Postprint: Anomaly Extraction Algorithm for GPS Data from Reference Stations Before and After Earthquakes

Authors: Li Nan, Kong Xiangzeng, Lin Ling

Date: 2018-05-18T00:00:00+00:00

Abstract

Currently, research on earthquake precursor anomalies primarily focuses on "thermal" and "electrical" aspects, with little attention given to GPS data from reference stations. However, scholars have demonstrated that GPS time series coordinate data from reference stations near the epicenter also contain precursor information for major earthquakes. This study investigates three representative earthquakes that occurred in the continental United States between 2001–2010, applies Martingale theory to GPS data processing for the first time, proposes an anomaly extraction algorithm, and analyzes GPS data from multiple reference stations near the epicenter before and after the earthquakes. Experimental results indicate that the algorithm can effectively reflect the anomalous variation trends in GPS data before and after major earthquakes, offering greater potential for forecasting major earthquakes using GPS data.

Full Text

Preamble

Title: Detection of Anomalies in GPS Data Before and After Earthquakes

Authors: Li Nan[1], Kong Xiangzeng[2], Lin Ling[2]
[1]College of Computer & Information Science, Fujian Agriculture & Forestry University, Fuzhou 350000, China
[2]College of Mathematics & Information, Fujian Normal University, Fuzhou 350000, China

Abstract: Most research studies on earthquake forecasting focus on thermal or electrical aspects. Scholars have proved that GPS time series also contains precursory information for large earthquakes. This paper investigates three typical earthquakes that struck the United States from 2001 to 2010 using GPS data

from several reference stations near the epicenters. We propose an algorithm based on Martingale theory to detect anomalies in GPS data. The experimental results show that the proposed algorithm can effectively reflect the change process in GPS data before and after large earthquakes, which offers more possibilities for earthquake forecasting using GPS data.

Key Words: earthquake; Martingale theory; anomaly detection

0 Introduction

Research indicates that GPS time series coordinate data from reference stations, after preliminary processing, can yield daily coordinate displacement solutions that reflect medium- and low-frequency crustal deformation information with good stability. Wang et al. used the Molchan diagram method to demonstrate that GPS deformation contains seismic precursor information. Sagiya et al. studied earthquakes and other crustal changes using time series coordinate data from multiple GPS reference stations. Yeha et al. established a GPS network based on multiple reference stations and found that GPS time series coordinate data showed disturbances before the 2013 Nantou and Hualien earthquakes in Taiwan. Wang et al. attempted to use a nonlinear filter to extract seismic-related anomaly signals from GPS data. Zhao et al. used data from several GPS reference stations in mainland China to study pre-seismic changes before Japan' s 2011 M9.0 earthquake, suggesting that the trend of slowing pre-seismic motion velocity might be a precursor.

Despite these efforts, research using GPS data as an indicator of seismic precursors remains relatively scarce. However, analyzing GPS data to identify seismic precursor anomalies is feasible. Existing achievements are primarily limited to the geography discipline, with analysis mainly based on manual observation, directly influenced by domain experts' knowledge and experience. This introduces significant uncertainty in identifying pre-seismic anomalies.

The purpose of data mining is to extract implicit, unknown, yet potentially useful knowledge and information from massive, incomplete, and noisy data. Currently, some scholars have applied anomaly detection methods from data mining to analyze changes in different indicators before earthquakes. For example, Marzocchi et al. used Bayesian estimation methods to analyze seismic data. Xiong et al. applied wavelet transforms to identify anomalies in OLR data before earthquakes. Konstantaras et al. used fuzzy-neural algorithms to detect pre-seismic electromagnetic anomalies. Li et al. proposed an anomaly mining algorithm for pre-seismic observation data based on errors and key points.

The main contribution of this paper is as follows: Research involving the use of GPS data from reference stations to analyze pre-seismic anomalies is still limited, and existing methods primarily rely on manual observation with high uncertainty in anomaly determination. Therefore, this paper utilizes GPS data provided by the Nevada Geodetic Laboratory and proposes an anomaly detection algorithm called ADM (Anomaly Detection based on Martingale theory for

GPS data) based on Martingale theory and the characteristics of GPS data. Applying this algorithm to three representative earthquakes in the United States, we analyze GPS time series data from reference stations near the epicenters before and after the earthquakes. The results not only further confirm that GPS data contains seismic precursor information but also provide a reference for applying data mining algorithms to geoscience.

