Automatic anomaly detection for swarm observations

BI Yaxin[1], CHRISTODOULOU Vyron[1], WILKIe George[1], ZHAO Guoze[2], NICHOLL Perter[1], HAN Bin[2], TANG Ji[2]

1. *School Computing, Faculty of Computing, Engineering and the Built Environment, Ulster University, Jordanstown Road, Newtownabbey, Co Antrim, UK*;
2. *Institute of Geology, China Earthquake Administration, Beijing* 100029*, China*

Abstract: The Swarm satellite mission was lunched on November 22, 2013, it is the first European Space Agency's constellation of three satellites, dedicated to monitoring geomagnetic field changes. The measurements delivered by the three satellite are very valuable for a range of applications, including earthquake prediction study. However, for more than 5 years, relatively little advancement has been achieved on establishing a systematic approach for detecting anomalies from the satellite measurements for predicting earthquakes. This paper presents the challenges of developing a pragmatic framework for automatic anomaly detection and highlights innovative features of functional components developed. Through a case study we demonstrate a functionality pipeline of the system in detecting anomalies, and present our solutions to coping with data sparsity and parameter tuning as well as insights into the differences between discovering seismic anomalies from periodic and non-periodic data observed by the Swarm satellites.

Key words: Anomaly detection; Swarm satellites; earthquake prediction study

Citation format: Bi Y X, Christodoulou V, Wilkie G, Zhao G Z, Nicholl P, Han B and Tang J. 2020. Automatic anomaly detection for swarm observations. Journal of Remote Sensing(Chinese). 24(S1): 173−182

1 INTRODUCTION

Approaches of monitoring earthquakes have evolved from conventional ground-based networks to space satellites. In the areas of seismology, geology and geophysics, scientists believe that the events leading up to earthquakes goes through a complex process and the process is somehow chaotic [1]. Understanding earthquakes requires a breakthrough from traditional approaches to utilizing advanced technology [2]. In fact, the seismology discipline has expanded the scope of earthquake study from conventional ground-based observations to space [3]. In particular, the fast advances of space-borne technology has paved a way to provide a range of measurements, including electromagnetic emission, total electron content (TEC), plasma density, powerful particle and particle temperature in the ionosphere as well as ground deformation, demonstrating numerous abnormal phenomena ahead of earthquakes [4]. However, their precise connection to the earthquakes has not yet been established.

It is commonly acknowledged that earthquake prediction is a long-standing issue across research communities. The literature suggests that an underlying of earthquake prediction study lies the way of capturing, analyzing and model anomalous phenomena and understanding the processes that cause those phenomena [5]. The precise recognition of the phenomena would provide evidence for developing risk assessments, safety measurements and better planning of emergency responses, thereby reducing economic losses and saving human lives. However, detecting seismic abnormal phenomena prior to earthquakes poses various challenges to scientists. It is extremely difficult to achieve accurate and reliable recognition of seismic phenomena in conjunction with developing models for estimating the occurrence time, precise location and magnitude of a potential earthquake. Except for a few earthquakes in recent history, for instance, the Haicheng Ms=7.3 earthquake occurred on the 4th February of 1975, almost all strong earthquakes have not been precisely predicted [6].

Realising the challenges of recognizing abnormal phenomena prior to earthquakes, the fundamental question has been posed since the beginning of earthquake study, that is: 'will earthquakes produce electromagnetic waves as they develop, prior to an event? '. Considerable amount of physic simulation work has been conducted [7], revealing when high temperature and pressure were gradually pushed on a rock, the rock started to fracture, meanwhile the frequency spectrum of electromagnetic radiation could be observed in bands of Hz to MHz. More findings were further found that the process of rock fracturing and frequency spectrum are dependent on the type of rocks. Although earthquakes do not occur in a scientifically controlled environment, these simulated experiments took into account the type of rock and tectonic characteristics of earthquake prone regions to a degree, thereby simulating the preparation behavior of earthquakes to some extent. There might still be a long way to go to establish a more secure proof of fractur-

Received: XXXX-XX-XX; **Accepted:** XXXX-XX-XX

FFirst and corresponding author biography: Yaxin Bi Yaxin,male,he interests in multiple supervised and unsupervised machine (deep) learning-based classification systems and ensemble methods in conjunction with the Dempter-Shafer (DS) theory of evidence; data analytics and decision making with uncertainty methods for satellite data exploitation with an emphasis on anomaly/change detection; and sentiment analytics for opinion mining and cyber-bullying detection

E−mail: y.bi@ulster.ac.uk

ing process under the ground, but at least these experiments revealed the existence of electromagnetic radiation when rocks fractured under laboratory conditions.

Electromagnetic radiation, resulting from ground rupture, could potentially be observed from the ground and space, but by contrast the observed electromagnetic variations may not necessarily come from the tremors in the ground. The study of ionospheric perturbation using the DEMETER satellite data demonstrated that solar activity could induce geomagnetic storms that cause electromagnetic changes. Thus, to detecting seismic anomalies, it is a precondition to exclude disturbance from other potential sources of anomalies. Magnetic activity indices can be used to measure solar activity, including the interaction of the solar wind with the magnetosphere, the ionosphere or the interaction between them. Two of the commonly used magnetic indices are the Dst (disturbance storm time) and Kp (Planetarische Kennziffer). The Dst index provides information about the solar activity by giving information about the strength of the earth's ring current [8]. The Kp index measures the daily variations in earth's electromagnetic field caused by solar radiation and its interaction with the magnetosphere, whereupon higher solar activity causes higher variations [9].

A vast amount of data from satellites for monitoring earthquakes is available and it is continually growing. Detecting anomalies from the acquired measurements requires a pragmatic approach to incorporating cut-edge data analytics technology to efficiently processing and understanding the collected data and making the technology of detecting seismic anomalies viable [11, 12, 13]. A large body of earthquake prediction work has been published in the literature, by comparison, relatively little is known about any automated analytics systems in place to provide daily services to allow scientists to have imminent analysis results from the space and ground-based observations other than taking into account historical observed data from ground-based networks. This work is an effort towards this goal, incorporating advanced data analytics underpinned with machine learning and artificial intelligence to developing a data analytics system, which can be used to automatically detect abnormal changes recorded in the Swarm satellite data. The analysis methods will be capable of learning anomaly patterns and avoid instinct sense when analysing data and recognise anomalies, achieving more objective results.

This system is able to gradually collect various anomalous cases from the Swarm data that will be complemented with CSELF (Control Source Extremely Low Frequency) and CSES (Chinese Seismo-Electromagnetic Satellite) data, looking at the details of each anomaly in order to establish normal and abnormal models for earthquake study. Through studying a large number of earthquake cases and associated electromagnetic measurements distilled from the satellite data, the system will incorporate more data to the existing historical observations by ground sensors to learn, and then adapt itself to form synergistic models for discovering seismic precursors. As a result, the system would be able to overcome false positive errors to the most extent, establishing a correlation between the anomalies and the earthquakes.

