ГЕОФИЗИЧЕСКИЕ ИССЛЕДОВАНИЯ, 2016, том 17, № 1, с.37-45

УДК 550.334

ЭНТРОПИЙНАЯ МЕРА ВЫБРОСОВ ВО ВРЕМЕННЫХ РЯДАХ СИГНАЛОВ *GPS*

© 2016 г. П.В. Яковлев

Российский государственный геологоразведочный университет им. С. Орджоникидзе, г. Москва, Россия

Предлагается метод выделения значимых выбросов во временных рядах сигналов *GPS*. Подход заключается в измерении отклонения в каждый момент времени относительно левой и правой частей анализируемого временного ряда такой статистики, как стандартное отклонение. Стандартное отклонение выбрано потому, что оно является чувствительным к малым изменениям значений временного ряда. В качестве определения меры вклада, привносимого выбросами в сигнал, используется нормализованная энтропия.

Приводятся примеры анализа рядов сигналов *GPS* с 30-минутным шагом дискретизации, зарегистрированных в Японии до и после мегаземлетрясения Тохоку (11.03.2011 г., M_W=9.0); представлены карты нормализованной энтропии, построенные для выявления аномальных зон. Показано, что эпицентральная область характеризуется пониженной энтропией выбросов как до, так и после сейсмической катастрофы. Если пониженная энтропия выбросов после события легко объясняется постсейсмическими эффектами и последствиями афтершоков, то выявление аномалии пониженной энтропии, возникшей до сейсмической катастрофы, представляется наиболее важным результатом проведенного анализа.

Ключевые слова: сигналы *GPS*, анализ временных рядов, прогноз землетрясений, выделение выбросов.

Введение

Динамическая система, наблюдения за которой мы регистрируем в виде дискретных временных рядов, часто представляет собой очень зашумленный сигнал. В состоянии покоя такие сигналы зачастую имеют четко выраженный тренд или колебания, на фоне которых мы видим шумовую составляющую, вызванную как посторонними явлениями или самим регистрирующим прибором, так и собственными случайными колебаниями системы. Одна из основных целей наблюдения и анализа полученных сигналов – прогноз событий, выводящих систему из состояния покоя, которыми для разных систем могут быть эпилептические припадки [*Osorio, Lyubushin, Sornette,* 2011], экономические кризисы [*Leung, Thulasiram, Bondarenko,* 2006], землетрясения [*Соболев, Любушин,* 2006; *Соболев, Любушин, Закржевская,* 2008] и т.д. С другой стороны, наличие в шумовой составляющей геофизических временных рядов хаотичных выбросов обеспечивает большую величину ширины носителя мультифрактального спектра сингулярности, что является признаком “здорового хаоса” и может быть индикатором безопасного состояния сейсмически активной области [*Любушин,* 2007, 2009, 2014; *Любушин и др.,* 2015; *Lyubushin,* 2012].

Сложность прогноза заключается в том, что мы не знаем точно, что следует искать во временных рядах. Предвестниками интересующего нас события могут быть изменение угла наклона тренда, увеличение частоты или амплитуды колебаний, выбросы, скачки сигнала или все это в совокупности. Может оказаться, что ничто из обнаруженного не поможет нам сделать правильный прогноз, так как еще одним препятствием в принятии решения является незнание верхнего и нижнего порогов – выделив в сигнале участки изменения частот колебаний или изломы тренда, мы зачастую затрудняемся сказать, насколько это аномально и аномально ли вообще.

Существует большой набор методов и алгоритмов, позволяющих выделять упомянутые выше особенности временных рядов и реализующих самые разные подходы, некоторые из которых описаны в [*Соловьев и др.*, 2012; *Aggarwal, Yu*, 2001; *Bay, Schwabacher*, 2003; *Sun, Chawla, Arunasalam*, 2006]. Так, в [*Aggarwal, Yu*, 2001; *Bay, Schwabacher*, 2003] описан распространенный алгоритм выделения выбросов, основанный на определении расстояний до "ближайших соседей". Популярность использования этого алгоритма связана с тем, что он не требует введения каких-либо распределений с целью выявления аномального поведения во временных рядах, что делает его подходящим для анализа практически любых сигналов.

