УДК 681.3

**Д. В. Ландэ[1], А. А. Снарский[2]**
Национальный технический университет Украины «КПИ»
проспект Победы, 37, 03056 Киев, Украина
[1]dwl@visti.net; [2]asnarskii@gmail.com

# Диаграмма отклонения элементов ряда измерений от локальных линейных аппроксимаций

*Предложен метод выявления и визуализации трендов, периодичностей, локальных особенностей в рядах измерений ($\Delta L$-метод), базирующийся на технологии DFA (Detrended Fluctuation Analysis). Суть метода состоит в отображении значений абсолютного отклонения точек ряда накопления значений измерений от соответствующих значений линейной аппроксимации. Показано, что метод $\Delta L$ в некоторых случаях позволяет лучше определять локальные особенности, чем вейвлет-анализ. Предложенный простой в реализации подход может применяться при анализе временных рядов в таких областях как экономика и социология.*

***Ключевые слова:*** *ряд измерений, линейная аппроксимация, вейвлет-анализ, Detrended Fluctuation Analysis, $\Delta L$-метод.*

В настоящее время для выявления и визуализации трендов, периодичностей, локальных особенностей в рядах измерений широко применяются методы фрактального и вейвлет-анализа. Один из таких методов — DFA (Detrended Fluctuation Analysis) [1, 2] — используется для выявления статистического самоподобия сигналов. Суть этого метода заключается в следующем. Пусть имеется ряд измерений $x_t$, $t \in 1,...,N$. Обозначим его среднее значение: $\langle x \rangle = \frac{1}{N}\sum_{k=1}^{N} x_k$. Из исходного ряда строится ряд накопления:

$$X_t = \sum_{k=1}^{t} \left( x_k - \langle x \rangle \right).$$

Затем ряд $X_t$ разделяется на временные окна длиной $L$, строится линейная аппроксимация ($L_{j,L}$) по значениям $X_{k,j,L}$ из $X_{j,L}$ внутри каждого окна (в свою очередь, $X_{j,L}$ — подмножество $X_t$, $j = 1,...,J$, $J = N/L$ — количество окон наб-

людения) и рассчитывается отклонение точек ряда накопления от линейной аппроксимации:

$$E(j,L)=\sqrt{\frac{1}{L}\sum_{k=1}^{L}\left(X_{k,j,L}-L_{k,j,L}\right)^2}=\sqrt{\frac{1}{L}\sum_{k=1}^{L}|\,\Delta_{k,j,L}\,|^2}\,,$$

где $L_{k,j,L}$ — значение локальной линейной аппроксимации в точке $t=(j-1)L+k$. Здесь $|\,\Delta_{k,j,L}\,|$ — абсолютное отклонение элемента $X_{k,j,L}$ от локальной линейной аппроксимации.

Далее вычисляется среднее значение:

$$F(L)=\frac{1}{J}\sum_{j=1}^{J}E(j,L)\,,$$

после чего, в случае $F(L)\propto L^{\alpha}$, где $\alpha$ некоторая константа, делаются выводы о наличии статистического самоподобия и характере поведения исследуемого ряда измерений.

Представляет интерес поведение абсолютного отклонения точек ряда накопления от линейной аппроксимации $|\,\Delta_{k,j,L}\,|$ (назовем его $\Delta L$-методом) для реальных процессов, например, отражающих интенсивность публикаций данной тематики в Интернете. Чаще всего временные ряды, соответствующие тематическим информационным потокам, обладают свойствами статистического самоподобия [3], что подтверждается, в частности методом DFA. Визуализация параметров $|\,\Delta_{k,j,L}\,|$ в зависимости от $t=(j-1)L+k$ и $L$ в виде «рельефной» диаграммы представляет собой определенный интерес для изучения локальных особенностей процесса, соответствующего исходному ряду измерений.

Следует заметить, что разделение исходного интервала значений $t\in 1,...,N$ на $J$ непересекающихся окон наблюдения приводит к некоторому «неравноправию» точек внутри этих окон, что не является принципиальным в случае суммирования и последующей приблизительной оценки, но существенно при анализе локальных значений и визуализации. Поэтому, не отказываясь от идеи линейной аппроксимации, предлагается выбирать для каждой точки $t$ такое окно наблюдения длиной $L$, чтобы данная точка оказывалась в его центре (или со смещением в 1 в случае четных $L$). Безусловно, с учетом этой поправки, замедляется скорость вычисления $|\,\Delta_{k,j,L}\,|$, что в значительной мере компенсируется простотой алгоритма.

