ENTROPY MEASURE OF OUTLIERS IN GPS TIME SERIES

P.V. Yakovlev

Ordzhonikidze Russian State Geological Prospecting University, Moscow, Russia

Abstract. The method of significant outliers detection in GPS time series is proposed. The approach consists in measuring deviation at each time point relative to both left and right parts of the original time series of such statistics as standard deviation. The standard deviation is chosen because of its sensitivity to small value changes in time series. Normalized entropy is used to define to what extend outliers affect a signal.

Examples of 30-minute GPS signal analysis before and after the mega-earthquake in Japan (March 11, 2011) and maps of normalized entropy, identifying anomalous zones, are presented in the article. It is shown that the epicenter is characterized by low entropy of outliers both before and after the seismic catastrophe. While the low entropy of outliers after the event is easily explained by post-seismic and aftershocks effects, the low entropy anomaly of outliers, appeared before the earthquake, is more important result of the carried out analysis.

Keywords: GPS signals, time series analysis, earthquake forecast, outliers detection.

Introduction

A dynamic system, which observations are recorded as a discrete time series, often represents a very noisy signal. At state of rest such signals often have a clearly defined trend or oscillations, against the background of which there is a noise component caused by irrelevant effects, recorder itself or by random oscillations of the system as well. One of the main goals of observation and analysis of the received signals is the forecast of events that disturb the system from the state of rest: it can be epileptic seizures [Osorio, Lyubushin, Sornette, 2011], economic crises [Leung, Thulasiram, Bondarenko, 2006], earthquakes [Соболев, Любушин, 2006; Соболев, Любушкин, Закржевская, 2008] etc. On the other side, the presence of the random outliers in the noise component in geophysical time series provides a greater width of singularity multifractal spectrum support that appears to be a sign of “a healthy chaos” and can be an indicator of a safe condition of a seismically active region. [Любушкин, 2007, 2009, 2014; Любушкин и др., 2015; Lyubushin, 2012].

The complexity of the forecast is that it is not exactly known what should be looked for in the time series. The signs of the event can be the change of the trend slope, the increase of oscillations frequency or amplitude, outliers, signal jumps or all they together. It may turn out that nothing discovered will help to make a correct forecast because another obstacle in making decision is an unknown lower and upper threshold. Once a change of oscillation frequency or trend break is detected, it is often difficult to say to what extend it is abnormal or whether it is abnormal.

There is a large set of methods and algorithms that allow to detect the time series features, that are mentioned above, and to implement a wide range of approaches, some of which are described in [Соловьев и др., 2012; Aggarwal, Yu, 2001; Bay, Schwabacher, 2003; Sun, Chawla, Arunasalam, 2006]. So, in [Aggarwal, Yu, 2001; Bay, Schwabacher, 2003] the widespread algorithm of outliers detection, based on the “nearest neighbors” distances, is described. The popularity of this algorithm is caused by the fact that it does not require the introducing of
any distributions to identify an abnormal behavior in the time series, making it suitable for analysis of almost any signal.

This paper presents a method based on the outliers’ randomness measure definition in the GPS time series that is considered as a normalized entropy of some outliers’ statistics. The similar measure of the stepwise part of GPS time series is considered in [Lyubushin, Яковлев, 2015], which suggested the threshold value of the normalized entropy statistics of the jumps (approximately equals to 0.89) and that allowed to determine the criterion for the presence of average level steps in the signals.

Comparing the outliers and jumps in the time series it could be said that the outliers are simpler anomaly in its nature as they have a point character and a weak impact on such statistic quantities as a mean value and standard deviation. Because of its simplicity, the outliers usually occur in the time series more often that is why its analysis can be theoretically more informative than the analysis of jumps.

In this article the analysis is carried out on the example of the GPS signals registered on the territory of the Japanese islands by 1248 stations in the period 30.02-26.03.2011 with a 30-minute time step. The used data is available in the Internet at the address http://quakesim.org/tools/timeseries. Every signal represents a set of three time series corresponding to the shifts on the east, north and verticals.

The aim of this research was to understand whether it is possible to identify the seismically dangerous zones, basing on time series outliers, and if so, to suggest a criterion of seismic activity zones identification.

The same data was considered previously in the works [Lyubushin, Yakovlev, 2014; Любулин, Яковлев, Родионов, 2015], the first one shows that the area of the future earthquake is distinguished by the anomaly of spectral exponent value estimated using orthogonal wavelets. The question of separation of the random and informative outliers is solved by consideration of several GPS signals at once taken from the nearby stations – if the outlier is presented in most signals then it can be considered informative.