В данной статье представлен метод, основанный на определении меры хаотичности выбросов во временных рядах сигналов *GPS*, в качестве которой рассматривается нормализованная энтропия некоторой статистики выбросов. Аналогичная мера скачкообразной составляющей временных рядов сигналов *GPS* рассмотрена в [*Любушин, Яковлев*, 2015], где было предложено пороговое значение нормализованной энтропии статистики скачков, приблизительно равное 0.89, что позволило определить критерий наличия скачков среднего уровня в сигналах.

Сравнивая выбросы и скачки во временных рядах, можно сказать, что выбросы являются более простой по своей природе аномалией, так как имеют точечный характер и намного слабее влияют на такие статистические величины, как математическое ожидание и стандартное отклонение. В силу своей простоты выбросы, как правило, встречаются во временных рядах намного чаще, в связи с чем гипотетически их анализ может оказаться более информативным, чем анализ скачков.

В статье анализ проводится на примере сигналов *GPS*, зарегистрированных на территории Японских островов 1248 станциями в период 30.02–26.03.2011 г. с 30-минутным шагом по времени. Использованные данные доступны в сети Интернет по адресу http://quakesim.org/tools/timeseries. Каждый сигнал представляет собой набор из трех временных рядов, соответствующих смещениям на восток, север и вертикальным.

Цель исследований состояла в том, чтобы, во-первых, в результате анализа понять, возможно ли выявление сейсмически опасных зон на основе выбросов во временных рядах, и, если да, то предложить критерий выделения зон повышенной сейсмической активности.

Ранее те же данные были рассмотрены в работах [*Lyubushin, Yakovlev*, 2014; *Любушин, Яковлев, Родионов,* 2015], в первой из которых показано, что область будущего землетрясения выделяется аномалией значений спектральной экспоненты, оцениваемой с помощью ортогональных вейвлетов. Вопрос разделения случайных и информативных выбросов решается рассмотрением сразу нескольких рядов сигналов *GPS*, взятых с близлежащих станций – если выброс присутствует в большинстве рядов, то это дает основание считать его информативным.

Статистика выбросов временных рядов

Пусть $X(t)$ – временной ряд, где $t=0, \ldots, N-1$. В каждый момент времени $t=i$ нас будет интересовать, насколько текущее значение сигнала отличается от значений, взятых слева и справа от него. В качестве сравнительной характеристики рассмотрим следующие разности:

$$\begin{aligned} L_i &= \left| L_\sigma^i - L_\sigma^{i-1} \right|, \\ R_i &= \left| R_\sigma^i - R_\sigma^{i+1} \right|, \end{aligned} \tag{1}$$

где L_σ^i, R_σ^i – стандартные отклонения, рассчитанные для значений $t\in[0, i]$ и $t\in[i, N–1]$ соответственно. Величины L_σ^{-1} и R_σ^N, выходящие за границы возможных значений t, положим равными нулю (пояснение этого условия изложено ниже). Величины (1) позволяют определить, сильно ли отклоняется текущее значение временного ряда относительно левой и правой частей в отдельности: L_i по своей сути является отражением эволюции временного ряда, в то время как R_i показывает, насколько значимый вклад данная эволюция привнесла в формирование последнего значения $X(N–1)$ временного ряда. Данная оценка является несимметричной, что может привести к ухудшению выделения выбросов по краям сигнала. Чтобы нивелировать этот факт, определим меру отклонения относительно всего сигнала в совокупности, введя взвешенную сумму приращений стандартных отклонений (1):

$$W_t = \frac{t}{N} L_t + \frac{N-1-t}{N} R_t. \qquad (2)$$

В момент времени $t=0$, когда возможно сравнение текущего значения только со значениями справа, мы видим, что первое слагаемое суммы (2) обращается в нуль; поэтому значение величины L_σ^{-1} может быть любым, в том числе и нулевым. В другом крайнем случае при $t=N–1$ по аналогии второе слагаемое суммы (2) равно нулю и, следовательно, R_σ^N также можно положить равным 0. Пример взвешенной суммы приращений стандартных отклонений сигнала *GPS* приведен на рис. 1.