2 RELATED WORK

The development of advanced data analytics for seismic anomaly detection starts with understanding how anomalous phenomena have been captured and processed in the seismology domain, particularly understanding the process of manual inspection of anomalous phenomena in the development cycle of earthquakes. An appropriate understanding will form a base of transforming the manual analysis ability into an automatic process and then allow the development of viable data analytics for detecting seismic anomalies. The section reviews the existing studies from two perspectives; methods for capturing abnormal changes before and after earthquakes and applications of advanced data analysis technology to ascertaining connections between anomalies and the earthquakes.

In [3] the authors collected all the 6.5 year's worth of daytime and night-time data from the DEMETER satellite and conducted analysis on ionospheric perturbations. They found there was no significant electromagnetic variation in the daytime data, the authors claimed because of the existence of more disturbance in the ionosphere, it is more difficult to detect small electromagnetic variations. As opposed to this, the daytime data revealed the electromagnetic intensity decreasing before the earthquakes. The study was undertaken further on the length of observed data, they concluded that the decrease was only observable over a long period of time, that means daytime measurements were not suitable for studying short-time earthquake prediction.

Studying electromagnetic variations using data acquired from ground and the DEMETER satellite was carried out in [16], their results indicated that anomalies were found from 1 to 5 days prior to the earthquake happening. In [17], the authors conducted an investigation on the effect of Ion density in ionosphere, they collected data from over 20,000 earthquakes, in which the distance of the DEMETER satellite footprint from these epicentres is in the range from 750 to 2000km, their results revealed that the anomalies appeared from 3 to 10 days before the earthquakes. Using data from the same satellite that was around 1500km and less far from these epicenters in [18], the authors discovered electromagnetic anomalies appearing from 5 to 8 days ahead of the earthquakes. The work published in [19] affirmed the correlation between earthquakes and electromagnetic anomalies, the authors noted geomagnetic disturbances could be observed even from a distance of 2000 to 4000km far away to the earthquake epicentres.

Using the interquartile (IQR) to examine the relation between seismic precursors and thermal electron content (TEC) measurements has been carried out in [13], the authors selected a median range as a threshold and found that there is a correlation in 16 out of 20 earthquakes of TEC anomalies appearing within 5 days prior to the earthquakes. A similar method was also used in a recent study reported in [14], the authors claimed to discover anomalies even 130 days prior to an earthquake in Mexico with MS8.2.

A statistical analysis of 650 earthquakes covered by the DEMETER satellite's data was reported in [15], the analysis was focused on interpreting a 3-year distribution of electromagnetic waves in the vicinity of these events. The authors also gathered another series of normal data acquired from the DEMETER satellite's measurements in a period of non-seismic activity. After comparing both datasets, the study shows that there is a statistically significant variation in night time intensity of 4—6 dB at ~4 hours before the earthquakes.

In the last few years there has been a big increase in the use of reconstruction based techniques for big data analysis [11]. Positive gains include the quick processing of large amounts of data, no parameter tuning and the ability to automatically model non-linear processes. Another advantage of this approach is that they are able to learn and also adapt to any changes in the underlying distribution of the data. When test data are presented to those algorithms it is viable to infer complex connections and adjust its reconstruction error to minimize the distance between the test and the expected results. However, their applications have not been so widespread because they add an additional layer of complexity to an already complicated problem, when ideally we want to understand the underlying mechanism and the reasons that electromagnetic anomalies might act as seismic precursors.

Most of the above studies treated the data analysis as a linear and parametric problem. A further analysis of the generative process was carried out in [20]. The authors experimented with state-

of-the-art algorithms including Decision Trees, Bagging, Random Forests and other non-parametric methods, demonstrating their effectiveness in detecting precursory phenomena that appeared from two weeks (the earliest) to four weeks before the main seismic event. A fuzzy logic system that uses the variance of signals obtained by the application of an Empirical Wavelet Transform on Ultra Low Frequency measurements was used in [21] as the anomaly detection method. The authors noted that the perturbation of geomagnetic field can be detected in a distance of 563km to 1473km from the epicentre.

Reviewing the existing studies illustrate that the process of detecting seismic electromagnetic anomalies before earthquakes involves a high degree of selectivity and parametrization [23], six parameters play a varied role in detecting anomalies from electromagnetic time series data: (a) magnitude, (b) time window, (c) spatial window, (d) confidence or accuracy of the prediction, (e) probability of a seismic event (f) earthquake prone area, an anomaly and distance of which the observation was made. These are some of the spatial and time components that may hinder the effect of seismic anomaly detection and have to be considered in the interpretation of results.

3 DESIGN OF AN AUTOMATIC ANOMALY DETECTION SYSTEM

Detecting seismic anomalies by our system involves data preparation and anomaly detection two major stages. Given an earthquake prone area under investigation, the orbit coverage of the satellites is not always able to provide sufficient data for the area due to distances between two satellite orbits and the length of revisit days of satellites. These factors therefore cause the coverage of the satellite data over the area not even or imbalanced. The consequence is that some part of the area might have more data, whereas others have less, or even have no data available at all. These factors result in sparsity of satellite data and artificial anomalies while generating continuous time series data for a study area, which have to be overcome. This system comprises the four pre-processing methods developed to alleviate the issue of data sparsity and minimize artificial anomalies to maximum extent.

3.1 FOUR PRE-PROCESSING METHODS

The data analysis on an area under investigation is conducted on a set of grids, into which the area is evenly divided. Given a predefined period of time, each grid is used to generate a time series dataset based on the time of that the satellites fly over the grid. However the orbits of the satellites could not fall in the grid, thus the grid could not have satellite data, that will result in missing data between the revisit cycles of the satellites. In this study, we have developed four pre-processing methods to ensure the quality of time series datasets on each grid.

Mean values: this method gathers all data point within a grid and takes a mean value for each date over all measurements.

Minimum values (First point): this method can be thought as a local mean within each grid. It first calculates what is the smallest number of readings per each grid, that is the day with the smallest overpass. It then calculates the mean using the minimum number, which might be different for each grid, separately.

Median values (Mode): This function takes the median for each date with observations. Because in most cases there are 52 to 53 readings for a full overpass the median and the mean have similar results.

User defined mean values: This function gives the user the ability to select how many data points they want to aggregate per day and evaluate artificial anomalies over the time series visually. It is a form of global mean that applies to every grid. However, if this mean is higher than the available measurements within a grid's date then this smaller number, which represents the smallest amount of readings available in that date, is selected instead of the user selected mean. The default value is the five first samples.

3.2 DETECTION ALGORITHMS

Outcome of each of the pre-processing functions will be nine sets of time series data. These datasets are fed into the detection algorithms implemented by the system, composing of four detection algorithms, which are detailed below.

The algorithm CE is a parametric statistics-based method, which involves the combination of the Cumulative Sum (CUSUM) and Exponentially Weighted Moving Average (EWMA) [24].