В качестве исследуемого временного ряда, на котором будем рассматривать возможности метода, будем использовать ряд из посуточного количества публикаций сообщений по определенной тематике в сети Интернет в течение года (рис. 1). Этот ряд получен с помощью системы контент-мониторинга InfoStream, регулярно сканирующей свыше 3000 онлайновых российских и украинских СМИ [4].

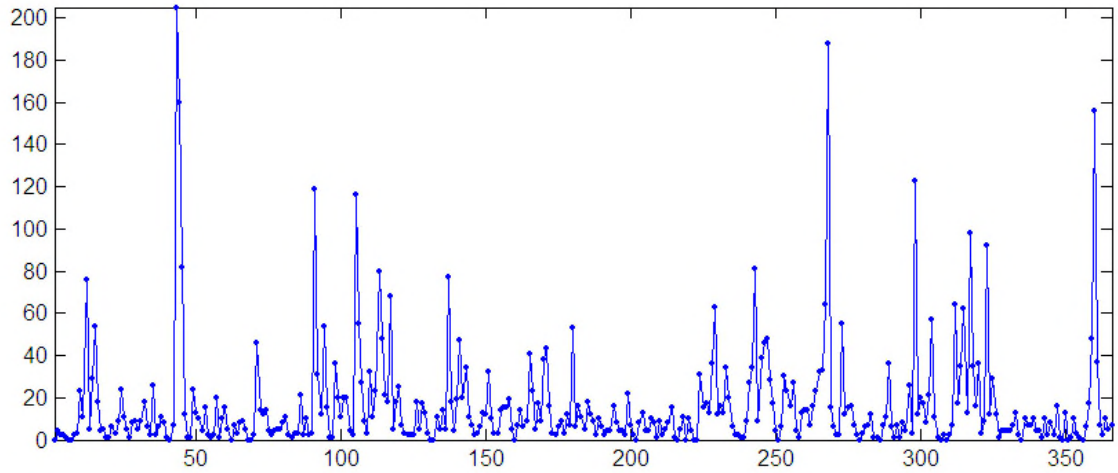



Рис. 1. Временной ряд интенсивностей публикаций по заданной тематике
(ось абсцисс — дни года, ось ординат — количество публикаций)

«Рельефные диаграммы», получаемые в результате предложенного метода (пример такой диаграммы приведен на рис. 2, где более светлые тона соответствуют большим значениям $|\Delta_{k,j,L}|$), напоминают скейлограммы, получаемые в результате непрерывных вейвлет-преобразований. Следует обратить внимание на то, что темные полосы в центре многих областей светлой закраски свидетельствуют о «стабилизации» больших значений рассматриваемого ряда на высоком уровне.

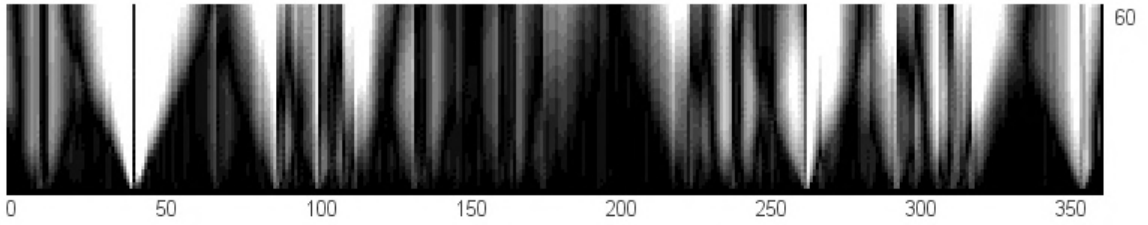



Рис. 2. $\Delta L$-диаграмма временного ряда интенсивности тематических публикаций
(ось абсцисс — дни года, ось ординат — величина окна измерений)

$\Delta L$-метод оказывается достаточно эффективным для выявления гармонических составляющих исследуемого ряда. На рис. 3 показана $\Delta L$-диаграмма ряда, соответствующего синусоиде ($y(i) = \sin(i\pi / 7)$, $i = 1,...,366$). Применение $\Delta L$-метода к ряду, составленному из количества публикаций, сосканированных системой InfoStream из Интернет без учета тематического деления, имеет явно выраженную гармоническую составляющую (общее количество публикаций зависит от дня недели), что можно видеть на рис. 4. Кроме того, на этой диаграмме заметны отклонения от общей динамики объемов публикаций в праздничные дни.