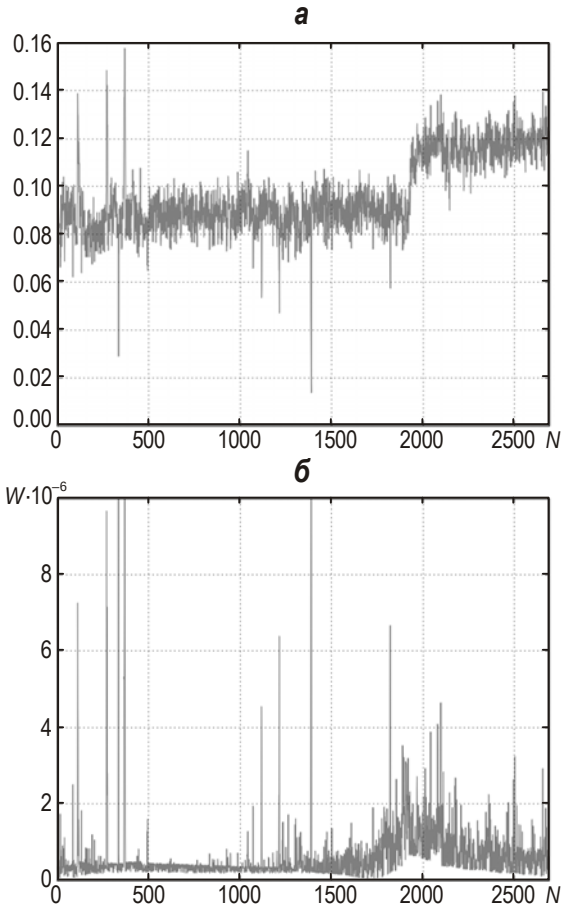

Рис 1. Исходный сигнал *GPS* (смещение северных координат), взятый с 30-минутным шагом по времени (*а*), и величина *W* для него (*б*). *N* – число отсчетов

Можно видеть (см. рис. 1, *б*), что величина *W,* вычисляемая по (2), чувствительна к скачкам и наличию трендов в исходном сигнале, в связи с чем анализ выбросов по получившемуся графику может оказаться неудобным. Чтобы избежать этих трудностей, в случае четко выраженного тренда временного ряда перед расчетом *W* необходимо этот тренд устранить или в общем случае просто перейти к приращениям. Таким образом мы получим очищенную статистику выбросов (рис. 2). Выброс, выделенный для приращений сигнала, всегда соответствует выбросу в исходном временном ряду, однако нужно учитывать, что обратное утверждение неверно, поэтому после перехода к приращениям статистика *W* качественно может отличаться от той, что рассчитана для исходного сигнала. Стоит отметить, что величина (2) для исходного сигнала при наличии крутых трендов, может иметь эффект завышения значений по краям.