1 Related Data

This study focuses on the largest Ms7.2 earthquake that occurred on U.S. mainland with depth less than 60km since 2001, along with two other earthquakes with relatively close epicenters. Earthquake information was obtained from the United States Geological Survey (USGS, https://earthquake.usgs.gov/earthquakes/), as detailed in Table 1.

Table 1 Earthquake Information

Date (y-m-d)	Latitude	Longitude	Depth (km)	Magnitude (JMA)
2001-02-28	47.15	-122.73	51.80	6.8
2008-02-21	32.35	-115.29	7.90	6.2
2010-04-04	32.26	-115.29	6.00	7.2

The GPS data used are reference station coordinates in three components: east-west, north-south, and vertical, obtained from the Nevada Geodetic Laboratory data sharing service (http://geodesy.unr.edu/). The data use the IGS08 framework. Considering station locations and data completeness, the GPS reference stations used in the experiments are listed in Table 2.

Table 2 GPS Station Information

Station	Latitude	Longitude	Height (m)
SEAT	47.58	-122.32	52.0
CLOV	32.96	-115.49	481.0
MEXI	32.03	-115.45	18.0
IID2	32.69	-115.93	-22.0
P500	32.50	-115.22	23.0

SEAT and CLOV stations correspond to two earthquakes occurring at different times (2001-02-28 and 2008-02-21) with relatively close epicenters, allowing analysis of GPS data from the same geographic location during different periods. MEXI, IID2, and P500 stations correspond to the M7.2 earthquake on 2010-04-04, enabling study of whether similar anomalies appear in GPS data from different reference stations near the same earthquake.

2 Research Methods

2.1 The ADM Algorithm

Based on the principle of minimizing differences among samples within the same cluster, let C represent the cluster center, calculated using Equation (1):

$$C = \frac{1}{num} \sum_{i=1}^{num} d_i$$

where d_i represents GPS data samples and num is the number of samples.

The offset degree of data d_t relative to center C can be measured using Equation (2):

$$s_t = \|d_t - C\|$$

where $\|\cdot\|$ can be any dissimilarity measure function; Euclidean distance is used in the experiments.

To analyze the distribution of s_t, the anomaly degree can be measured based on the number of nearest neighbor objects exceeding threshold θ in set $S = \{s_1, s_2, ..., s_t\}$. Therefore, the ADM algorithm uses Equation (3) to measure the stability of GPS data in the specified direction corresponding to d_t:

$$p_t = \frac{|\{s_j | s_j > \theta, j = 1, ..., t\}|}{|\{s_j | j = 1, ..., t\}|}$$

where θ is a random number in $[0, 1]$, and $|\{\#\}|$ is a function returning the number of samples satisfying the given condition. From Equation (3), we see that $p_t \in [0, 1]$, and larger p_t indicates d_t better conforms to the distribution of previous samples, meaning smaller likelihood of anomaly in the current day' s data.

However, GPS data inevitably contains noise (flicker noise, random walk noise, etc.). A relatively small p_t value on a single day does not necessarily indicate overall GPS data anomaly. Martingale theory must be used to comprehensively consider the trend of the entire time series.

Martingale theory originated in gambling and probability theory and is one of the earliest financial asset pricing models, now widely applied in decision optimization and investment analysis. To accurately reflect continuous changes in GPS data, combined with Martingale theory and GPS data characteristics, the randomized Martingale value (hereinafter M_t value) corresponding to each data d_t is calculated using Equation (4):

$$M_t(p) = \prod_{i=1}^{t} \varepsilon_i$$

where ε_i is a random variable in $[0, 1]$. According to Vovk et al., we set $\varepsilon_i = 0.82$ in our experiments.

From Equation (4), larger M_t values indicate more significant anomalies in GPS data over a certain previous period.

In clustering algorithms, the selection of initial cluster centers significantly affects results. Similarly, the choice of the first sample d_1 position in $Data$ substantially impacts anomaly extraction. Therefore, the algorithm sets an initial cluster sample parameter num. Intuitively, this uses the average of the first num days of data as the initial center and begins analysis after num days to reduce result uncertainty.

Additionally, since crustal movement is relatively intense before and after large earthquakes, GPS data may fluctuate significantly in short periods. According to Equation (4), M_t values would rapidly increase to uncontrollable levels. To prevent this, the algorithm sets a stop parameter h. If $M_t > h$, the algorithm restarts from d_{t+1}. The effects of parameters num and h on algorithm performance are further analyzed in the experimental section.

