correlation of antipsychotic agents for the treatment of chronic schizophrenia in Foucart [17] and Alonso *et al.* [2], along with diverse studies in learning methods [22]. In addition, the CCA was used in flight control systems [14], marketing [5,19], environment [27], epilepsy [48,50], finance [40], genomics [24,45,10] and geology [7]. The CCA was used by Theodoulidis *et al.* [39] to quantify the correlation between earthquake damage distribution and ambient noise spectral ratio.

The aim of our work is to investigate correlation patterns using the CCA in successive earthquakes (preshocks) that occurred before an upcoming main shock ($M \geq 5.5$). Particularly, the pairs of preshocks ordered according to their temporal closeness to the main shock are 1st–2nd, 2nd–3rd, 3rd–4th and 4th–5th. The sets (vectors) in CCA comprise two components, the magnitude of the current preshock and the temporal, in days, succession to the next preshock for preshocks of order 2–5, and to the main shock for the first preshock. The analysis is done separately at each one of the 10 well-defined seismic zones of the Greek territory. We investigate the correlation among, only, the last five successive earthquakes that occurred just before a particular main shock, which form four pairs (i.e. 1st–2nd, 2nd–3rd, 3rd–4th and 4th–5th). The analysis is not feasible for all main shocks with $M \geq 5.5$ that occurred in the study period, because there are cases where the number of preshocks between two main shocks is smaller than five. The statistical significance of the canonical correlation is determined using a standard parametric test and two proposed randomization tests, one using the randomization based on shuffling and one designed for the setting of our study making use of all the ordered preshocks. A methodological task of the study is to compare the proposed randomization tests with the standard parametric test. The significance tests are assessed in a systematic simulation study involving scenarios with and without inter-dependencies in seismic data. Then, the CCA with the suitable significance tests is applied to the observations of the regional seismic catalog that covers the period 1964–2018.

We first present in Section 2 the theory upon which the CCA is based and the significance tests for its validation. The data are presented in Section 3. The simulation study and the results are shown and discussed in Section 4, and the results of the application to the Greek seismicity are presented and discussed in Section 5. The concluding remarks and discussion are given in Section 6.

2. Methodology

In the following, the CCA and the standard parametric significance test for CCA are briefly presented. Then, the proposed methodology with two approaches for the generation of the randomized data sets for the significance test of CCA is described.

2.1. Canonical correlation analysis (CCA)

The CCA quantifies the linear correlation of two multivariate (vector) variables and seeks for a pair of transforms as linear combinations of the components of each vector variable, such that the transformed scalar variables are maximally correlated [5,21]. The linear transform can be seen as a transform of the physical base of the state space of each vector variable to a new base with respect to the best alignment, i.e. correlation, of the points of the two vector variables in the new corresponding bases. The dimensionality of these new bases is equal to or less than the smallest dimensionality of the two variables. The linear combinations are called canonical variates and their correlations are called canonical correlations. A significant property of canonical correlations is that they are invariant with respect to affine transformations of the two variables while the classical correlation analysis depends on the base in which the variables are described.

Suppose that we have two sets (vectors) of random variables, $X = (X_1, \ldots.., X_m)$ and $Y = (Y_1, \ldots.., Y_l)$, and a sample of n observations for each vector variable X and Y given as $\{x_{k,1}, \ldots, x_{k,n}\}$ for $k = 1, \ldots, m$ and $\{y_{k,1}, \ldots, y_{k,n}\}$ for $k = 1, \ldots, l$, respectively. Let the sample covariance of X and Y be S_{XX} and S_{YY}, respectively, and the sample cross-covariance of X and Y be S_{XY}. The goal of CCA is to find pairs of linear combinations of the variables in X, say $w_X^T X$ (where the T in the superscript denotes transpose), and in

Y, say $w_Y^T Y$, which are maximally correlated:

$$((w_X^*, w_Y^*) = \arg\max)$$
$$= \arg\max_{w_X, w_Y} \text{Corr}(w_X^T X, w_Y^T Y) = \text{argmax}_{w_X, w_Y} \frac{w_X^T S_{XY} w_Y}{\sqrt{w_X^T S_{XX} w_X w_Y^T S_{YY} w_Y}}. \quad (1)$$

The correlation coefficient, Corr, in Equation (1) is not affected by the scaling of w_X and w_Y and the denominator can be set to one. Hence, one can solve the equivalent problem:

$$(w_X^*, w_Y^*) = \text{argmax}_{w_X, w_Y} w_X^T \Sigma_{XY} w_Y, \quad (2)$$

where Σ_{XY} is the cross-covariance matrix of random variables X and Y. When finding multiple pairs of vectors (w_X^i, w_Y^i), which construct the new base of the space of dimension $\min(m, l)$ for each one of X and Y, subsequent transforms are constrained to be uncorrelated with previous ones. To achieve this, the same maximization step is repeated on the residuals of the pair found in the previous step. Analytic expressions can be derived for the pair (w_X^*, w_Y^*) at each step up to $\min(m, l)$. Here, we are only interested in the first canonical correlation given by the first eigenvalue, r_1, of matrix $S_{XX}^{-1/2} S_{XY} S_{YY}^{-1/2} S_{YX} S_{XX}^{-1/2}$.

2.2. The parametric significance test for CCA

For the parametric significance test of the first eigenvalue r_1, we note that if the true first canonical correlation is zero then all the $\min(m, l)$ true canonical correlations are also zero, and thus the null hypothesis yields the true cross-covariance Σ_{XY}, i.e. $H_0 : \Sigma_{XY} = 0$, where 0 is the zero matrix of dimension $m \times l$. The test statistic [4] is given by

$$\chi^2 = -(n - 1 - 0.5(m + l + 1)) \ln \prod_{i=1}^{\min(m,l)} (1 - r_i^2), \quad (3)$$

where r_i is the ith canonical correlation defined in the same way as r_1. The χ^2 is asymptotically distributed as chi-square with ml degrees of freedom, $\mathcal{X}_{ml}^2$. Here after, we use the first canonical correlation as the CCA measure and denote it as r. For small sample size n, the distribution of χ^2 may deviate significantly from $\mathcal{X}_{ml}^2$, and the parametric test may not be accurate.

2.3. The randomization significance tests

We propose two randomization significance tests applied directly to the CCA measure r, i.e. assuming the same H_0 of no correlation among variables and having r (rather than χ^2) as test statistic. Both tests rely on generating an ensemble of randomized samples consistent to H_0. On these samples the measure r is computed, and in this way the null empirical distribution of r is formed. The randomized samples are generated using a simple constrained realization approach, the randomization based on shuffling in the first test and a typical realization approach based on all the data from seismic catalog in the second test [38].

For the first test, the n multivariate data points of each vector variable X and Y are randomly shuffled, such that any relationship may exist among the data points of X and Y at the same index $i = 1, \ldots, n$ is destroyed. Each realization of a new pair of randomly shuffled samples of size n considers independent X and Y and is consistent to H_0. We refer to this randomization test as *shuffling test*.