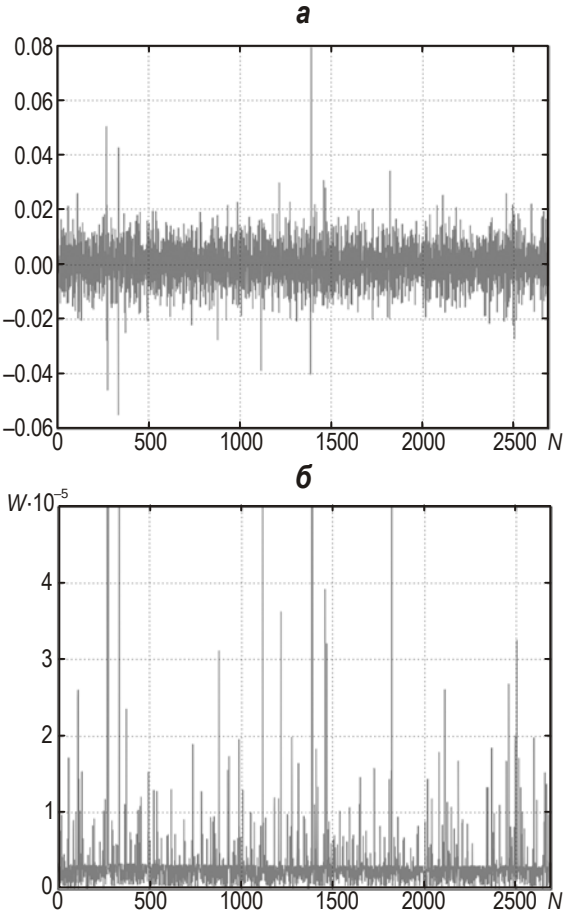

Рис 2. То же, что на рис. 1, после перехода к приращениям

На графиках *W* для исходного сигнала до перехода к приращениям (рис. 1, *б*) и после него (рис. 2, *б*) наблюдается ярко выраженная полоса колеблющихся около нуля значений, которая отображает шум, не влияющий на стандартное отклонение сигнала, т.е. естественные колебания системы. Полагая, что естественные колебания во временном ряду должны преобладать, т.е. выбросы – более редкое явление в сигнале, найдем пороговое значение этих колебаний. Возьмем в качестве него правую границу интервала, на которой гистограмма взвешенной суммы приращений стандартных отклонений достигает своего максимума. В случае, если встречается несколько одинаковых максимумов, выбирается наиболее удаленная от нуля граница интервала гистограммы. После этого можно положить равными нулю все значения *W*, которые меньше найденного

порога, что позволяет в дальнейшем работать только с выбросами. Изменяемым параметром поиска порогового значения является число бинов гистограммы – чем оно выше, тем более детальна оценка плотности распределения выбросов. Таким образом, может появиться ложный максимум гистограммы. Одной из наиболее распространенных оценок оптимального числа бинов [*Новицкий, Зограф,* 1991; *Чернецкий,* 1994], которая используется и в данной статье, является

$$n = \left[\sqrt{N}\right], \tag{3}$$

где n – число бинов; […] – целая часть числа.

Результат расчета порогового значения, отделяющего полосу естественных колебаний от резких отклонений приращений сигнала, представлен на рис. 3. Пороговое значение приблизительно равно $3.57 \cdot 10^{-6}$.