2.2 Algorithm Steps

Algorithm Name: ADM (Anomaly Detection based on Martingale theory for GPS data)

Input: GPS data in a specified direction from a reference station; stop parameter h; initial cluster sample number num

Output: M_t value for each sample in $Data$

Step 1: Initialize sample center - Calculate initial cluster center C using the first num samples - Set current processing sample index $t = num + 1$ - Set loop stop condition $isStop = false$ - Set current batch start index $front = 1$

Step 2: While $isStop = false$ 1. Calculate offset s_t using C and Equation (2) 2. Calculate stability p_t using θ and Equation (3) 3. Calculate Martingale value M_t using p_t and Equation (4) 4. Update cluster center C using d_t and Equation (2) 5. If $M_t > h$ then - Set $isStop = true$ 6. $t = t + 1$ 7. If $t > |Data|$ then - Exit loop 8. End if

Step 3: If $isStop = true$ - Start a new batch: set $front = t - num$ - Reinitialize M values: $M_{front}, M_{front+1}, ..., M_t = 1$ - Set $isStop = false$ - Return to Step 2

Step 4: Output M_t **values**

3 Results and Analysis

3.1 Analysis of SEAT and CLOV Stations

The original GPS time series coordinate data and corresponding M_t values for SEAT and CLOV stations are shown in Figure 1 [Figure 1: see original paper] and Figure 2 [Figure 2: see original paper]. These two reference stations correspond to two earthquakes occurring at different times with relatively close epicenters. In the ADM algorithm, parameters are set as $num = 45$ and $h = 1000$. The studied data span from October of the year before the earthquake to one month after. To reduce the impact of random θ values, the figures show averages from 10 experimental runs. For better observation of GPS data trends, the original values use only the fractional parts while ignoring identical integer parts. Yellow vertical lines indicate earthquake occurrence times.

From Figures 1(d) and 2(d), we observe that M_t values show almost no change in the early stage, indicating stable GPS data. Near the earthquake time, both stations show significant changes in east-west and north-south components' M_t values, rapidly reaching peaks about one week after the earthquake. This further confirms that GPS data contains earthquake-related information. Notably, the reason M_t values fluctuate before the earthquake but peak after may be that GPS anomalies can appear only a short time before earthquakes. However, to reflect trends and reduce noise effects, the algorithm requires accumulation of a certain number of anomalous samples before M_t shows significant changes. Therefore, pre-seismic GPS anomalies may manifest as M_t peaks shortly after the earthquake. While setting h to smaller values could advance peak timing, this might cause false alarms, as verified in experiment 3c.

Since the two earthquakes have relatively close epicenters and similar causes, the extracted M_t variation patterns in east-west and north-south components are similar, demonstrating the reliability of the ADM algorithm. Moreover, the original GPS data images show no obvious anomaly trends, confirming the algorithm' s effectiveness in extracting GPS anomalies. Because vertical GPS measurement errors are much larger than horizontal components, vertical M_t fluctuations are less pronounced and show no clear earthquake-related patterns.

3.2 Analysis of MEXI, IID2, and P500 Stations

The original GPS data and corresponding M_t values for MEXI, IID2, and P500 stations near the 2010-04-04 earthquake epicenter are shown in Figures 3 [Figure 3: see original paper], 4 [Figure 4: see original paper], and 5 [Figure 5: see original paper]. In these experiments, $num = 45$ and $h = 1000$. Because GPS data at each station showed "cliff-like" changes shortly after this earthquake, the studied period extends from October of the previous year to two months after the earthquake. To clearly show anomaly trends in three components, Figures 3-5 separately display M_t values for each direction.

Comparing Figures 3(c)(d) reveals that due to post-earthquake "cliff-like" GPS

data changes in specific directions, corresponding M_t values peak during subsequent periods. Similar patterns appear in Figures 4(a)(b) and 5(c)(d). The delayed peak appearance demonstrates that the ADM algorithm truly reflects GPS data anomalies and suggests that anomalies causing M_t peaks in experiment 1 may have occurred before the earthquake. In Figures 1(a) and 1(b), MEXI station' s east-west component doesn' t show similar patterns because potential violent GPS changes occurred before the earthquake, causing M_t to exceed the stop parameter h and triggering algorithm restart, which then fitted this "cliff" data.