The second test is not defined generally for any vector variables X and Y, but refers specifically to our setting of ordered preshocks, so that X and Y denote a preshock of any of the orders from one to five getting two components, its magnitude and its temporal distance from the next preshock ($m = l = 2$). For the pair of X and Y of a maximum order, say p, the randomized sample for the variables X and Y of size n is randomly selected from the pool of all earthquakes of order larger than p (and after the previous main shock, thus here we use the data for all earthquakes of magnitude $M \geq 4.0$ in the seismic catalog). For example, if the pair of the 3rd–4th preshocks is examined (i.e. $p = 4$) then the randomized sample of X and Y is derived from different randomly selected order $p \geq 5$ (i.e. 5th–6th, 6th–7th, etc.) preshocks. In this way, the possible correlation due to the preshock order is destroyed because the randomized sample of X and Y includes successive earthquakes with random temporal relevance to the main shocks. This second randomization significance test is called *adaptive test*.

For any of the two randomization significance tests, we generate $B = 100$ randomized samples and compute the CCA measure r on the original sample and on the B randomized samples, denoted as $r^{*1}, \ldots, r^{*B}$. The p-value for the one-sided test (supposing that it is improbable to have r on the left tail of the null distribution) is given by

$$p = 1 - \frac{i_0 - 0.326}{B + 1 + 0.348}, \tag{4}$$

where i_0 is the rank of r in the ordered list of $r, r^{*1}, \ldots, r^{*B}$, and the correction is used for the empirical cumulative distribution [49].

The proposed approach with the two tests, i.e. the shuffling test and the adaptive test, is used because the underlying distribution of the observed variables is unknown and thus the null distribution of the test statistic is not known, as opposed to the parametric test where the null distribution is taken to be asymptotically the chi-square χ^2. To this respect, the proposed randomization tests are more general.

3. Data

We utilize the seismic catalog compiled from the central Seismological Station of the Geophysics Department of the Aristotle University of Thessaloniki (doi:10.7914/SN/HT). In our analysis, we focus on the last five preshocks that occurred until 180 days before each of 271 main shocks ($M \geq 5.5$) during 1964-2018. In this period from 1964 to 2018 occurred 341.453 crustal earthquakes (focal depth less than 50 Km) of any magnitude. We consider these earthquakes, that are 23.908, which ensure the magnitude of completeness which is $M_{\text{cthrhh}} = 4.0$ (Figure 1) as identified by the goodness-of-fit method [47].

The zonation scheme (based on faulting type, seismic moment rate and other seismotectonic criteria) of Papaioannou and Papazachos [34] is adopted for the division of the study area into smaller seismic zones. In our analysis, we integrated several of the suggested 67 seismic zones keeping the aforementioned criteria, into 10 seismic subareas for getting a sufficient number of earthquakes in each one of them (Figure 2). The simulation study is

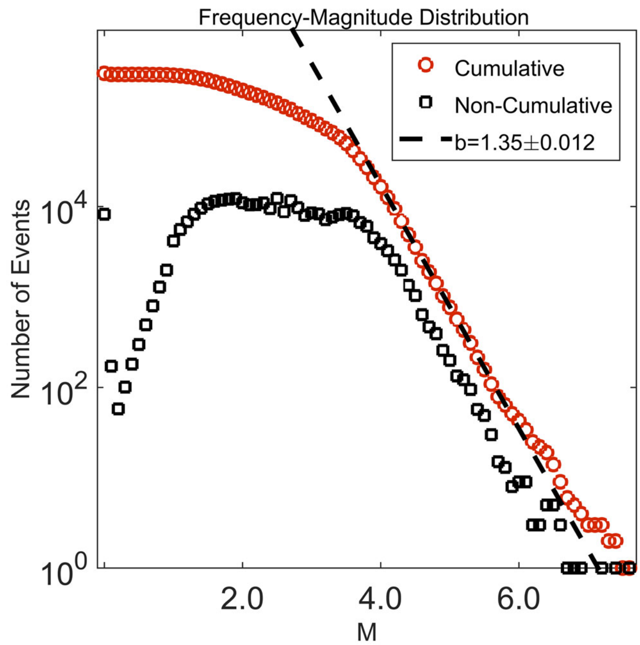

Figure 1. Maximum Likelihood estimation for the completeness magnitude, which equals to 4.0, as calculated from the goodness-of-fit method.

performed on each one of the 10 larger seismic subareas (Section 4) and then the CCA is applied to the data set in each seismic subarea (Section 5).

4. Simulation study

4.1. The simulation setup

The simulation study is performed using the data on the 271 main shocks ($M \geq 5.5$) that occurred in the 10 seismic subareas and the five most temporally close preshocks for each main shock. We consider scenarios of different lower magnitude threshold M_{thr} for the foreshocks ranging as $4.0 \leq M_{\text{thr}} \leq 4.7$, i.e. $M_{\text{thr}} = 4.0$, $M_{\text{thr}} = 4.1, \ldots$,$M_{\text{thr}} = 4.7$, and their possible upper magnitude is $M = 5.4$, because the magnitude threshold for the main shocks is $M \geq 5.5$. For each of the five ordered foreshocks, $i = 1, \ldots, 5$, we collect the data on their magnitude and time (days to the next earthquake) for all main shocks in the catalog at each seismic zone and compute the mean and standard deviation of the magnitude, $\mu_{i,m}$ and $\sigma_{i,m}$, and time, $\mu_{i,t}$ and $\sigma_{i,t}$, respectively. We consider pairs of successive ordered foreshocks, i and $i+1$, for $i = 1, \ldots, 4$, and generate a bivariate sample of size n of magnitude and time for foreshocks of order i and $i+1$, denoted as $(m_{i,j}, t_{i,j})$ and $(m_{i+1,j}, t_{i+1,j})$, respectively, for $j = 1, \ldots, n$. The magnitudes of foreshocks of order i and $i+1$ are drawn from a bivariate normal distribution as

$$\begin{bmatrix} m_{i,j} \\ m_{i+1,j} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{i,m} \\ \mu_{i+1,m} \end{bmatrix}, \begin{bmatrix} \sigma^2_{i,m} & \rho_m \sigma_{i,m} \sigma_{i+1,m} \\ \rho_m \sigma_{i,m} \sigma_{i+1,m} & \sigma^2_{i+1,m} \end{bmatrix} \right), \text{ for } j = 1, \ldots, n, \quad (5)$$