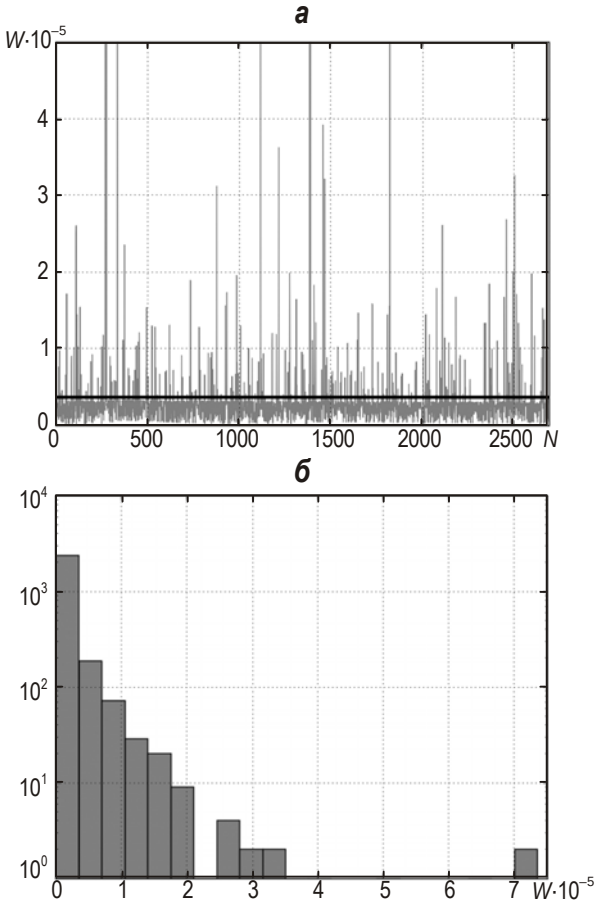

Рис 3. Величина W для исходного сигнала *GPS* после перехода к приращениям (*а*) и её гистограмма (*б*). Горизонтальная линия на *а* соответствует пороговому значению, равному $3.57 \cdot 10^{-6}$. Ось Y на *б* в логарифмическом масштабе

Построение карт рассчитанной статистики

Рассмотрим, как выглядит анализируемая статистика выбросов после перехода от исходных сигналов к приращениям на сети размером 50×50 узлов, покрывающей исследуемый регион Японии. Рассчитаем взвешенную сумму приращений стандартных отклонений для 30-минутных временных рядов, получившихся в результате регистрации вертикальных смещений координат каждой из 1248 станций *GPS* в период с 30.01.–

10.03.2011 г. (40 сут), т.е. до землетрясения Тохоку, и построим усредненные карты W (рис. 4, 5). Значения для каждого узла карты вычислим путем усреднения рассчитанных статистик выбросов для ближайших к узлу станций. Число таких станций (“ближайших соседей”) положим равным 10. Расчет будем проводить в движущемся временном окне, размеры которого составляют 7 сут (336 отсчетов).

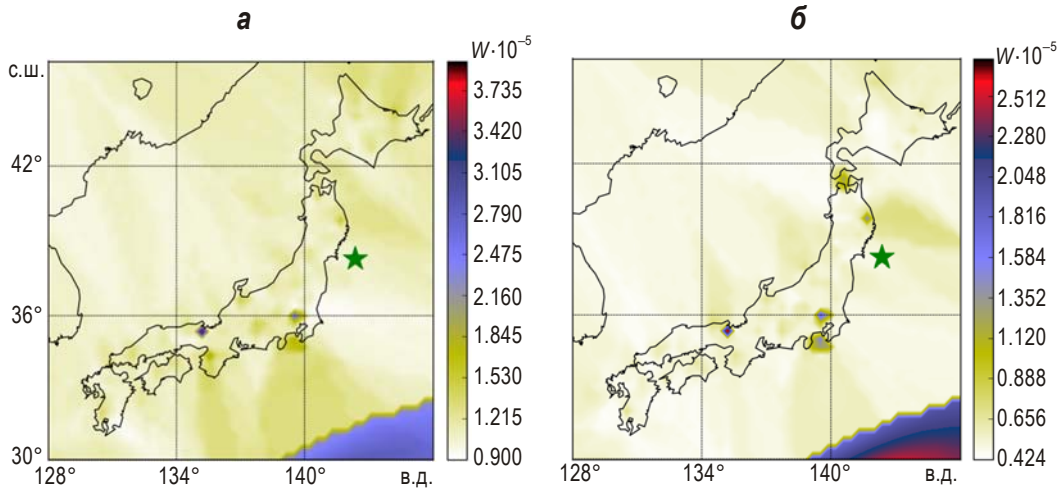

Рис 4. Усредненные карты W вертикальных (*а*) и восточных (*б*) координат, построенные для периода 30.01–10.03.2011 г. Звездочка – эпицентр землетрясения Тохоку 11.03.2011 г.

На картах, представленных на рис. 4, практически невозможно проследить какие-либо явные признаки приближающейся катастрофы. Как говорилось выше, выбросы носят точечный характер, следовательно, основной вклад в усреднение привносят естественные колебания системы. Таким образом, эти карты, не являются информативными для выделения аномальных зон.