CUSUM formulates a sequence of partial sums for a given process and monitors changes in the mean of a process. By contrast the EWMA takes present and past observations into account by a memory element, rendering that larger values of the memory element mean that past values count for less, while smaller values mean that they count for more. As being a geometric moving average, this motivates to use EWMA to replace the mean of a process formulated by CUSUM, thereby resolving false positives produced by CUSUM and EWMA, respectively, in detecting complicate anomalies. The algorithm has been evaluated on benchmark datasets, the F-Score results demonstrate its promising performance in detecting anomalies from time series data

The fuzzy inspired algorithm consists of functional components that are aggregated in a serialized fashion to achieve anomaly detection in electromagnetic sequential datasets [12]. Each of the component methods adds an element towards anomaly detection, i. e. a smoothing filter removes any unwanted noise, an automated peak finding with Fast Fourier Transformation and correlation reduces the dimensionality of the signal, as well as a fuzzy inference system encodes the signal before the final comparison and its respective output. In practice, the peak distance has to be selected in direct proportion to the size of the input dataset. The peak distance parameter affects the selectivity and the final number of peaks after the pre-processing signal. The evaluation results on benchmark data and SWARM satellite data demonstrate its effectiveness for the detection of anomalies in electromagnetic time series datasets.

The SAX algorithm was developed on the basis of Symbolic Aggregate approXimation (SAX) [25]. SAX allows a time series of arbitrary length to be reduced to a string with variable length, and uses an intermediate representation between the raw time series and the symbolic strings. The underlying idea of the algorithm is to transform time series data into the Piecewise Aggregate Approximation (PAA) representation and then symbolizes the PAA representation into a discrete string. The transformed data allows distance measures to be defined, in which the lower bound corresponds to a distance measure defined on the original series. The latter feature is particularly effective because it allows certain data analytics algorithms to efficiently manipulate the symbolic representation, and produce identical results as operating on the original data.

The fundamental part of the SAX algorithm is to use the average value of the time series for each segment, regardless the situation where two segments have very close averages but behave differently. For such a situation, instead of generating two averages for two segments, the two segments should be quantized into the same symbol. Specifically, a segment with an increasing trend can be mapped into the same bin as another segment with a decreasing trend if their respective means are close. Based on this idea [26], the authors developed a new representation, called 1D-SAX, which incorporates the SAX representation with the trend of the time se-

ries on each segment. The 1D-SAX consists of three main steps: dividing a time series into segments with a defined length; computing the linear regression of the time series on each segment; and quantizing these regressions into a symbol from an alphabet of size. In practice, the PAA and the slope in 1D-SAX have to be based on values of the power of two, and the PAA has to be larger than the slope.

3.3 GRAPHIC USER INTERFACE (GUI) OF THE SYSTEM

Fig. 1 presents the GUI for the system consists of three sections. Section 1 is used to define areas under investigation. Section 2 integrates the functions of the four pre-processing and heatmap, which will be used to generate continuous time series data and visualise the intensity distribution of electromagnetic waves over an area defined in Section 1. Section 3 is composed of the four algorithms described above with parameter input.

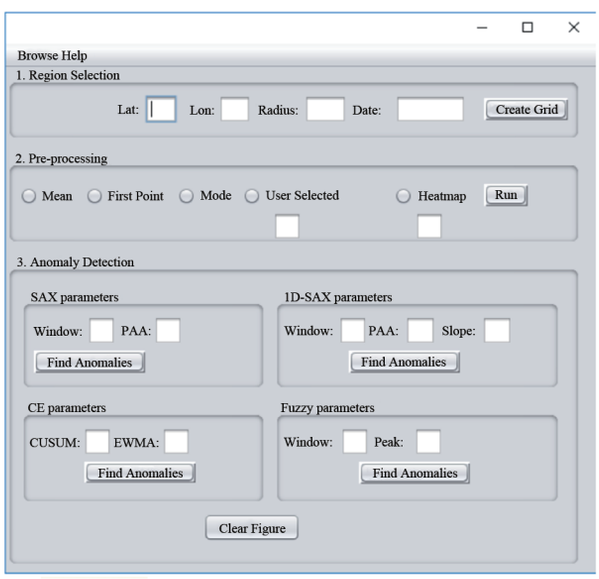


Fig. 1 The graphical user interface of the system

The detection process starts with browsing the local drive to upload a satellite data file, enters the latitude and longitude of the epicentre of an earthquake, the date when the earthquake occurred and the length of radius for the earthquake prone region, the system will then extract the satellite data from the file in accordance with the entered values. As the default setting, the study area is evenly divided to a set of nine grids when 'Create Grid' button is clicked, thus nine respective time series datasets can be simultaneously generated using the four pre-processing functions, corresponding to the nine even grids defined with respect to the radius.

Once one pre-processing function is chosen and run, 9 time series datasets will be generated and displayed over each of the grids of the area. Additionally, the time series data over the grids can be rendered by the HeatMap, showing the distribution of the averaged intensities of electromagnetic emission over the area of interest. These functions provide users with the descriptive summary of the observed data by the Swarm satellites and allow users to have a perception on the emission and possible variations before anomaly detection over them being conducted.

In Section 3 of Fig. 1, each algorithm has two or three input fields, allowing users to input parameter values based on the individual requirements. Different values chosen could have direct impact on the performance of the algorithms. The following briefly interprets what role of the parameters play in each of the algorithms:

For CE the range of λ values has a direct relation to sizes of input datasets because it denotes the amounts of historical values that the algorithm is going to take into account.

For the Fuzzy Shape-based method, the peak distance has to be selected in direct proportion to the sizes of input datasets. The peak distance parameter affects the selectivity and the final number of peaks, which describe the reduced, after the pre-processing signal.

For SAX, all parameters have to be selected in proportion to the sizes of input datasets but there is no other requirement.

1D-SAX can accept only parameters on the power of two. The conversion stage to symbols is based on a binary conversion. The PAA has to be of a larger value than the slope.

Parameters have to be tuned through experiments in order to obtain an optimal one. That means selecting parameter values is dependent on individual applications, one set of parameter values only can be regarded a reference, which may not be directly used to other applications. Table 1 presents a range of possible values for the parameter taking on, which are obtained through the experiments on 20 benchmark datasets and preliminary experiments on the Swarm satellite data. These suggested values have been used for a range of analyses in detecting seismic anomaly exercise by the algorithms, receiving acceptable performance.

Table 1 Ranges of parameters for the anomaly detection algorithms

	CE (K, λ)	Fuzzy (P, w)	SAX (P, w)	1D-SAX (p, s, w)
range	0.1-10.0,1-N	1-10,1-10	1-10,1-10	2-8,2-8,2-8

Given a set of optimal parameter values, a further optimization should be conducted as applying four detection algorithms onto four sets of time series data generated from the pre-processing functions will produce 4x4 analysis results. At the current stage, the optimization of selecting one algorithm is manual work, an automatic and viable method for this is under development.