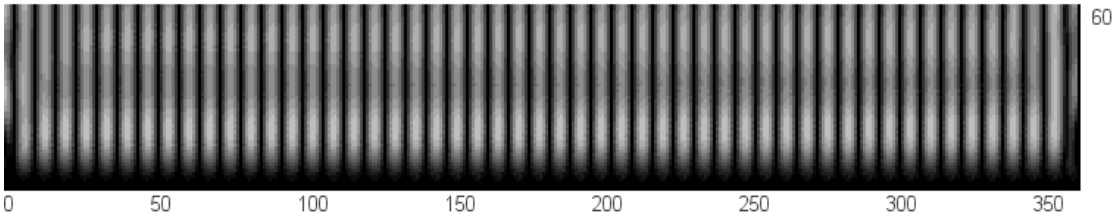



Рис. 3. $\Delta L$ -диаграмма синусоиды

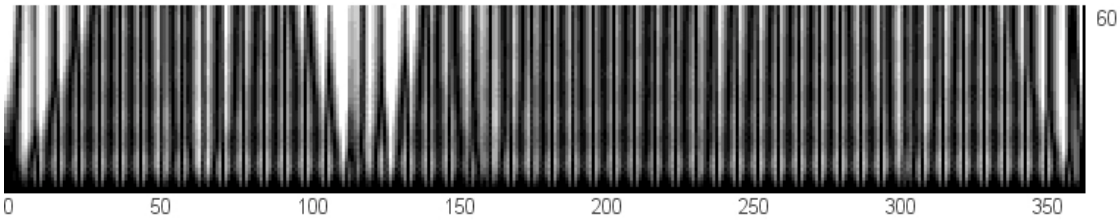



Рис. 4. $\Delta L$ -диаграмма ряда из количества публикаций, сосканированных ежесуточно системой InfoStream в 2008 году

$\Delta L$ -диаграммы внешне похожи на скейлограммы, получаемые в результате вейвлет-анализа рядов измерений. Основная идея вейвлет-преобразований состоит в том, что некоторый числовой ряд, как и в рассмотренном выше методе, разделяется на «окна наблюдения», и на каждом из них генерируется набор коэффициентов, являющихся функциями двух переменных: времени и частоты, и поэтому также могут быть представлены в виде «рельефных» диаграмм, так называемых, скейлограмм. По своей природе вейвлет-коэффициенты представляют собой определенную степень близости исследуемого ряда измерений с некоторой специальной функцией, называемой вейвлетом [5, 6].

Непрерывное вейвлет-преобразование для функции $f(t)$ строится с помощью непрерывных масштабных преобразований и переносов вейвлета $\psi(t)$ с произвольными значениями масштабного коэффициента $a$ и параметра сдвига $b$:

$$W(a,b) = (f(t),\psi(t)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi^*\left(\frac{t-b}{a}\right)dt .$$

На рис. 5 приведена скейлограмма — результат непрерывного вейвлет-анализа (вейвлет Гаусса) временного ряда, который соответствует исследуемому процессу.

Предложенный метод визуализации абсолютных отклонений $\Delta L$, как и метод вейвлет-преобразований, позволяет (и как показано на примере — не хуже) выявлять единичные и нерегулярные «всплески», резкие изменения значений количественных показателей в разные периоды времени. Следует отметить, что метод вейвлет-преобразований может применяться с использованием разнообразных вейвлетов. В частности, применение вейвлета Хаара (рис. 6), по-видимому, более подходит к анализу рассматриваемой последовательности. Однако, даже применение вейвлета Хаара не позволило идентифицировать особенность (локальный

максимум) исходного ряда измерений в последние дни 2008 года, по меньшей мере, эта особенность не показана как скелетон на рис. 6б.