Рассчитаем для величины W нормализованную энтропию выбросов, перейдя от самих выбросов к их вероятностям

$$p(t) = \frac{W_t}{\sum_k W_k} \tag{4}$$

и введя нормализованную энтропию согласно формуле

$$En = -\frac{1}{\ln(N-1)} \sum_t p(t) \ln p(t) . \tag{5}$$

В аргументе логарифма стоит N–1, а не N, так как мы перешли к приращениям, уменьшив, таким образом число отсчетов на единицу. В общем случае, когда рассчитывается статистика для исходного сигнала, под логарифмом следует указывать число отсчетов, равное N.

Построим карты нормализованной энтропии для трех компонент сигнала *GPS* (восточной, северной, вертикальной) без использования скользящего временного окна для двух периодов – до землетрясения (30.01.–01.03.2011 г.) Тохоку и после него (12.03.–26.03.2011 г.). Для чистоты эксперимента из периода до землетрясения исключены последние десять дней перед катастрофой, чтобы в анализе не участвовали форшоки, произошедшие в этот промежуток времени.

Чем больше выбросов содержит величина W, тем выше вероятность того, что помимо естественных колебаний в наблюдаемой системе присутствуют другие, порождаемые неизвестным источником.

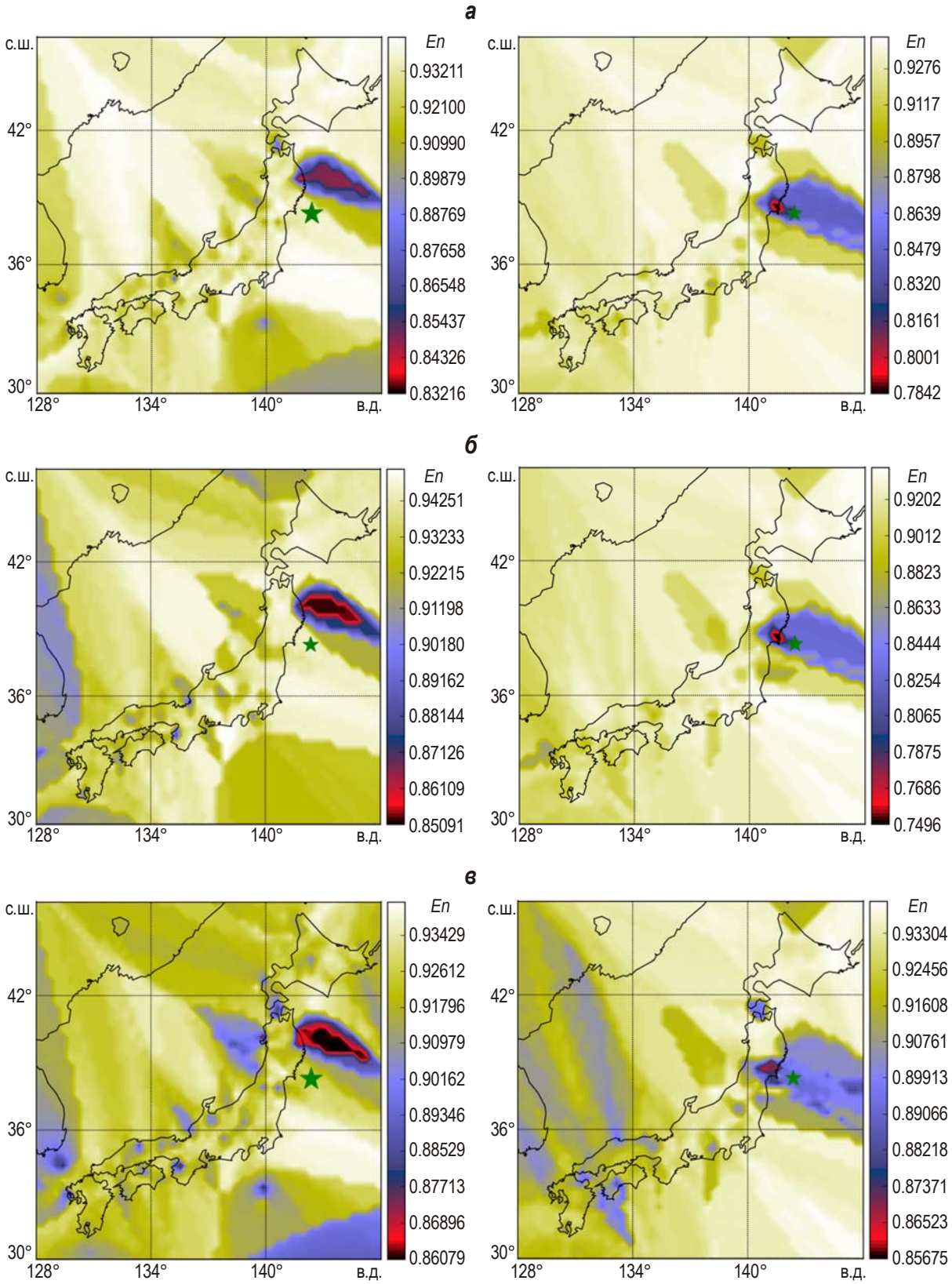

Рис. 5. Карты нормализованной энтропии *En* для восточной (*а*), северной (*б*) и вертикальной (*в*) компонент сигнала *GPS* до землетрясения 30.01.–01.03.2011 г. (*левый столбец*) и после него 12.03.–26.03.2011 г. (*правый*). Звездочка – эпицентр землетрясения Тохоку 11.03.2011 г.

При подготовке землетрясений происходит разрушение пород, находящихся на стыках литосферных плит. Данным событиям может соответствовать как скачкообразное поведение сигнала, так и появление в нем выброса. Основываясь на этом, можно предположить, что увеличение выбросов в сигнале – следствие более активного движения земных блоков. В связи с этим наиболее интересными для исследования представляются области пониженных значений нормализованной энтропии, характерные для временных рядов, содержащих повышенное количество выбросов.