4 A CASE STUDY

The case study demonstrates how the system will be used to detect anomalies over the area under investigation. As illustrated in Fig. 1, the detection process will go through three sections. For this case study, we select an earthquake occurred in Peru with 6.8 magnitude. The occurrence of the earthquake was on 24 August 2014 and its epicentre is located at 14.598° S latitude and 73.571° W longitude. Before entering these information onto the input fields in the GUI, we need to acquire about one year satellite data into three files, namely Swarm A, B and C, and then upload one into the system each time. We also use the epicentre and 500 kilometers as the centre and radius, respectively, to draw a square as the area under investigation, it is evenly divided into a set of 9 grids.

On the basis of the entered parameters in Section 1 of the GUI, we can select a pre-processing function in Section 2, such as 'Mean', then nine time series datasets are generated and rendered in Fig. 2, which represents the mean values of every five days in each grid for 10 months. The red dash line in each of the grids marks the date when the earthquake occurred. By a visual inspection, we see more variations of intensities in the second and third columns com-

pared with the first column, in which anomalies could be thought and likely detected. The intensities of electromagnetic radiation over the grids can be visualized by the heatmap as shown in Fig. 3. The values rendered in the heatmap represent the average of the number of days, a value 5 entered in input fields means, for instance, the mean of 5 days.

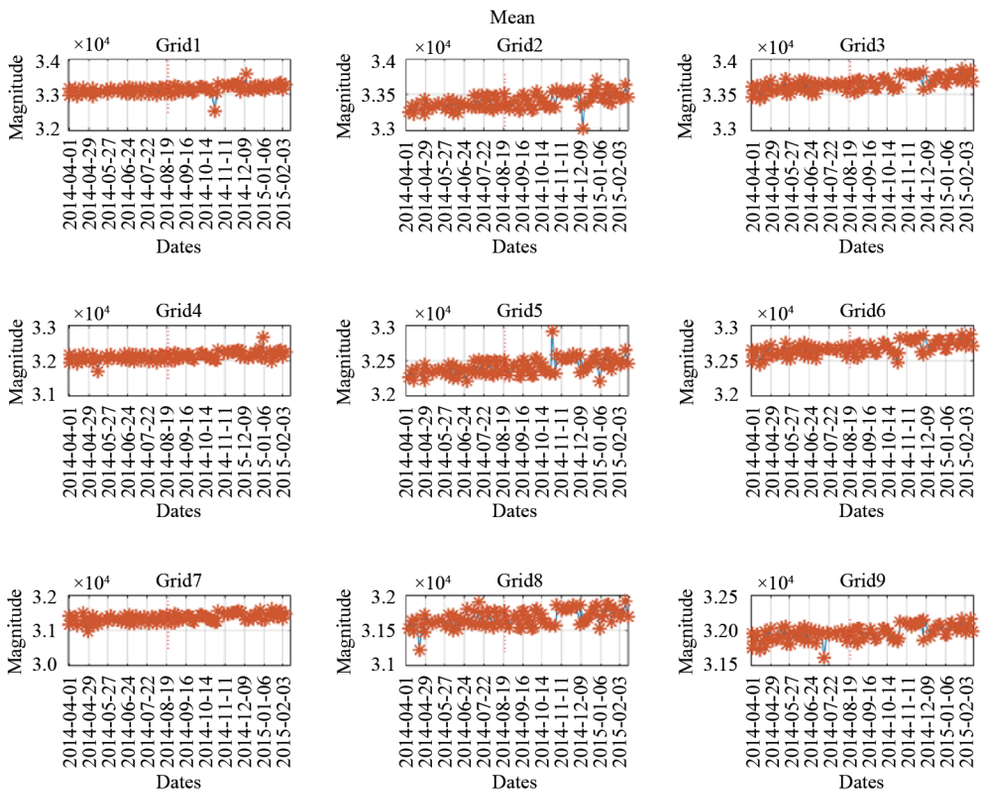


Fig. 2 Mean visualization

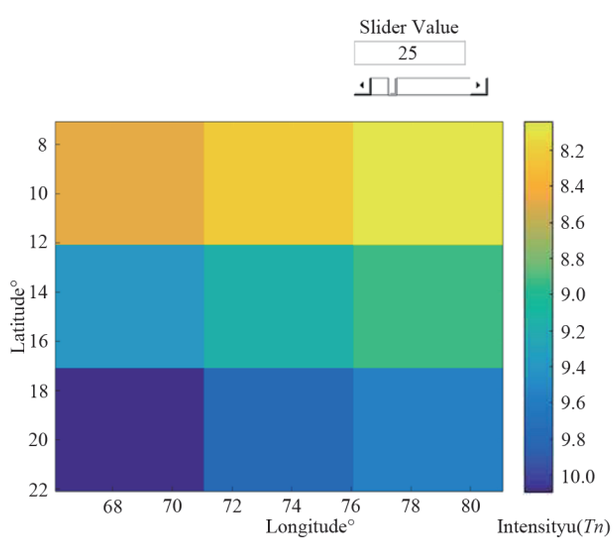


Fig. 3 The intensity heatmap of the area of the Peru earthquake

Once the time series data generated, it is ready to conduct anomaly detection on the data given in Fig. 2 using the algorithms shown in Fig. 1. A set of selected parameters for these algorithms are given in Table 2 below.

Selecting the algorithm CUSUM-EWMA (CE) with the input parameters to analyse the time series data, Fig. 4 presents the detecting results that are overlaid on the time series data in each grid and marked with red lines that indicate the anomalies detected by the algorithm. Looking at these data in the 9 grids only, we can note some variations of intensities, particularly in the second and third columns, but it is not straightforward to recognize where there are abnormal changes before and after the earthquake.

Table 2 Parameter values for the algorithms

	CE (K, λ)	Fuzzy (P, w)	SAX (P, w)	1D-SAX (p, s, w)
value	(1, 6)	(4, 2)	(6, 4)	(8,4,2)

As seen the red lines appear in each grid, but in grids 1, 2, 3, 5 and 6, they are found after the occurrence of the earthquake, whereas the red lines in grids 7, 8 and 9 found before the earthquake. Exceptionally in grid 4, the anomalies appear both before and after the earthquake. Linking these anomalies to the time of the earthquake occurring, they appear in grids 4, 7 and 8 about 4 months ahead of the earthquake, and about 1 month in grid 9. As opposed to this, the anomaly appears in grid 1 is about 3 months later than the earthquake, but for the others, they appear about 4 months or even more after the earthquake.

Compared with the anomalies detected on the same data using the algorithm, they are not consistent with what inspected by a visual inspection approach. The visual inspection mainly seeks sharp fluctuations along the mean of a process, that is where the strong variations of intensities occur, and then perceives them part as anomalies. However, when the CE algorithm detecting anomalies, it assumes that the generative process of means obeys a Gaussian distribution. Through adjusting a pair parameters (K, λ), the algo-

rithm calculates the CUSUM and EWMA statistics in addition to the number of past values in the process, finally ascertains whether there are anomalies within the process. Apparently the latter is be much more robust than the former, and the process of determining anomalies is traceable and explainable, however a number of anomalies detected could not be perceived as anomalies manually.