**Statistics of the time series outliers.**

Let $X(t)$ be a time series where $t=0,\ldots, N-1$. In each moment of time $t=i$ the main point of interest will be to what extent the current signal value differs from the values on the left and right of it. As a comparative characteristic, the following differences can be regarded:

$$L_i = \left| L_\sigma^i - L_\sigma^{i-1} \right|, \quad R_i = \left| R_\sigma^i - R_\sigma^{i+1} \right|,$$

(1)

where $L_\sigma^i, R_\sigma^i$ are standard deviations calculated for $t\in[0, i]$ and $t\in[i, N-1]$ correspondingly. Values $L_\sigma^{i-1}$ and $R_\sigma^{N}$, that are out of possible $t$ values range, are considered equal to zero (explanation of this condition is given below). Values (1) help to determine whether the current signal value strongly deviates from the left and right sides separately: $L_i$ by its nature is a reflection of the time series evolution while $R_i$ shows how considerable is the contribution that this evolution introduced to the last value formation $X(N-1)$. The fact that this estimation is asymmetric can lead to the worsening of outliers detection at the edges of the signal. To offset this fact the measure of deviation relative to the whole signal should be defined by introducing the weighted sum of standard deviations increments (1):

$$W_i = \frac{t}{N} L_i + \frac{N-1-t}{N} R_i,$$

(2)
At the moment of time \( t=0 \), when the comparison of the current value only with values on the right is possible the first summand (2) turns to zero; therefore value of \( L_{\sigma}^{-1} \) can be anything including zero. In another extreme case when \( t=N-1 \) by analogy the second summand (2) equals to zero and therefore \( R_{\sigma}^{N} \) can be considered as zero as well. An example of the weighted sum of standard deviations increments of the GPS signal is demonstrated in Fig.1.

![Fig. 1. Original GPS signal (shift of the north coordinates) with (a) 30-minute time step and (b) \( W \) for it. N is the number of samples](image)

It can be seen (see Fig. 1, b), that the value \( W \), calculated with (2), is sensitive to jumps and trends in the original signal, therefore the outliers analysis from the given schedule can be inconvenient. To avoid such difficulties in the case of a clearly defined trend of the time series before calculating \( W \) this trend should be eliminated or, in general case, just come to increments. Thus, the clear statistics of outliers can be received (Fig.2). The outlier, allocated to the increments of the signal, always corresponds with the outlier in the original time series, however it should be considered that the reverse statement is not true, therefore after coming to increments the \( W \) statistics can qualitatively differ from which is calculated for the original signal. It should be noted that the value (2) for the original signal can have the effect of overstating values at the edges in the presence of steep trends.
Fig. 2. The same that is in the Fig. 1 after coming to increments

On the graphs of W for the original signal before coming to increments (Fig. 1, б) and after it (Fig. 2, б) there is a clearly defined line of fluctuation around zero values which displays the noise that doesn’t affect the standard deviation of the signal, i.e. natural oscillations of the system. Considering that the natural oscillations in the time series should prevail, i.e. the outliers are a rare phenomenon in the signal, the threshold value of these oscillations can be found. As the threshold value can be taken the right border of the interval where the histogram of the weighted sum of standard deviations increments reaches its maximum. In case there are several equal maximums, the most distant bound of the histogram interval from zero must be chosen. After this all values of W that are less than the found threshold can be considered equal to zero, that allows working further only with the outliers. The variable search parameter of the threshold value is a number of histogram bins – the higher it is the more detailed the estimation of the outliers’ distribution density is. This way a false maximum of the histogram can appear. One of the most common estimates of the optimal number of bins [Новицкий, Зограф, 1991; Чернецкий, 1994], which is also used in this paper is

\[ n = \left\lfloor \sqrt{N} \right\rfloor, \]

where \( n \) is a number of bins; \( \left\lfloor \ldots \right\rfloor \) is the integer part of the number.

The result of the threshold calculation that separates the natural oscillations from the sharp deviations of the signal increments is shown on Fig. 3. The threshold value is approximately equal to \( 3.57 \times 10^{-6} \).
Mapping the calculated statistics

Let us have a look at analyzed outliers’ statistics after coming from the original signals to increments on the grid of 50×50 nodes, which covers the investigated Japan region. One can calculate the weighted sum of standard deviations increments for 30-minute time series resulting from the registration of vertical coordinates shift of each 1248 GPS stations during the period 30.01.—10.03.2011 (40 days), i.e. before the Tohoku earthquake and plot the averaged maps of $W$ (Fig. 4, 5). The values for each node are calculated by averaging the calculated statistics of the outliers for the nearest stations to the node. The number of such stations (‘the nearest neighbors’) consider equal to 10. The calculation will be carried out in the moving time window, which length equals 7 days (336 samples).