На картах, построенных до землетрясения (см. рис. 5, *левый столбец*), для всех координат видно, что "нормальным" значением *En* является величина, бóльшая 0.9. Однако в области, находящейся в непосредственной близости к эпицентру землетрясения (зеленая звездочка), прослеживаются пониженные значения *En*, т.е. аномальная область тяготеет к эпицентру будущего события. Можно сделать вывод, что появление на картах областей пониженных значений нормализованной энтропии (в нашем случае $En<0.9$) может интерпретироваться в качестве предвестника возможной катастрофы.

На картах, построенных после землетрясения (см. рис. 5, *правый столбец*), зона пониженных значений *En* сместилась в область наиболее сильных толчков. Это подтверждает гипотезу о том, что пониженная нормализованная энтропия статистики *W* отражает нестабильность анализируемой системы. Наблюдаемая аномалия в области эпицентра землетрясения объясняется наличием афтершоков.

Выводы

Предложенный метод оценки меры выбросов, основанный на вычислении взвешенной суммы приращений стандартных отклонений временных рядов сигналов *GPS*, позволил на примере землетрясения Тохоку (Япония, 11.03.2011 г., M_W=9.0) выделить аномальную зону, расположенную в непосредственной близости к эпицентру катастрофы и характеризующуюся пониженными значениями энтропии выбросов. Данный факт можно интерпретировать следующим образом. Подготовка землетрясения происходила именно в этом регионе, а не в области, где расположен эпицентр события, т.е. породы, находящиеся на стыке литосферных плит, начали разрушаться именно там. Подобные разрушения отражаются в виде выбросов во временных рядах сигналов *GPS*, которые в свою очередь понижают значение энтропии.

Литература

Любушин А.А. Анализ данных систем геофизического и экологического мониторинга. М.: Наука, 2007. 228 с.

Любушин А.А. Тренды и ритмы синхронизации мультифрактальных параметров поля низкочастотных микросейсм // Физика Земли. 2009. № 5. С.15–28.