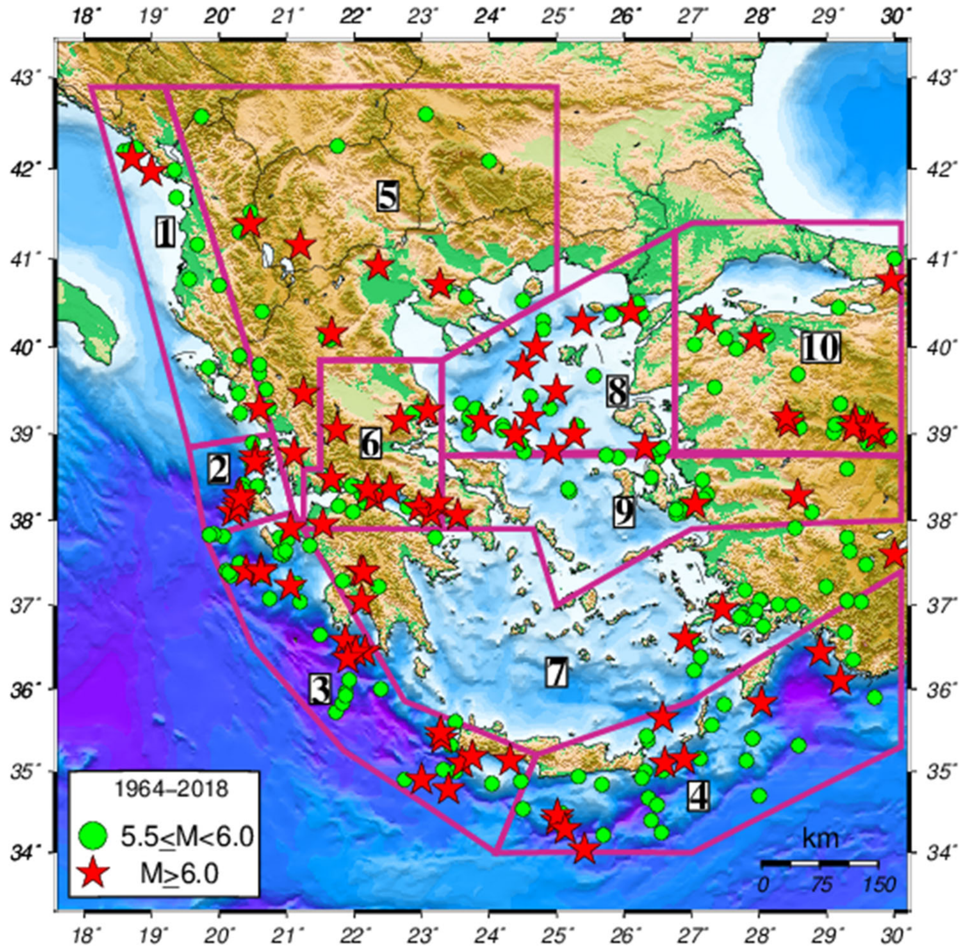

Figure 2. Epicentral distribution of the 271 strong ($M \geq 5.5$) earthquakes that occurred in 1964–2018 in the broader area of Greece, upon which the definition of the ten subareas is based.

and the same holds for the times

$$\begin{bmatrix} t_{i,j} \\ t_{i+1,j} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{i,t} \\ \mu_{i+1,t} \end{bmatrix}, \begin{bmatrix} \sigma^2_{i,t} & \rho_t\sigma_{i,t}\sigma_{i+1,t} \\ \rho_t\sigma_{i,t}\sigma_{i+1,t} & \sigma^2_{i+1,t} \end{bmatrix}\right), \text{ for } j = 1, \ldots, n, \qquad (6)$$

where ρ_m and ρ_t are the Pearson correlation coefficients of the magnitudes and times of the successive ordered foreshocks, respectively. We use the bivariate normal distribution for the construction of simulated data as it can describe any set of random variables (here, magnitude and time) each of which clusters around a mean value.

In the simulations, we consider a single free parameter which defines the correlation among vector variables and set $\rho = \rho_m = \rho_t$. We introduce this simplification in our simulation study to examine only two settings, one for independence ($\rho_m = \rho_t = 0$) and one for dependence ($\rho_m = \rho_t = 0.8$). We set the sample size n to be equal to the number of the pairs of the true successive foreshocks of orders i and $i + 1$ in the catalog for the specific seismic subarea. This means that as the order increases the sample size may be smaller due to the lack of foreshocks of higher orders for certain main shocks. In this way, we generate synthetic pairs of bivariate samples of magnitude and time, which preserve the real data conditions for each pair of foreshock orders and seismic zone and have a correlation determined by the free parameter ρ.

We consider two main settings with regard to the correlation of the ordered foreshocks, assigning $\rho = 0$ to assess the specificity of the CCA measure r where we expect to have low in significant values of r, and $\rho = 0.8$ to assess the sensitivity of r where we expect high significant values of r. For $\rho = 0$ we investigate whether the parametric and the two randomization significance tests obtain the correct size of the test (probability of rejection of H_0, being true, at the nominal significance level, here set equal to $\alpha = 0.05$). For $\rho = 0.8$ we compare the power attained with each of the three tests quantified by the probability of rejection of H_0, which is not true. To obtain statistical results for the CCA, we perform 100 Monte Carlo realizations for each setting (i.e. $\rho = 0$ and $\rho = 0.8$) and for a range of values of the magnitude threshold M_{thr} ($4.0 \leq M_{\text{thr}} \leq 4.7$). The statistical results on the synthetic data that preserve the data conditions of the real foreshocks allow us to investigate the limitations in the estimation of the CCA measure and the significance tests to be applied to the real data.

We show two exemplary settings to illustrate the simulation study for the pair of foreshocks of 2nd and 3rd order using $M_{\text{thr}} = 4.0$ in the seismic subarea 3. Figure 3 shows five of the totally $n = 44$ main shocks of seismic subarea 3 which are shown together with the foreshocks, where the foreshocks of 2nd and 3rd order are highlighted.

Bivariate samples of magnitudes and times of the same size $n = 44$ are generated by Equations (5) and (6) and $\rho = 0$, based on the statistics for the 44 foreshocks of 2nd and 3rd order that occurred before the 44 main shocks in the seismic subarea 3. The scatter plot, which illustrates the degree of correlation between two variables, of the magnitudes of the 2nd and 3rd preshocks is shown in Figure 4(a), and the respective scatter plot for times in Figure 4(b).

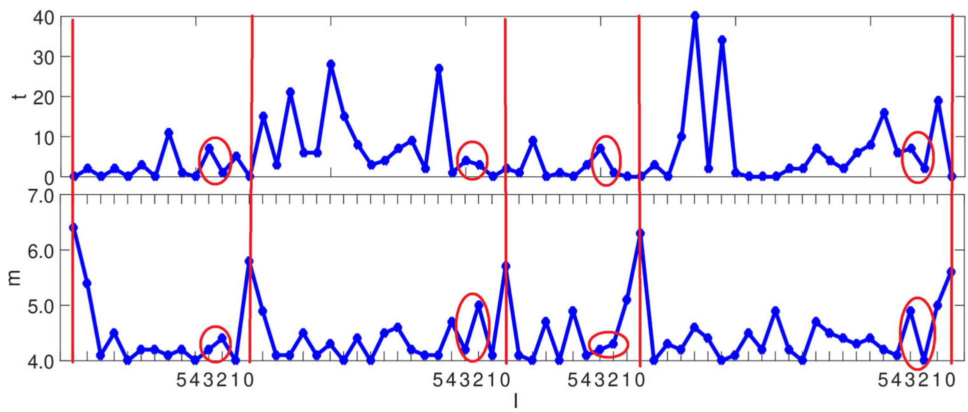

Figure 3. The interevent times (upper panel) and magnitudes (lower panel) of the preshocks of the 44 main shocks ($M \geq 5.5$) in seismic subarea 3 denoted by the vertical lines. The magnitudes and interevent times of the foreshocks of 2nd and 3rd order are highlighted in circles. The scale in the abscissa is for the time order and not for real time intervals.