On the maps presented in Fig.4 it is almost impossible to identify any clear signs of the upcoming catastrophe. As mentioned above, the outliers have the point character and therefore the main contribution to the averaging is made by the natural oscillations of the system. Thus, these maps are not informative for allocation of the anomalous zones.

The normalized entropy of the outliers for the value of $W$ is calculated by coming from the outliers to their probabilities

$$p(t) = \frac{W_i}{\sum_k W_k}$$

and introducing the normalized entropy according to the formula
Fig. 4. Averaged maps of $W$ for (a) vertical and (b) east coordinates plotted for the period 30.01.–10.03.2011. The star is the epicenter of the earthquake 11.03.2011.

In the argument of the logarithm is $N–1$, and not $N$, since we came to increments thereby reducing the number of counts per unit. In the general case, when statistics is calculated for the original signal, under the logarithm should be the number of counts equal to $N$.

Let us plot maps of normalized entropy for three components of the GPS signal (east, north, vertical) without using the moving time window for two periods – before the Tohoku earthquake (30.01.–01.03.2011) and after it (12.03.–26.03.2011). For the accuracy of the experiment, the last ten days before the disaster are excluded from the period before the earthquake so that the foreshocks that occurred in this period of time were not involved in the analysis.

The more outliers contains the value $W$, the higher the probability that there are other oscillations in the observed system caused by an unknown source in addition to the natural oscillations.

The preparation of the earthquakes starts with the destruction of rocks located at the junctions of lithospheric plates. These events may correspond to the stepwise behavior of the signal as well as the appearance of outliers. Based on this, it can be assumed that the increase of outliers in the signal is a result of more active movements of the earth blocks. In this regard, the most interesting regions for investigation seem to be the regions with the low values of normalized entropy characteristic for the time series with the increased amount of outliers.

Maps before the earthquake (Fig. 5, left column) demonstrate for all coordinates the “normal” $En$ value is the value more than 0.9. However, in the area close to the epicenter (green star), can be seen the low values of $En$, i.e. the anomalous area tends to be the epicenter of the future event. It can be concluded that the appearance of regions of low values of normalized entropy on the maps (in our case $En<0.9$) can be interpreted as a sign of potential disaster.

Maps made after the earthquake (Fig. 5, right column) demonstrate that the zone of the low $En$ values shifted to the zone of the strongest aftershocks. This confirms the hypothesis that the low normalized entropy of $W$ statistics reflects the instability of the analyzed system. The observed anomaly in the region of the earthquake epicenter is explained by the presences of aftershocks.
Fig. 5. Maps of normalized entropy of (a) east, (b) north and (c) vertical components of GPS signal before the earthquake 30.01.–01.03.2011 (left column) and after it 12.03.–26.03.2011 (right column). The star is the epicenter of the earthquake 11.03.2011.
Conclusions

The introduced method of the outliers estimation, based on the calculation of the weighted sum of standard deviations increments of GPS signals time series, allowed to identify the anomalous area on the example of the Tohoku earthquake (Japan, 11.03.2011, $M_W=9.0$), located close to the epicenter of the disaster and characterized by the low values of the outliers entropy. This fact can be interpreted as follows. The preparation of the earthquake occurred in this region and not in the region where the epicenter of the event was located, i.e. the rocks located at the junctions of lithospheric plates began to crumble right there. Such destructions are reflected as outliers in the time series of GPS signals that in turn lower the value of entropy.

Литература

Bay S.D. and Schwabacher M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule, Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 2003
Chernetskiy V.I. Matematicheskoe modelirovanoe stokasticheskikh system (Mathematic modelling of stochastic systems), Izd. of Petrozavodsk State University, 1994 (in Russ).
Lyubushin A.A. Prognostic properties of random fluctuations of geophysical characteristics, Biosphere, 2014, vol. 6, no. 4, pp. 313–332
Novitskiy P.V. and Zograf I.A. Estimation of measurement results errors, Energoatomizdat, 1991
Энтропийная мера выбросов во временных рядах сигналов GPS

Sun P., Chawla S., and Arunasalam B. Outlier detection in sequential databases, Proc. SIAM International Conference on Data Mining, 2006