Любушин А.А. Прогностические свойства случайных флуктуаций геофизических характеристик // Биосфера. 2014. № 4. С.319–338.

Любушин А.А., Яковлев П.В., Родионов Е.А. Многомерный анализ параметров флуктуаций GPS сигналов до и после мегаземлетрясения 11 марта 2011 г. в Японии // Геофизические исследования. 2015. Т. 16, № 1. С.14–23.

Любушин А.А., Копылова Г.Н., Касимова В.А., Таранова Л.Н. О свойствах поля низкочастотных шумов, зарегистрированных на Камчатской сети широкополосных сейсмических станций // Вестник Камчатской региональной ассоциации "Учебно-научный центр" (КРАУНЦ). Сер. Науки о Земле. 2015. № 2, вып. 26. С.20–36.

Любушин А.А., Яковлев П.В. Энтропийная мера скачкообразной составляющей временных рядов GPS // Физика Земли. 2016. Т. 52, № 1. С.98–107.

Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. Л.: Энергоатомиздат, 1991. 304 с.

Соболев Г.А., Любушин А.А. Микросейсмические импульсы как предвестники землетрясений // Физика Земли. 2006. № 9. С.5–17.

Соболев Г.А., Любушин А.А., Закржевская Н.А. Асимметричные импульсы, периодичности и синхронизация низкочастотных микросейсм // Вулканология и сейсмология. 2008. № 2. С.135–152.

Соловьев А.А., Агаян С.М., Гвишиани А.Д., Богоутдинов Ш.Р., Шулья А. Распознавание возмущений с заданной морфологией на временных рядах. II. Выбросы на секундных магнитограммах // Физика Земли. 2012. № 5. С.37–52.

Чернецкий В.И. Математическое моделирование стохастических систем. Петрозаводск: Изд-во Петрозаводского государственного университета, 1994. 488 с.

Aggarwal C.C., Yu P.S. Outlier detection for high dimensional data // In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Santa Barbara, California, USA, 2001. P.37–46.

Bay S.D., Schwabacher M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule // In Proceedings of the ninth ACMSIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 2003. P.29–38.

Leung C., Thulasiram R., Bondarenko D. An Efficient System for Detecting Outliers from Financial Time Series // Lecture Notes in Computer Science. 2006. V. 4042. P.190–198. doi: 10.1007/11788911_16.

Lyubushin A. Prognostic properties of low-frequency seismic noise // Natural Science. 2012. V. 4, N 8A. P.659–666. doi: 10.4236/ns.2012.428087.

Lyubushin A., Yakovlev P. Properties of GPS noise at Japan islands before and after Tohoku mega-earthquake // Springer Plus. 2014. V. 3, N 364. doi: 10.1186/2193-1801-3-364, http://www.springerplus.com/content/3/1/364.

Osorio I., Lyubushin A., Sornette D. Automated seizure detection: Unrecognized challenges, unexpected insights. Epilepsy & Behavior, Volume 22, Supplement 1, December 2011. P.S7–S17. The Future of Automated Seizure Detection and Prediction. doi: 10.1016/j.yebeh.2011.09.011.

Sun P., Chawla S., Arunasalam B. Outlier detection in sequential databases // In Proceedings of SIAM International Conference on Data Mining, 2006. P.94–105.

Сведения об авторе

ЯКОВЛЕВ Павел Викторович – аспирант, Российский государственный геологоразведочный университет им. С. Орджоникидзе. 117997, Москва, ул. Миклухо-Маклая, д. 23. Тел.: 8(916) 998-51-25. E-mail: paulyakovlev@gmail.com

ENTROPY MEASURE OF OUTLIERS IN GPS TIME SERIES

P.V. Yakovlev

Ordzhonikidze Russian State Geological Prospecting University, Moscow, Russia

Abstract. The method of detection of significant outliers 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 also maps of normalized entropy that identify 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, revealing the anomaly of low entropy of outliers that appeared before the earthquake is more important result of the analysis carried out.

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