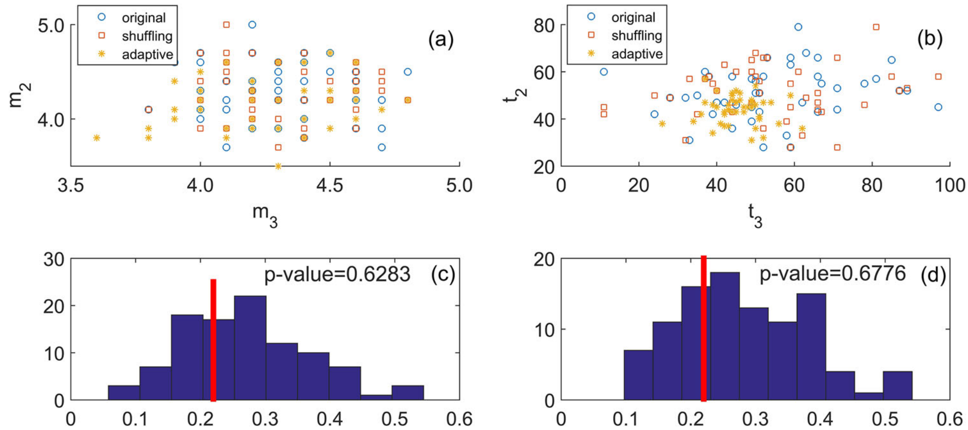

Figure 4. Scatter plot of the magnitudes in (a) and interevent times in (b) for bivariate samples generated on the basis of the 2nd and 3rd preshocks order for 44 main shocks in the seismic subarea 3 and using $\rho = 0$. Superimposed are the data points of the corresponding randomized samples for the shuffling and the adaptive tests, as shown in the legend. The histogram of the 100 r values are shown in (c) for the randomized data of shuffling type and in (d) for the adaptive randomized data, respectively, along with the r value (vertical line) for the original sample.

Figure 4, shows superimposed the points from the same quantities but for the randomized data for the shuffling and the adaptive test. The scatter is the same for both the originally generated data, obtained by the Equations (5) and (6) with $\rho = 0$, and for the randomized data of both types (shuffling and adaptive) as the original data are uncorrelated. The values of the CCA measure r are 0.22 for the original data, 0.34 for the randomized data of shuffling type and 0.38 for the adaptive randomized data, suggesting similar levels of correlation of the vector variables of magnitude and time for the 2nd and 3rd preshocks. We performed the shuffling and the adaptive test using $M = 100$ randomized samples and the histogram of the 100 r values are shown in Figure 4(c,d), respectively, along with the r value for the original sample. Obviously, the r value (red line in Figure 4(c,d)) for the original sample is well within the null empirical distribution for the shuffling (blue bars in Figure 4(c)) and adaptive test (blue bars in Figure 4(d)) giving p-values 0.63 and 0.68 by the Equation (4), respectively, while the parametric test gave 0.67 by the Equation (3), all well above the significance level of $\alpha = 0.05$. We note here that whereas the magnitude and the time distribution of the randomly shuffled data is identical to the original ones, this does not hold for the adaptive randomized data having distributions of smaller mean and standard deviation than shuffled data for both magnitudes and times. This happens because the adaptive test generates the randomized data from bigger sample than the other tests as it includes all the preshocks with larger order. The latter holds for both vector variables (time and magnitude) in the same way, so that it does not add bias in the computation of r.

The same example is repeated but now for correlated magnitudes and times of the 2nd and 3rd preshocks, using $\rho = 0.8$, instead of $\rho = 0$. Both scatter plots, in Figure 5(a,b) for the magnitudes and times, respectively, show that the points for the originally generated data are close to the diagonal (correlated) while for the randomized data of both types are scattered in the same way as for the case of $\rho = 0$ (Figure 4).

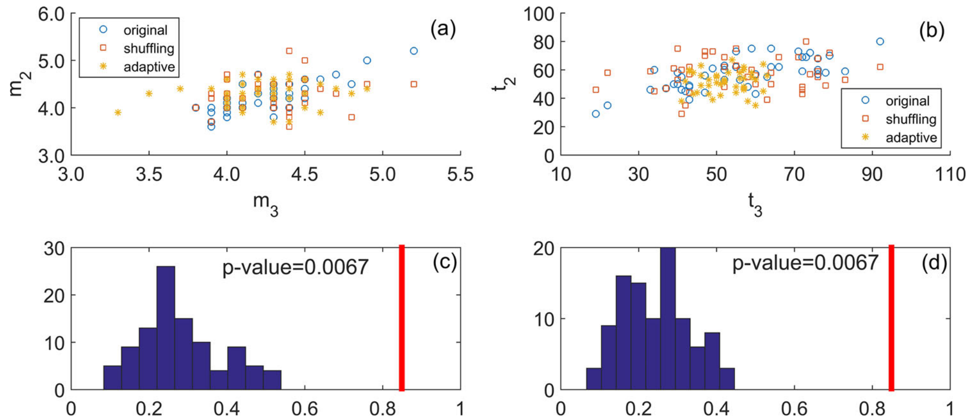

Figure 5. Same as Figure 4 but for the dependent vector variables using $\rho = 0.8$.

The r for the original data gets a value of 0.85, much larger than the r values of 0.41 and 0.21, for the shuffling and adaptive randomized data, respectively. The histograms of the 100 r values for the shuffling test and for the adaptive test, shown in Figure 5(c,d), respectively, are placed well below the r value for the original sample. Thus, the H_0 of uncorrelated vector variables of magnitudes and times is rejected with high confidence and the p-values are equal to 0.0067, 0.0067 and 0.0009 for the parametric test, the shuffling test, and the adaptive test, respectively. Note that the value of 0.0067 is the smallest p-value that can be obtained using $B = 100$ and implies that the r value on the original data set is larger than the 100 r values on the randomized data set (see Equation (4)).

4.2. Results on independent vector variables

In the first part of the simulation study, the vector variables are independent, and the synthetic data are generated for $\rho = 0$ by the Equations (5) and (6). The setting of independent vector variables is consistent with the null hypothesis H_0. The H_0 for the randomization test is incorrectly rejected when p - value < 0.05 in Equation (4) while for the parametric test the p - value is derived from χ^2 in Equation (3). Then, Monte Carlo realizations are generated using as parameters in Equations (5) and (6) the statistics of the earthquakes in each subarea and different M_{thr}. Having $B = 100$ realizations for each scenario we compute the percentage of rejection H_0, P, for the parametric and the two randomization significance tests. We report summary results for the specificity of each test from all subareas.