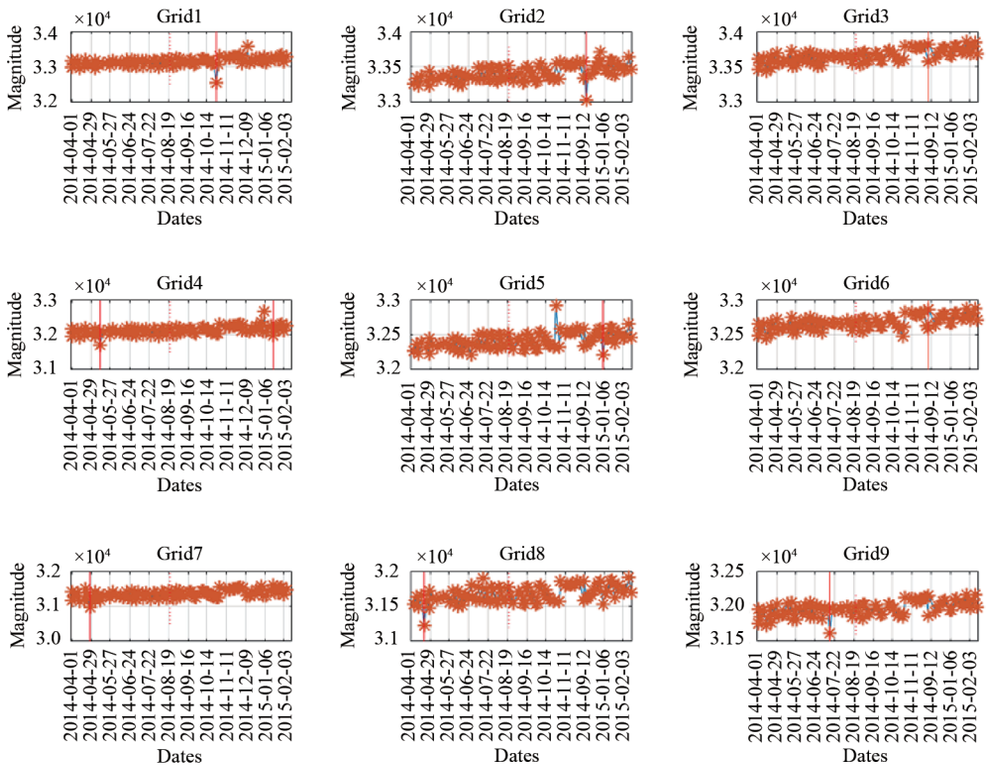
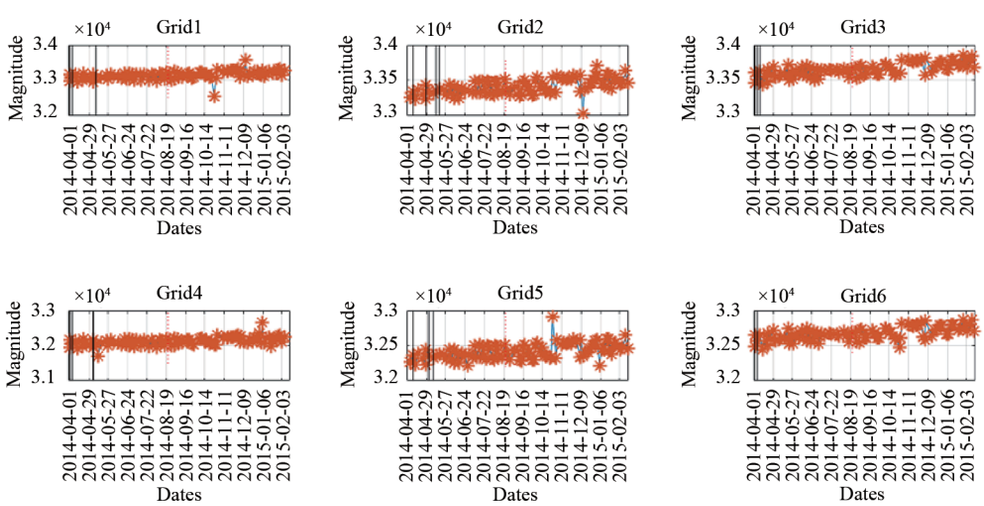


Fig. 4 Mean pre-processing with Algorithm CE

Following the same process on the same data above along with the parameters given in Table 2, the analyses have been conducted using the Algorithms of Fuzzy, SAX and 1D-SAX, respectively. The analyzing results are overlaid on the time series data as presented in Fig. 5, 6 and 7, respectively, in which the detected anomalies are marked with black, green and blue lines. It is worth noting that the anomalies detected by Fuzzy and 1D-SAX in Figs 5 and 7 all appear before the earthquake, and more than one anomaly were detected in each grid. The times of anomaly appearing against the occurrence of the earthquake are varied, they are from 1 to 4 months before the earthquake occurring.

Examining the times of the anomaly appearing in detail, the anomalies discovered by the algorithm 1D-SAX are more closer to when the earthquake occurred. For instance, in grid 5 of Fig. 7, the anomaly detected in July is about one month away from the time of the earthquake occurring. On the other hand, the distribution of the anomalies across these grids in Fig. 5 and 7 are not even, no regular pattern can be found. As opposed to Fig. 5 and 7, there are one or two anomalies found in the grids, except grids 4 and 7 in Fig. 6, all anomalies are found before the earthquake, about 4 months ahead of the earthquake.