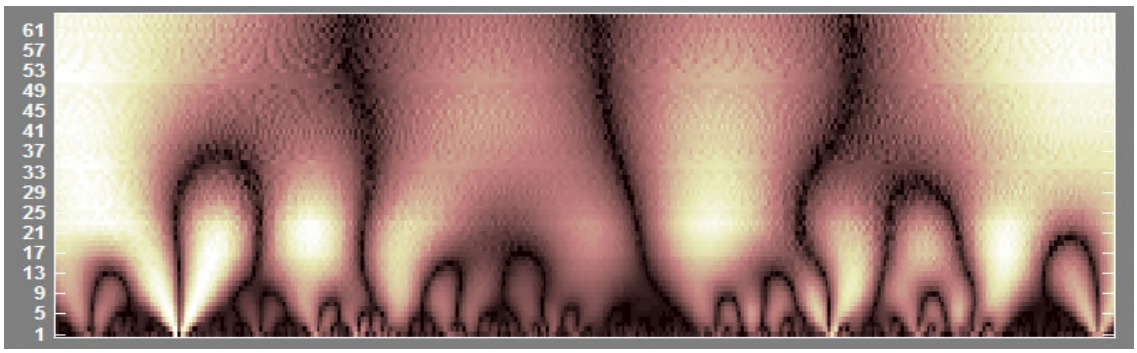
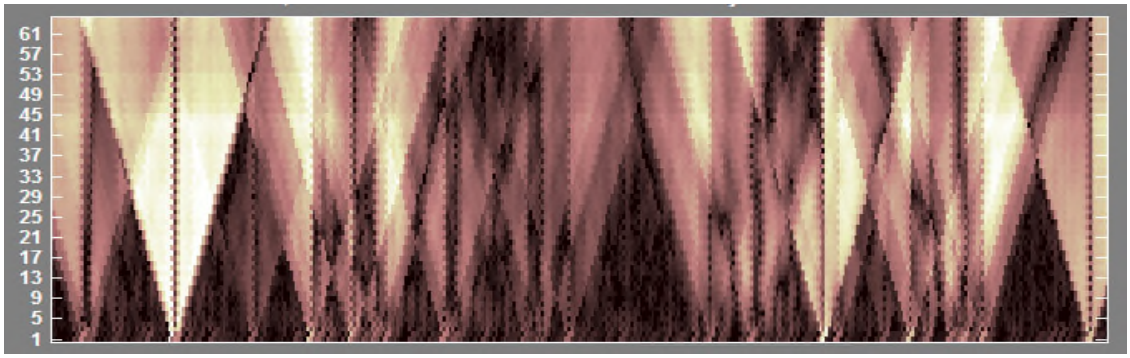
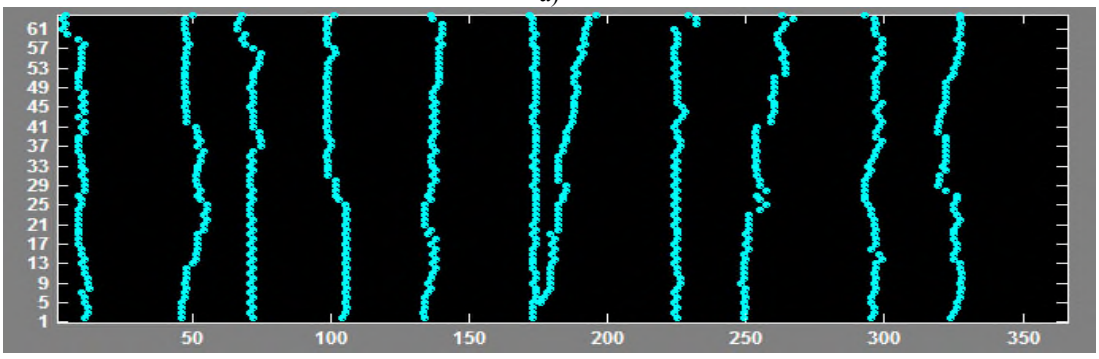



Рис. 5. Скейлограмма временного ряда (вейвлет Гаусса), приведенного на рис. 1




а)




б)

Рис. 6. Скейлограмма временного ряда (вейвлет Хаара), приведенного на рис. 1:
а) скейлограмма (ось абсцисс — день года, ось ординат — частота);
б) линии локальных маскимумов скейлограммы

Метод $\Delta L$, реализация которого существенно проще, тем не менее, позволил определить эту аномалию. Кроме того, выбор подходящего вейвлета для анализа всегда является сложной задачей, которую не требуется решать в случае применения метода $\Delta L$. Предложенный подход достаточно прост в программной реали-

зации и, как показывает опыт, может эффективно применяться при анализе временных рядов в таких областях как экономика и социология.

1. *Peng C.K.* Mosaic Organization of DNA Nucleotides / Peng C.K, Buldyrev S.V, Havlin S, Simons M, Stanley H.E, Goldberger A.L. // Phys. Rev. E. — 1994/ — **49**(2). — P. 1685–1689.

2. *Peng C.K*. Quantification of Scaling Exponents and Crossover Phenomena in Nonstationary Heartbeat Time Series / Peng C.K, Havlin S., Stanley H.E., Goldberger A.L. // Chaos. — 1995. — Vol. 5. — P. 82.

3. *Додонов А.Г.* Самоподобие массивов сетевых публикаций по компьютерной вирусологии / А.Г. Додонов, Д.В. Ландэ // Реєстрація, зберігання і оброб. даних. — 2007. — Т. 9, № 2. — С. 53–60.

4. Система контент-мониторинга InfoStream. URL [Электронный ресурс]: http://infostream.ua/

5. *Астафьева Н.М.* Вейвлет-анализ: основы теории и примеры применения / Н.М. Астафьева // Успехи физических наук. — 1996. — Т. 166, № 11. — С. 1145–1170.

6. *Buckheit J.* Wavelab and Reproducible Research / Buckheit J., Donoho D. // Stanford University Technical Report 474: Wavelets and Statistics Lecture Notes, 1995. — 27 p.

Поступила в редакцию 26.02.2009