The average of the percentage of rejection of H_0, $\bar{P}$, in the 100 realizations in each subarea is shown as a function of the M_{thr} in Figure 6 for the three significance tests and for the four preshock pairs (one at each panel). The parametric test tends to reject H_0 with higher probability than $\alpha = 0.05$ (large type I error) when the magnitude threshold gets large ($M_{\text{thr}} \geq 4.3$), as shown in Figure 6. This is observed for any of the four pairs of ordered foreshocks, for some magnitude thresholds with $M_{\text{thr}} \geq 4.3$, that considered before each main shock.

The parametric test fails in the cases of small data size n (attained for large M_{thr}), for which the chi-square approximation of the null distribution of the parametric test statistic is not established. Hence, for these cases the parametric test tends to identify spurious

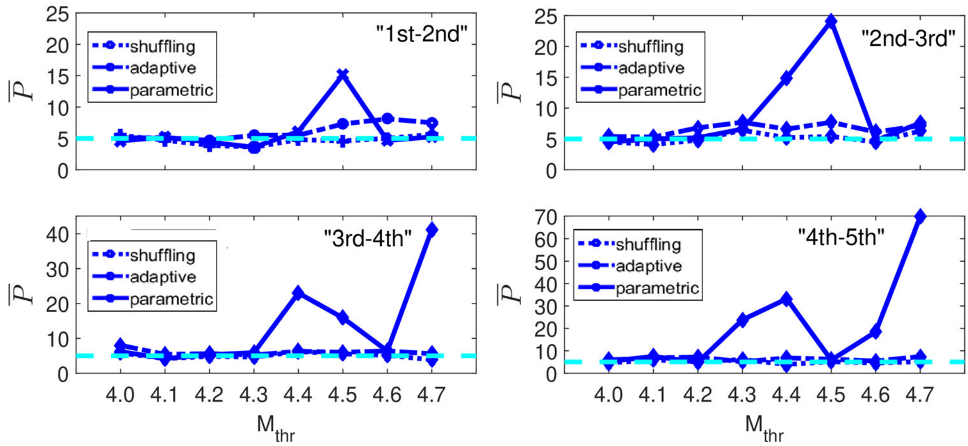

Figure 6. Summary results on the significance of the CCA measure r for simulations using independent vector variables ($\rho = 0$). The average percentage of rejection of H_0, $\bar{P}$, from 100 realizations over the 10 seismic subareas as a function of M_{thr} is given for the four pairs of ordered preshocks in the four panels. The horizontal gray (cyan online) stippled line is at the nominal significance level $\alpha = 0.05$ (5%).

correlation between the vector variables and has low specificity. On the other hand, the two proposed randomization tests always obtain $\bar{P}$ at the nominal significance level $\alpha = 0.05$ attaining the correct size of the test and having high specificity.

4.3. Results on dependent vector variables

In the second setting of the simulation study, we examine the sensitivity of CCA based on the three significance tests using dependent vector variables for $\rho = 0.8$ in the Equations (5) and (6). In this case, the null hypothesis H_0 is not true. The simulation setting is the same as for the independent vector variables and we report summary results for the sensitivity of each test from all 10 subareas. As for the independent vector variables, the average of the percentage of rejection of H_0, $\bar{P}$, over the 10 subareas is shown as function of the magnitude threshold M_{thr} in Figure 7 for the three significance tests and for the four ordered pairs (one at each panel). Here, H_0 is false because the variables are dependent, and we expect high rejection rates.

Indeed, high-rejection rates are obtained with all three significance tests and their power is at the maximum for all pairs of ordered foreshocks when the magnitude threshold is comparatively low ($M_{thr} \leq 4.3$), as shown in Figure 7. The $\bar{P}$ decreases in the same way for all three significance tests as the magnitude threshold increases ($M_{thr} \geq 4.4$) and the sample size decreases ($n \leq 10$). The small sample size ($n \leq 10$), because corresponding to high magnitude thresholds ($M_{thr} \geq 4.4$), does not allow for the correlation structure to be significantly estimated by CCA leading to failure of all statistical tests (parametric test, shuffling test and adaptive test).

4.4. Remarks of the simulation study

Putting together the results from the simulations on independent and dependent vector variables, it turns out that the two proposed randomization significance tests perform

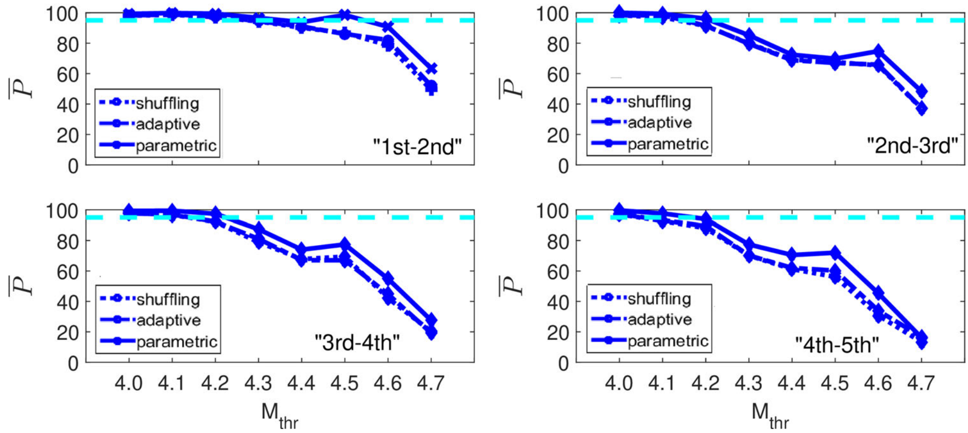

Figure 7. Same as Figure 6 but for simulations with dependent vector variables using $\rho = 0.8$.

appropriately for all data sets involving the statistical characteristics of the preshocks' magnitude and times in all subareas. Both randomization tests attain the correct size of the test for all magnitude thresholds and preshocks pairs (Figure 6) and get power that increases with the data size (Figure 7), i.e. for low magnitude thresholds ($M_{\text{thr}} \leq 4.3$), both being the favorable properties of a statistical test. On the other hand, the standard parametric significance test has a tendency to detect correlation in the vector variables when these are independent and give higher probability of rejection of the H_0 than the nominal significance level when the sample size n is small, i.e. when the magnitude threshold is high ($M_{\text{thr}} \geq 4.3$). This shortcoming of the parametric significance test questions its use in applications, especially for small sample sizes. The fact that for dependent vector variables all three significance tests perform similarly, further approves the use of the two randomization tests as it asserts that they are not conservative and attain the same power as the parametric test.