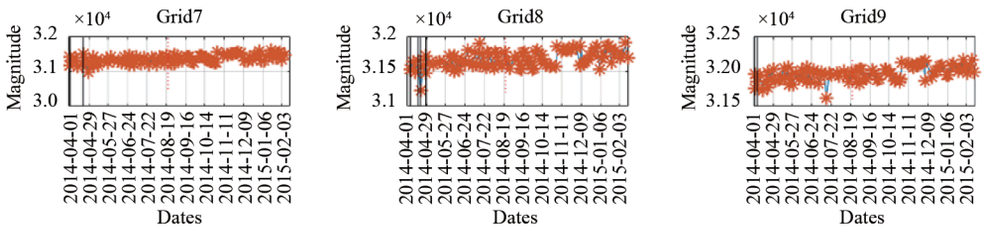


Fig. 5 Mean pre-processing with fuzzy inspired algorithm

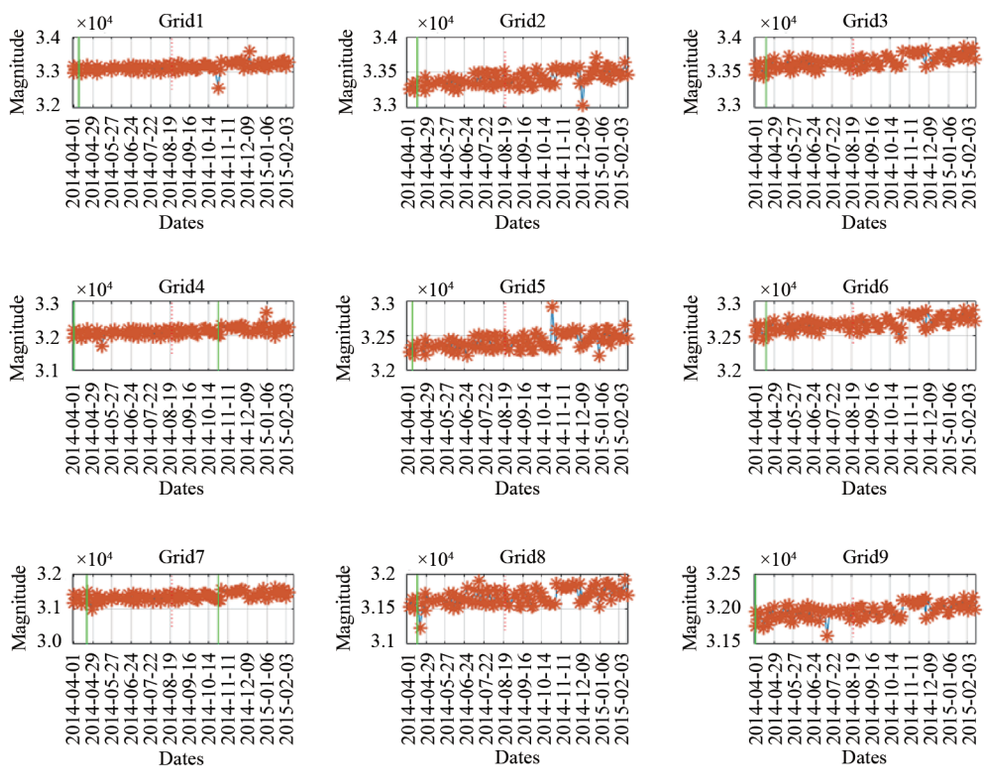
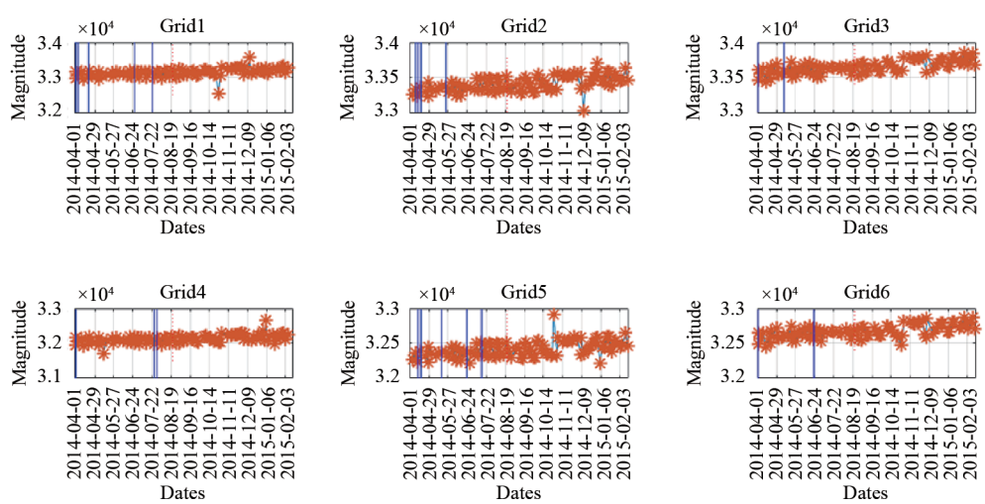


Fig. 6 Mean pre-processing with Algorithm SAX

The preceding figures demonstrate the process of applying the four algorithms to detect anomalies on the time series data produced by the Mean pre-processing function, as well as the insights into the detected results, the remaining step is to examine how a single algorithm behave over the datasets generated by four different pre-processing functions.



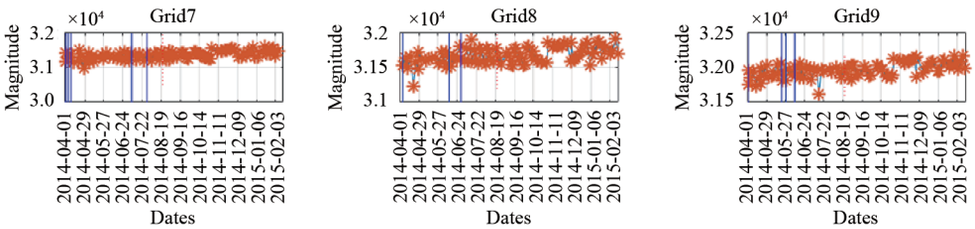


Fig. 7 Mean pre-processing with Algorithm 1D-SAX

Fig. 8 puts together the four detecting results by the CE algorithm on the four time series data produced by the 'Mean', 'First point', 'Mode' and 'User selected' pre-processing functions, respectively. It can be observed the resulting anomalies appear at different times, before and after the date of the earthquake occurring. There are more anomalies detected on the data derived by Mode' and 'User selected' compared with those Mean' and 'First point', but they behave quite differently, and also it is not straightforward to identify regular patterns from the figure. The similar results have been obtained by the algorithms of Fuzzy, SAX and 1D-SAX over the same data generated by the four pre-processing functions.