Like MEXI and IID2 stations, P500 station's north-south M_t values also changed before the earthquake, though less dramatically, possibly due to the station' s specific location. Similar to experiment 1, vertical component M_t values show no consistent patterns across the three stations. Only IID2 station' s vertical M_t shows clear fluctuations around the earthquake time. MEXI station' s vertical fluctuations may stem from "cliff-like"data changes, while P500 station's vertical M_t peak appears some time after the earthquake.

Geographically, all three stations are near the epicenter, so no significant differences exist in the first M_t peak timing for east-west and north-south components (with smaller measurement errors). However, MEXI station (closest to the epicenter) shows earlier first-peak timing in both components than IID2 and P500 stations, indicating that closer stations may experience earlier GPS anomalies.

3.3 Relationship Between Earthquake Magnitude and M_t Values

Further analysis of the relationship between earthquake magnitude, occurrence time, and M_t values from east-west and north-south components reveals that for the largest 2010-04-04 earthquake (Ms7.2), all three stations' M_t values first peaked about three months before the earthquake, with changes continuing until the event. For the relatively smaller 2001-02-28 (Ms6.8) and 2008-02-21 (Ms6.2) earthquakes, M_t values began changing about one month before each earthquake, peaking shortly after. This suggests that larger earthquakes require more energy accumulation, so GPS anomalies may occur earlier, causing M_t values to fluctuate sooner. The 2001-02-28 earthquake' s M_t fluctuations appeared later than the 2008-02-21 earthquake, possibly because the former' s focal depth (51.80km) is much greater than the latter' s (7.90km), making pre-seismic GPS changes less obvious.

3.4 Parameter Sensitivity Analysis

Effect of parameter h: To evaluate h' s impact, we tested values of 250, 500, 1000, and 2000 on MEXI station' s north-south component with $num = 45$ (Figure 6 [Figure 6: see original paper]). When GPS anomalies occur, M_t grows exponentially. With the same num, smaller h values advance the first M_t peak timing. However, excessively small h may cause false alarms, making $h = 1000$ a reasonable choice.

Effect of parameter *num*: We tested *num* values of 25, 35, 45, and 55 on MEXI station' s north-south component with $h = 1000$ (Figure 7 [Figure 7: see original paper]). When GPS data is stable, sufficiently large *num* ensures similar initial cluster centers C across different *num* values, so *num* has minimal impact on first peak timing. However, when the algorithm restarts after detecting anomalies, different *num* values produce different initial cluster centers, causing variations in second peak timing. Although $num = 55$ yields the earliest second peak, excessively large *num* may overly extend the interval between two peaks in practical applications, so we use $num = 45$.

3.5 Comparison with Sigma Criterion

To compare ADM with conventional methods using domain expert knowledge, we applied the $k\sigma$ criterion to SEAT station's north-south data. In this criterion, $\bar{x}$ represents the mean and σ the standard deviation; data outside $[\bar{x}-k\sigma, \bar{x}+k\sigma]$ are considered anomalous. Following reference [22], we set $k = 2$ (Figure 8 [Figure 8: see original paper]).

The $k\sigma$ criterion failed to detect pre-seismic anomalies, instead showing multiple exceedances weeks after the earthquake. In contrast, Figure 1(d) shows M_t peaks appearing shortly before and after the earthquake, clearly reflecting anomaly trends. Thus, the ADM algorithm demonstrates superior performance.

4 Conclusion

This paper 首次 applies Martingale theory to GPS data, proposing the ADM algorithm for anomaly extraction and analyzing GPS time series from reference stations near earthquake epicenters. The results show that M_t value changes can truly reflect GPS data' s potential variation trends, unlike traditional geography methods limited to expert observation. Even when original GPS values show no obvious anomalies, corresponding M_t values change significantly around earthquake time, potentially peaking before the event. This confirms that GPS deformation contains seismic precursor information and offers possibilities for earthquake forecasting using GPS data.

However, based on current data volume and research level, exploring earthquake forecasting using reference station GPS data remains a long-term endeavor. While the ADM algorithm reveals preliminary patterns in M_t anomalies, further verification across more earthquake cases is needed to statistically analyze relationships between M_t trends and earthquake parameters (magnitude, location, time). This represents our future research direction.

References

[1] Saradjian M R, Alhoondzadeh M. Thermal anomalies detection before strong earthquakes (M>6.0) using interquartile, wavelet and Kalman filter method [J]. Natural Hazards and Earth System Sciences, 2011, 11(1): 1099-1108.

[2] Honkura Y, Oshiman N, Matsushima M, et al. Rapid changes in the electrical state of the 1999 Izmit earthquake rupture zone [J]. Nature Communications, 2013, 4(2116): 1-10.