5. Application to the seismic data

The vector variables in the CCA include the earthquake magnitudes and interevent times of earthquakes with $M \geq 4.0$ that occurred in the Greek territory during 1964–2018. The final goal of our work is to investigate the existence of correlation patterns for magnitude and time in the Greek seismicity using CCA in the last five preshocks. The catalog contains 271 main shocks ($M \geq 5.5$) and the CCA is repeated for different sets of preshocks of magnitudes $M_{\text{thr}} \leq \text{M} \leq 5.4$, where the magnitude threshold varies in $4.0 \leq M_{\text{thr}} \leq 4.7$. The datasets of preshocks are formed separately for each of the 10 well-defined seismic zones of the Greek area. As in the simulation presented in Section 4, for the statistical significance of the CCA measure r, the parametric test and the two randomization tests are used. The null hypothesis H_0, for the parametric test and the two randomization tests, is that there is no correlation of preshocks of a specific order, for any of the order pairs 1st-2nd, 2nd-3rd, 3rd-4th and 4th-5th. H_0 for the adaptive test tends to be stricter in comparison with the other two tests, i.e. the parametric and shuffling test, because the correlation of the successive preshocks of a specific order (1st–2nd, 2nd–3rd, 3rd–4th and 4th–5th) is

Table 1. The cases where the H_0 of zero canonical correlation is rejected (p-value < 0.05) are shown as follows: the seismic subarea and toponym (1st column), the preshock pair (2rd column), the magnitude threshold (3th column), and type of significance test (4th column).

Seismic subarea and toponym	Preshock pair	M_{thr}	Test
1. Albania-Montenegro-West Greece	3rd–4th	4.2	Parametric
	4th–5th	4.0	All
2. Lefkada-Kefalonia	3rd–4th	4.1	All
		4.2	
	4th–5th	4.3	Parametric
3. Zante-Southwest Peloponnese-West Crete	2nd–3rd	4.1	Parametric
	3rd–4th	4.1	Shuffling
			Parametric
	4th–5th	4.0	Shuffling
			Parametric
		4.1	Shuffling
			Parametric
4. East Crete	1st–2nd	4.0	All
	3rd–4th	4.1	All
5. Macedonia	1st–2nd	4.0	All
		4.1	All
	2nd–3rd	4.1	All
	4th–5th	4.0	Parametric
6. Thessaly-Corinthian gulf	1st–2nd	4.0	Parametric
	2nd–3rd	4.1	Shuffling
			Parametric
	3rd–4th	4.0	Shuffling
			Parametric
		4.1	Parametric
Seismic subarea and toponym	Preshock pair	M_{thr}	Approach
7. Peloponnese-South Aegean	1st–2nd	4.0	Parametric
		4.2	Shuffling
			Parametric
		4.3	Shuffling
			Parametric
	2nd–3rd	4.0	Shuffling
			Parametric
		4.1	Shuffling
			Parametric
	3rd–4th	4.0	Shuffling
			Parametric
	4th–5th	4.0	Shuffling
			Parametric
		4.1	Shuffling
			Parametric
		4.2	All
		4.3	Shuffling
			Parametric
8. North Aegean	1st–2nd	4.0	Shuffling
			Parametric
		4.1	Shuffling
			Parametric
		4.2	Parametric
	3rd–4th	4.0	All
		4.3	Adaptive
	4th–5th	4.2	Parametric
Seismic subarea and toponym	Preshock pair	M_{thr}	Approach

(*continued*).

Table 1. Continued.

Seismic subarea and toponym	Preshock pair	M_{thr}	Test
9. Athens-South Aegean	1st–2nd	4.1	Shuffling
			Parametric
		4.2	All
	3rd–4th	4.1	Adaptive
		4.2	Shuffling
			Parametric
	4th–5th	4.0	Shuffling
			Parametric
		4.1	Shuffling
			Parametric
10. Northwest Turkey	2nd–3rd	4.0	Shuffling
			Parametric
		4.1	Shuffling
			Parametric
	4th–5th	4.3	All

compared with randomly, also, successive preshocks of any order larger than the tested one (presumably the high order pairs are not correlated). We recall that for larger magnitude thresholds, $M_{thr} > 4.3$, resulted to small sample size n, the parametric test does not attain the proper size and all the three tests have small power, and for this reason we stay on magnitude thresholds in the range $4.0 \leq M_{thr} \leq 4.3$ as the simulation study presents the proper results for these magnitude thresholds (i.e. acceptance and rejection of H_0 for independent and dependent variables, respectively). We consider the three significance tests for $M_{thr} = 4.0,\ 4.1,\ 4.2,\ 4.3$, the four pair orders, and the 10 seismic subareas and report in Table 1 the cases for which the H_0 of no correlation of successive preshocks is rejected (p - value < 0.05). We note that for each subarea there is statistical evidence for the presence of correlation in successive preshocks but for specific pair orders and M_{thr} and not always obtained with all three significance tests. Expect two the cases shown in Table 1, in all other cases where H_0 is rejected, this is obtained with the parametric test either alone or along with the one or both randomization tests. It is somehow anticipated, as the simulation study also revealed, that the parametric test presents a bias towards rejecting the H_0 (lower specificity). Even when we focus on the randomization significance tests, there is still evidence of correlation in successive preshocks for each subarea but not in any systematic way with respect to M_{thr} or pair order. In addition, the cases of rejections of the H_0 with the adaptive test indicate that the successive preshocks differ, in time and magnitude, from randomly selected pairs of successive earthquakes further away in time from the main shocks. As shown in Table 1, for all but seismic subareas 3 and 6, there are cases where the H_0 is rejected with the adaptive test, suggesting strict hierarchical organizations of preshocks. In subareas 3 and 6, the hierarchical organizations are shown with the least rigorous statistical tests, i.e. the parametric and shuffling tests, which only compare earthquakes occurring in the specific order by chance. Still, these findings might be considered promising for the seismic hazard assessment as there are statistically significant patterns of preshocks correlation in the study area.

6. Conclusion and discussion

The Canonical Correlation Analysis (CCA) proved to be a useful statistical tool to decode complex dependency structures, as the seismic activity, in multivariate data and identify groups of interacting variables. We use the CCA measure r to detect temporal correlation of preshocks for main shocks with $M \geq 5.5$ that occurred in Greece during 1964–2018. The presence of any such correlation is expected to be weak because of the complexity of seismicity and therefore we focus on the statistical significance of the CCA measure r. Hence, we consider a standard parametric test and two proposed randomization tests, one randomly shuffling the given multivariate data (called shuffling test) and one adapting a random resampling to the setting of all ordered foreshocks (called adaptive test).