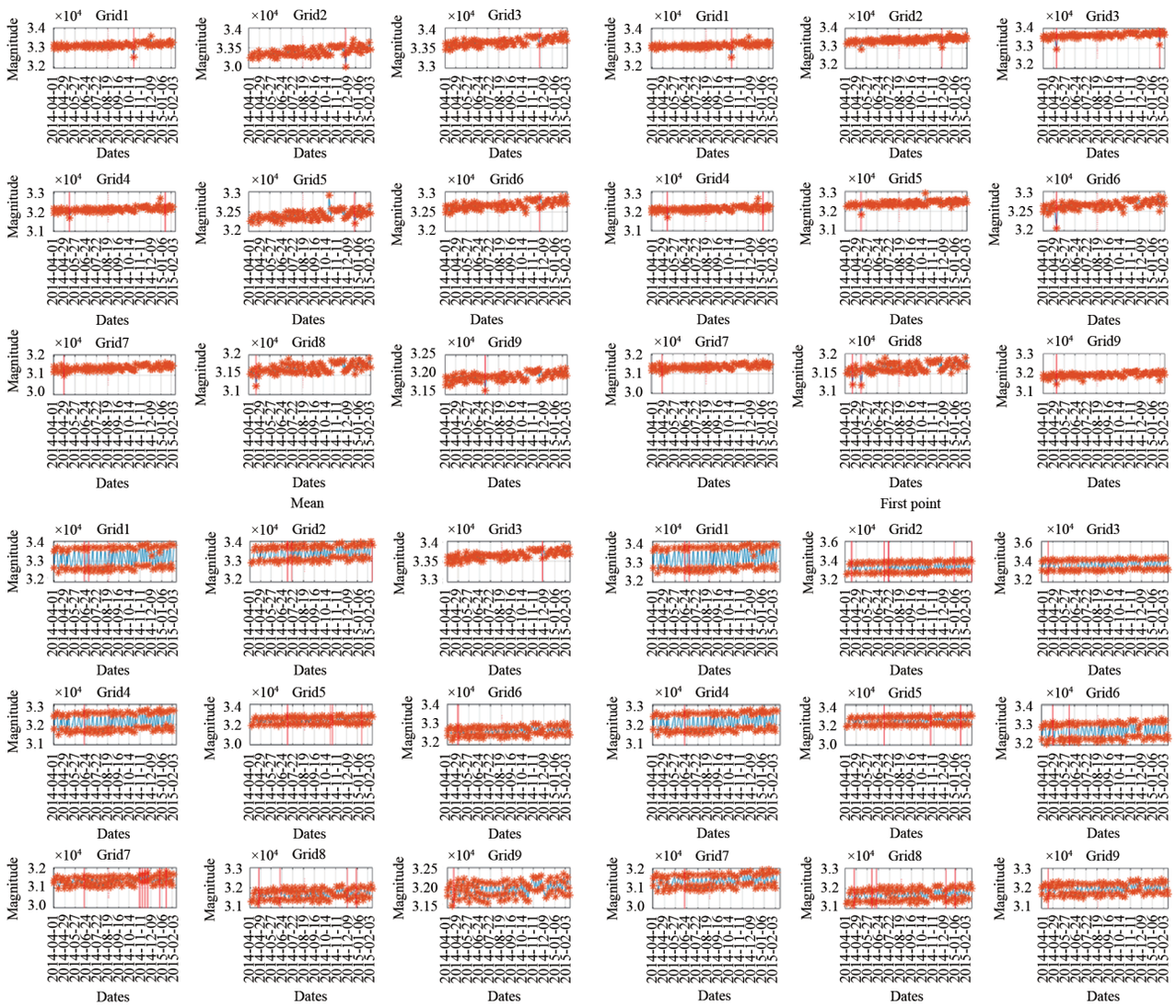


Fig. 8 Algorithm CE working on Mean, First Point, Mode and User selected point pre-processing functions

5 DISCUSSIONS

Fig. 4—Fig. 7 present the analysis effectiveness of the algorithms developed by ourselves and the comparative performance of two similar algorithms SAX and 1D-SAX. From these figures, it is evident that these algorithms have the different ability of detecting

anomalies, i. e. Fuzzy and 1D-SAX detecting more anomalies, whereas CE and SAX detecting less, which are more cautious. The differences between them could come from different characteristics of these algorithms. The first is that these algorithms have the different representation of input time series data. The CUSUM-EWMA (CE) does not have any constraints on the representation of input data, but it treats input data in the form a generative process that follows the Gaussian distribution. The Fuzzy-Inspired method transforms time series data by a fuzzy function defined on the basis of the shape of the data. SAX and 1D-SAX require a symbolic approximation on time series data and both are considered state-of-the art algorithms.

The second is that these algorithms require different parameters. Table 1 shows the range values of the parameters used. The range values provided for the algorithms are based on extensive experimental study on real and benchmark data [12, 25]. The Fuzzy and SAX algorithms coincidentally have a similar function range, that does not mean they hence have the same parameter tuning procedure. According to [26], the best parameters for 1D-SAX are for slope 4 and PAA 2. The word length was in-line to the previous symbolic representational methods at 4. In the algorithm the alphabet is hard-coded to the value of 3 based on the best experimental results suggested in [25], all of which have been considered in this case study.

The suitability of the algorithms to the data is a pressing matter. As the Swarm satellites fly along obits, due to the latitudinal difference on the orbits, this results in the periodic oscillation in observations. Specifically, when the satellites approach the equator, i.e. the latitude is close to zero degree, the magnetic field is getting the weakest, whereas approaching the poles, i.e. the latitude is close to ±90 degree, the magnetic field is getting the strongest. However, its footprint has a varying length of gaps between observations on two consecutive obits and the way of generating time series data does not take a periodic feature into consideration. From this perspective the Swarm satellite data is regarded as non-periodic. In the literature, most algorithms consider data as a periodic one, which is something that is easy to address with a predefined window for the symbolic approximation. However, our system has the ability to handle both periodic and nonperiodic situations, that is CE, FSB and 1D-SAX are able to deal with both periodic and nonperiodic data while SAX addresses only periodic data.

The above aspects are possible factors that might make these algorithms to perform differently in detecting anomalies from the Swarm satellite data. The performance of the algorithms have been evaluated against benchmark data and measured by a novel metric, denoted by R, which was detailed in [27]. The distinguishing feature of metric R is to take into account the subsequence length in addition to the anomalous location predicted by the algorithms and the true one.

The cross-comparison on the four pre-processing functions by the algorithm of CE is presented in Figures 8. As the time series data are identical for those algorithms, we confine the attention on Figure 8 only, examining the effectiveness and the relation of preprocessing functions. At first glance, the data generated by 'Mean' is similar to one by 'First point', where except grids 2 and 5, other grids across two methods correspondingly have very similar shape. Moreover, the data produced by 'Mode' is also similar to one by 'User selected' function in some aspects, for instance, both data contains periodic oscillation in addition to grids 1, 4, 5 and 8 across two methods.

On the other hand, when examining the anomalies detected over these grids, it can be postulated that the former two groups of the time series data contains less changes than the latter two groups of the data, which is easier to be perceived by visual inspection. For example, the maximum number of anomalies detected in the former groups of the data is 2, by contrast, the maximum number is 7 in the other two groups of the data. However, this postulation could not be supported with the results produced by the other three algorithms.

6 CONCLUSION

At the heart of this paper is an illustration of the system developed for detecting seismic anomalies in data observed by the Swarm satellites for a particular area under investigation. From this illustration we describe the functional components and demonstrate an automatic process of detecting abnormal changes within the data, from pre-processing, symbolic representation of data to finally detecting seismic anomalies, which is entirely inspired and guided by the principle of big data analytics. Each part involves a considerable amount of studies, dealing with the challenges faced by the state-of-the-art studies in the related multi-disciplines and optimization of integrating them together in order to achieve better detecting performance of the system.

Through running a case study, we present analysis results and comparisons for both, across the four detecting algorithms and the four pre-processing functions. This case poses two pressing issues to be resolved in the next phase.

As seen, an area under investigation is defined around an epicentre, choosing a radius is based on the Dobrovolsky's principle in [29], and distance from the centre of one grid to another is typically 300 kilometers, but the data distribution in one grid is quite different from another as shown in Figures 8, which are caused by data sparsity from the satellites. These differences will directly affect the performance of the algorithms. Thus, it is imperative to develop effective interpolation methods to accurately represent electromagnetic waves in the vicinity.