[3] Kong X, Bi Y, Glass D H. Detecting seismic anomalies in outgoing long-wave radiation data [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015, 8(2): 649-660.

[4] Han P, Hattori K, Hirokawa M, et al. Statistical analysis of ULF seismomagnetic phenomena at Kakioka, Japan during 2001-2010 [J]. Journal of Geophysical Research: Space Physics, 2014, 119(6): 4998-5011.

[5] Němec F, Santolík O, Parrot M. Decrease of intensity of ELF/VLF waves observed in the upper ionosphere close to earthquakes: A statistical study [J]. Journal of Geophysical Research: Atmospheres, 2009, 114(4): 1-10.

[6] Liu J Y, Chen Y I, Huang C H, et al. A statistical study of lightning activities and M 5.0 earthquakes in Taiwan during 1993–2004 [J]. Surveys in Geophysics, 2015, 36(6): 851-859.

[7] Kuo C L, Lee L C, Heki K. Preseismic TEC changes for Tohoku-Oki earthquake: Comparisons between simulations and observations [J]. Terrestrial, Atmospheric & Oceanic Sciences, 2015, 26(1): 63-72.

[8] Zhang Y, Wu Y, Duan W, et al. Relationship between GPS long-term trends and large earthquakes [J]. Geomatics and Information Science of Wuhan University, 2012, 37(6): 675-678.

[9] Wang T, Zhuang J, Kato T, et al. Assessing the potential improvement in short-term earthquake forecasts from incorporation of GPS data [J]. Geophysical Research Letters, 2013, 40(11): 1-5.

[10] Sagiya T, Miyazaki S, Tada T. Continuous GPS array and present-day crustal deformation of Japan [C]//Proc of the 1st Workshop of the APEC Cooperation for Earthquake Simulation. IEEE Press, 2000: 2303-2308.

[11] Yeha Y L, Chenga K C, Wanga W H, et al. Very short-term earthquake precursors from GPS signal interference based on the 2013 Nantou and Rueisuei earthquakes, Taiwan [J]. Journal of Asian Earth Sciences, 2015, 114(2): 312-320.

[12] Wang T, Bebbington M. Identifying anomalous signals in GPS data using HMMs: An increased likelihood of earthquakes [J]. Computational Statistics and Data Analysis, 2013, 58(1): 27-44.

[13] Zhao G, Li P. Coseismic displacement and pre-seismic changes measured by GPS continuous observation stations in mainland China for Japan' s 9.0 earthquake [J]. Earthquake, 2012, 32(2): 129-134.

[14] Marzocchi W, Zechar J, Jordan T. Bayesian forecast evaluation and ensemble earthquake forecasting [J]. Bulletin of the Seismological Society of America, 2012, 102(6): 2574-2584.

[15] Xiong P, Shen X, Bi Y, et al. Study of outgoing longwave radiation anomalies associated with Haiti earthquake [J]. Natural Hazards & Earth System Sciences, 2010, 10(10): 2169-2178.

[16] Konstantaras A, Varley M, Vallianatos F, et al. Detection of weak seismoelectric signals upon the recordings of the electrotelluric field by means of neurofuzzy technology [J]. IEEE Geoscience & Remote Sensing Letters, 2007, 4(1): 161-165.

[17] Li Z, Chen J, Wang L, et al. An anomaly mining algorithm for seismic precursor observation data based on errors and key points [J]. Computer Application Research, 2011, 28(8): 2987-2901.

[18] Ho S S, Wechsler H. A Martingale framework for detecting changes in data streams by testing exchangeability [J]. IEEE Transactions on Software Engineering, 2010, 32(12): 2113-2127.

[19] Rebischung P, Griffiths J, Ray J, et al. IGS08: the IGS realization of ITRF2008 [J]. GPS Solutions, 2012, 16(4): 483-494.

[20] Kanungo T, Mount D M, Netanyahu N, et al. An efficient k-means clustering algorithm: Analysis and implementation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 881-892.

[21] Vovk V, Nouretdinov I, Gammerman A. Testing exchangeability on-line [C]//Proc of the 12th International Conference on Machine Learning. Morgan Kaufmann, 2003: 768-775.

[22] Wu L X, Qin K, Liu S J. GEOSS-Based thermal parameters analysis for earthquake anomaly recognition [J]. Proceedings of the IEEE, 2012, 100(10): 2891-2907.

Note: Figure translations are in progress. See original paper for figures.