The three significance tests are assessed in simulations generating uncorrelated ($\rho = 0$) and correlated ($\rho = 0.8$) multivariate data based on the statistics of the observed ordered preshocks. The results showed a bias of the parametric test for the detection of the correlation, as it gave larger percentage of rejections of the null hypothesis of no correlation when it was true $\rho = 0$, increasing with the decrease of sample size n/M_{thr}. On the other hand, the two proposed randomization tests attain high specificity, i.e. the percentage of rejection is at the nominal significance level ($\alpha = 0.05$) when $\rho = 0$ for any sample size. The parametric test and both randomization tests indicate high sensitivity, i.e. high percentage of rejection when $\rho = 0.8$, for the cases that the magnitude threshold is low ($M_{\text{thr}} \leq 4.3$). The failure of all statistical tests (parametric test, shuffling test and adaptive test) for high magnitude thresholds ($M_{\text{thr}} \geq 4.4$), when the sensitivity of each statistical test is examined, is due to the very low number of the time series observations. The parametric test attains somewhat higher sensitivity for small sample sizes but given the very low specificity (bias towards rejection) it cannot be trusted. The results from the simulated data show that they do not depend on the magnitude threshold M_{thr} but rather on the sample size n. Figure 6 reveals that when the sampling size n is sufficiently low (i.e. for high values of magnitude threshold M_{thr}) M_{thr} for the pairs '3rd-4th' and '4th-5th' both randomization approaches provide suitable results because the percentage of rejection fluctuates around the nominal significance level while the parametric test fails due to the small length n of time series.

We further use the two proposed randomization tests and the standard parametric test for investigating the correlation patterns in magnitude and time among the last five foreshocks occurring before each main shock in the study area using the tool of CCA. The three significance tests were performed for each pair of successive preshocks, considering also different M_{thr}, and for all subareas. We found in an adequate number of cases (combinations of pair order and M_{thr}), statistically significant CCA measure r mainly with the parametric and the shuffling test. The adaptive test is the most conservative among the three significance tests and gave the fewest rejections of the H_0, as it compares the successive preshocks close to the main shock (of orders one to five) to other successive earthquakes further away in time from the main shock. On the contrary, the other two significance tests only compare earthquakes taken by chance (earthquakes in pairs taken in any order for the shuffling test). These results reveal the presence of some type of hierarchical organization in the preshock activity. Hence, the CCA analysis as implemented here, may contribute to the short-term seismic hazard assessment because the statistical significance of the CCA measure r for the five ordered preshocks indicates the presence of correlation structure among them. For example, in the subarea 1 the 2^{nd}-3^{rd} preshock pair with $M_{\text{thr}} = 4.2$ and the parametric statistical test rejects the null hypothesis H_0 of zero canonical correlation (Table 1). This means that this pair presents statistical significance correlation and can be seen as a warning for the main shock occurrence after an earthquake of 1st order with $M_{\text{thr}} \geq 4.2$.

If there is any correlation in times and magnitudes between succession of earthquakes before the main shock this is more likely to be found among the temporally closest earthquakes to the main shock. Therefore, we restrict our analysis to the first four pairs (five earthquakes before the main shock). We do consider higher order pairs when we do the adaptive randomization test.

The CCA approach constitutes a good method to find linear relationships between observed vector variables but it is not able to extract nonlinear relationships. For this, extensions of CCA can be used, such as the nonlinear CCA and kernel-CCA (K-CCA) that are inspired by the concepts of sparse multiple kernel learning and kernel CCA, respectively. The analysis based on nonlinear correlation, in the same way as done here with CCA, is left for future work.

One more immediate extension of this analysis, left for future work is to involve another set of main shocks in the same seismic subarea as a test set, e.g. splitting the seismic catalog in the learning and testing period that includes 80% and 20% of earthquakes, respectively. For the cases where statistical significance of CCA was established (Table 1), assessment of whether significance also persists in the test set will be investigated. The investigation whether the non-successive preshocks (e.g. 2nd–4th) present stronger correlation than the successive ones, as they have been considered in this study, is also left for future work.

Acknowledgments

The map is generated using the Generic Mapping Tool [46]. All the calculations and figures are made using the Matlab software (www.mathworks.com/products/matlab). The Geophysics Department of Aristotle University of Thessaloniki is also gratefully acknowledged by authors DC and EP. Contribution number 950.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was financially supported by the European Union and Greece (Partnership Agreement for the Development Framework 2014–2020) for the project ‘Development and application of time-dependent stochastic models in selected regions of Greece for assessing the seismic hazard’, MIS5004504.

References

[1] S. Abe, and N. Suzuki, *Small-world structure of earthquake network*. Phys. A 337 (2004), pp. 357–362.
[2] A. Alonso, H. Geys, G. Molenberghs, M. Kenward, and T. Vangeneugden, *Validation of surrogate markers in multiple randomized clinical trials with repeated measurements*. J. Biomed 45(8) (2003), pp. 931–945.
[3] Y. Altınok, and D. Kolçak, *An application of the semi-Markov model for earthquake occurrences in North Anatolia, Turkey*. J. Balkan Geophys. Soc 2 (1999), pp. 90–99.
[4] M.S. Bartlett, *The general canonical correlation distribution*. Ann. Math. Statist. 18 (1947), pp. 1–17.
[5] S. Basfirinci, *An evaluation and an application of using canonical correlation analysis in marketing research*. Inter. J. Econ. Adm. Stud 18(3) (2009), pp. e23131.
[6] M. Bath, *Seismic risk in Fennoscandia*. Tectonophysics 57 (1979), pp. 285–295.