The case study illustrates complexity of parameter tuning and selection of these algorithms, which makes the system quite difficult to use in practice. Optimizing combinations of the parameters needs to be automated. Meanwhile, we will investigate whether the current more complex methods with less parameter tuning are appropriate for anomaly detection and how they perform in a constrained real-world environment, such as for Swarm data analysis.

The work undertaken is appealing for scientists as it works towards transforming the satellite observations into knowledge for understanding the process of earthquake detection. The next step of work will be integrated into the ground-based data platform of CSELF electromagnetic observation network to establish a viable three-dimensional data analytics framework, a disruptive technology, for studying earthquake prediction.

REFERENCES

Fraser-Smith A.C., Bernardi A., McGill P.R., Ladd M.E., Helliwell R. A., VillardJr. , O.G., 1990. Low-frequency magnetic field measurements near the epicenter of the Ms 7.1 Loma Prieta earthquake. Geophys. Res. Lett. 17, 1465-1468.

Bleier T, Freund F. 2005. Impending earthquakes have been sending US warning signals and people are starting to listen. IEEE Spectrum INT, 3:3-7.

Piša D, Němec F, Santolik O, et al. 2013. Additional attenuation of natural VLF electromagnetic waves observed by the DEMETER spacecraft resulting from preseismic activity. J Geophys Res, 118: 5286-5295.

MasashiHayakawa, alet, 2010, Current status of seismo-electromagnetics for short-term earthquake prediction, Geomatics, Natural Hazards and Risk, 1:2,115-155

Guoze Zhao, Yaxin Bi, alet, 2015. Advances in alternating electromagnetic field data processing for earthquake monitoring in China,

SCIENCE CHINA, Earth Sciences Vol.58 No.2: 172-182.

Chen C.S., Chen C.C., 2000. Magnetotelluric soundings of the source area of the 1999 Chi-Chi earthquake in Taiwan: evidence of fluids at the hypocenter. Terr. Atmos. Ocean. Sci. 11, 679-688.

F. Freund, "Pre-earthquake signals: Underlying physical processes," Journal of Asian Earth Sciences, vol. 41, no. 4-5, pp. 383-400, 2011.

SugiuraM., Hourly values of equatorial Dst for the IGY. 1963.

A. J. Dessler and J. Fejer, "Interpretation of kp index and m-region geomagnetic storms," Planetary and Space Science, vol. 11, no. 5, pp. 505-511, 1963.

Xiangzeng Kong, Yaxin Bi, and GlassDavid H.. Detecting Seismic Anomalies in Outgoing Long-Wave Radiation Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 8 (2): 649 - 660, 2015

Kong X., Bi, GlassY. &, D. Detecting Anomalies in Sequential Data Augmented with New Features. Artificial Intelligence Review, 2019

ChristodoulouVyron, Yaxin Bi, George Wilkie and Guoze Zhao. A Fuzzy Shape-Based Anomaly Detection and its Application to Electromagnetic Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 11(9): 3366-3379, 2018.

LiuJ. Y., ChuoY., ShanS., TsaiY., ChenY., PulinetsS., and YuS., "Pre-earthquake ionospheric anomalies registered by continuous gps tec measurements,vol". 22, no. 5, pp. 1585-1593, 2004.

Marchetti and MD.. Akhoondzadeh, *Analysis of Swarm satellites data showing seismo-ionospheric anomalies around the time of the strong Mexico (Mw= 8*.2 earthquake of 08 September 2017. Elsevier, 2018.

N ˇ emecF., ParrotM., SantolıkO., RodgerC., RycroftM., HayoshM., ShklyarD., and DemekhovA., "Survey of magnetospheric line radiation events observed by the demeter spacecraft," Journal of Geophysical Research: PhysicsSpace, vol. 114, no. A5, 2009.

AkhoondzadehM., ParrotM., and SaradjianM., "Electron and ion density variations before strong earthquakes (m¿ 6.0

) using demeter and gps data," Natural Hazards and Earth System Sciences, vol. 10, no. 1, pp. 7-18, 2010.

M. Li and M. Parrot, "Statistical analysis of an ionospheric parameter as a base for earthquake prediction," Journal of Geophysical Research: Space Physics, vol. 118, no. 6, pp. 3731-3739, 2013.

BhattacharyaS., SarkarS., GwalA., and ParrotM., "Observations of ulf/elf anomalies detected by demeter satellite prior to earthquakes," 91.30. Px, 2007.

HaoY. Q., XiaoZ., and ZhangD. H., "Teleseismic magnetic effects (TMDs) of 2011 earthquakeTohoku," Journal of Geophysical Research: Space Physics, vol. 118, no. 6, pp. 3914-3923, 2013.

AkhoondzadehM., tree"Decision, bagging and random forest methods detect tec seismo-ionospheric anomalies around the time of the chile,(mw= 8.8) earthquake of 27 february 2010,

Advances in Space Research, vol. 57, no. 12, pp. 2464-2469, 2016.

AlegriaO. C., Valtierra-RodriguezM., Amezquita-SanchezJ. P., Millan-AlmarazJ. R., RodriguezL. M., MoctezumaA. M., Dominguez-GonzalezA., and Cruz-AbeyroJ. A., "Empirical wavelet transform-based detection of anomalies in ulf geomagnetic signals associated to seismic events with a fuzzy logic-based system for automatic diagnosis," in Wavelet transform and some of its real-world applications, InTech, 2015.

XiaoG., Yuan-ShengZ., Mei-JiaoZ., Wen-RongS., and Cong-XinW., "Variation characteristics of olr for the wenchuan earthquake," Chinese Journal of Geophysics, vol. 53, no. 6, pp. 980-988, 2010

C. R. Allen, "Responsibilities in earthquake prediction: To the seismological society of america, delivered in edmonton, alberta, may 12, 1976," Bulletin of the Seismological Society of America, vol. 66, no. 6, pp. 2069-2074, 1976.

Christodoulou V, Bi Y. A combination of CUSUM-EWMA for Anomaly Detection in time series data. In Data Science and Advanced Analytics (DSAA). 2015;2015(36678):1-8

Keogh E, Lin J, Fu A. HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. In: Proceedings of the Fifth IEEE International Conference on Data Mining. IEEE Computer Society; 2005. p. 226-233.

Malinowski S, Guyet T, Quiniou R, Tavenard R. 1D SAX: A symbolic representation for time series. In: International Symposium on Intelligent Data Analysis, Oct 17, Heidelberg. Berlin: Springer; 2013. p. 273-284.

Christodoulou V, Bi Y, Zhao GA. Fuzzy Inspired Approach to Seismic Anomaly Detection. In: International Conference on Knowledge Science, Engineering and Management. Oct 28 . 2015, Cham; p. 575-587.

Dobrovolsky IP, Zubkov SI, Vi M. Estimation of the size of earthquake preparation zones. Pure and Applied Geophysics. Sep. 1979; 1 (117):5