[7] J. Borovsky, *Canonical correlation analysis of the combined solar wind and geomagnetic index data sets*. J. Geophys. Res 119(7) (2014), pp. 5364–5381.
[8] D. Brillinger, *Some bounds for seismic risk*. Seismol. Soc. Am. Bull 72 (1982), pp. 1403–1410.
[9] A. Busuioc, R. Tomozeiu, and C. Cacciamani, *Statistical downscaling model based on canonical correlation analysis for winter extreme precipitation events in the Emilia-Romagna region*. Int. J. Clim 28(4) (2008), pp. 449–464.
[10] D.S. Cao, S. Liu, W.B. Zeng, and Y.Z. Liang, *Sparse canonical correlation analysis applied to omics studies for integrative analysis and biomarker discovery*. J. Chemom. 29 (2015), pp. 371–378.
[11] P. Chingtham, A. Tiwari, and A.K. Yadav, *Forecasting seismic activity rates in northwest Himalaya through multivariate autoregressive forecast of seismicity algorithm*. Acta Geophys. 67 (2019), pp. 465–476.
[12] D. Chorozoglou, D. Kugiumtzis, and E. Papadimitriou, *Testing the structure of earthquake networks from multivariate time series of successive main shocks in Greece*. Phys. A 499C (2018), pp. 28–39.
[13] E. Daskalaki, K. Spiliotis, C. Siettos, G. Minadakis, and G.A. Papadopoulos, *Foreshocks and short-term hazard assessment of large earthquakes usingcomplex networks: the case of the 2009 L'Aquila earthquake*. Non. Proc. Geophys. 23 (2016), pp. 241–256.
[14] A. Degani, M. Shafto, and L. Olson, *Using Canonical Correlation Analysis: Use of Composite Helio Graphs for Representing Multiple Patterns*, Diagram conference, Stanford, CA, 2006, p. 166.
[15] S. De Iaco, *A new space-time multivariate approach for environmental data analysis*. J. Appl. Stat 38(11) (2011), pp. 2471–2483.
[16] J. Dunnand, and G. Doeksen, *Canonical correlation analysis of selected demo graphic and health personnel variables*. S. J. Agricult. Econom. (1977), pp. 95–99.
[17] T. Foucart, *Multiple linear regression on canonical correlation variables*. Biom. J 41(5) (1999), pp. 559–572.
[18] F. Galton, *Natural inheritance*. Science 13(322) (1889), pp. 266–267.
[19] A. Gohary, and K. Hanzaee, *Personality traits as predictors of shopping motivations and behaviors: A canonical correlation analysis*. Arabian J. Bus 9 (2014), pp. 166–174.
[20] B. Gutenbergand, and C.F. Richter, *Frequency of earthquakes in California∗*. Seismol. Soc. Am. Bull 34 (1944), pp. 185–188.
[21] W. Härdle, and L. Simar, *Canonical correlation analysis*. J. Multiv. Anal., Springer, Berlin, Heidelberg Chp 14 (2007), pp. 321–330.
[22] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, *Canonical correlation analysis: An overview with application to learning methods*. Neur. Comput. 16 (2004), pp. 2639–2664.
[23] C. Herrera, F.A. Nava, and C. Lomnitz, *Time-dependent earthquake hazard evaluation in seismogenic systems using mixed Markov chains: An application to the Japan area*. Earth. Space. 58 (2006), pp. 973–979.
[24] S. Hong, X. Chen, L. Jin, and M. Xiong, *Canonical correlation analysis for RNA-seq co-expression networks*. Nuc. Acid Res. (NAR) 41(8) (2013), pp. e95.
[25] H. Hotelling, *Relation between two sets of variates*. Biom 28 (1936), pp. 321–377.
[26] K. Hsu, *Time series analysis of the interdependence among air pollutants*. Atmos. Environ 26(4) (1992), pp. 491–503.
[27] R. Iaci, T. Sriram, and X. Yin, *Multivariate association and dimension reduction: A generalization of canonical correlation analysis*. Biometrics 66(4) (2010), pp. 1107–1118.
[28] C. Lomnitz, *Global Tectonics and Earthquake Risk*, Elsevier Scientific Publishing Co, Amsterdam-London-New York, 1974.
[29] C. Lomnitz, and F. Nava, *The predictive power of seismic gaps*. Seismol. Soc. Am. Bull 73 (1983), pp. 1815–1824.
[30] S. Masters, and A. Wallston, *Canonical correlation reveals important relations between health locus of control, coping, affect and values*. J. Heal. Psych 10(5) (2005), pp. 720–731.
[31] F.A. Nava, C. Herrera, J. Frez, and E. Glowacka, *Seismic hazard evaluation using Markov chains: application to the Japan area*. Pure Appl. Geophys 162 (2005), pp. 1347–1366.
[32] F. Omori, *On the aftershocks of earthquakes*. J. Col. Science 7 (1894), pp. 111–120.

[33] B Orlecka-Sikora, S. Cielesta, and S. Lasocki, *Tracking the development of seismic fracture network from The Geysers geothermal field*. Acta Geophys. 67 (2019), pp. 341–350.
[34] C.A. Papaioannou, and B.C. Papazachos, *Time-independent and time-dependent seismic hazard in Greece based on seismogenic sources*. Seismol. Soc. Am. Bull 90 (2000), pp. 22–33.
[35] K. Pearson, *Mathematical contributions to the theory of evolution III, regression, heredity and panmixia*. Phil. Trans. 187 (1896), pp. 253–318.
[36] K. Pearson, *Mathematical contributions to the theory of evolution. On the reconstruction of the stature of prehistoric races*. Phil. Trans. (1898), pp. 169–244.
[37] C. Pertsinidou, G. Tsaklidis, E. Papadimitriou, and N. Limnios, *Application of hidden semi-Markov models for the seismic hazard assessment of the North and South Aegean Sea, Greece*. J. Appl. Stat 44(6) (2017), pp. 1064–1085.
[38] J. Theiler, and D. Prichard, *Constrained-realization Monte-Carlo method for hypothesis testing*. Phys. D 94(4) (1996), pp. 221–235.
[39] N. Theodoulidis, G. Cultrera, V. De Rubeis, F. Cara, A. Panou, M. Pagani, and P. Teves-Costa, *Correlation between damage distribution and ambient noise H/V spectral ratio: the SESAME project results*. Bull. Earth. Eng. 6 (2008), pp. 109–140.
[40] K. Todros, and A. Hero, *Measure Transformed Canonical Correlation Analysis with Application to Financial Data*, Conference: sensor array and multichannel signal processing workshop (SAM), IEEE 7th, 2012.
[41] T. Utsu, *A statistical study on the occurrence of aftershocks*. Geophy. Magaz. 30 (1961), pp. 521–605.
[42] T. Utsu, Y. Ogata, and R.S. Matsura, *The centenary of the Omori formula for a decay Law of aftershock activity*. J. Phys. Earth 43 (1995), pp. 1–33.
[43] I. Votsi, N. Limnios, G. Tsaklidis, and E. Papadimitriou, *Hidden Markov models revealing the stress field underlying the earthquake generation*. Phys. A 392 (2013), pp. 2868–2885.
[44] J. Walden, J.P. Smith, and R.V. Dackombe, *The use of simultaneous R- and Q-mode factor analysis as a tool for assisting interpretation of mineral magnetic data*. Math. Geol. 24(3) (1992), pp. 227–247.
[45] R. Wang, K. Jiang, J. Feldman, J. Bickeland, and H. Huang, *Interringgene association networks using sparse canonical correlation analysis*. Ann. Math. Statist 9(1) (2014), pp. 300–323.
[46] P. Wessel, W.H.F. Smith, R. Scharroo, J.F. Luis, and F. Wobbe, *Generic Mapping tools: improved version released*. EOS Trans. AGU 94(45) (2013), pp. 409–410.
[47] S. Wiemer, and M. Wyss, *Minimum magnitude of completeness in earthquake catalogs: examples from Alaska, the Western United States, and Japan*. Seismol. Soc. Am. Bull 90(4) (2000), pp. 859–869.
[48] G. Wu, F. Chen, D. Kang, X. Zhang, D. Marinazzo, and H. Chen, *Multiscale causal connectivity analysis by canonical correlation: theory and application to epileptic brain*. IEEE Trans. Biom. Eng. 58(11) (2011), pp. 3088–3096.
[49] G.H. Yu, and C.C. Huang, *A distribution free plotting position*. Stoch. Res. Risk As. (SERRA) 15 (2001), pp. 462–476.
[50] Y. Zhang, G. Zhou, J. Jin, X. Wang, and A. Cichocki, *Frequency recognition in ssvep-based BCI using multiset canonical correlation analysis*. Int. J. Neur. Syst. (IJNS) 24(4) (2014), pp. 1450013.