

УДК 004.9

А. Н. Грайворонская, Д. В. Ландэ
Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина

Элементы нелинейного анализа информационных потоков

Рассмотрены методы нелинейной динамики, которые применяются для анализа временных рядов, соответствующих информационным потокам в сети Интернет. Большинство из этих методов базируются на корреляционном, фрактальном, мультифрактальном, вейвлет- и Фурье-анализе. Детально описаны особенности этих методов, их взаимосвязь. Представленные методы и соответствующие алгоритмы могут быть использованы для выявления особенностей в динамике развития информационных процессов, выявления периодичностей, аномалий, свойств самоподобия, а также взаимной коррелированности и прогнозирования развития различных информационных процессов. Представленные методы могут быть рассмотрены как основа для выявления информационных атак, кампаний, операций, войн.

Ключевые слова: информационные потоки, временные ряды, нелинейная динамика, корреляционный анализ, фрактальный анализ, прогнозирование.

Введение

Для исследования информационных потоков в Интернете, т.е. потока сообщений, которые публикуются на страницах веб-сайтов, в социальных сетях, блогах и т.п., должен применяться современный инструментарий. Так, известные методы обобщения информационных массивов (классификация, фазовое укрупнение, кластерный анализ и т.д.) уже не всегда пригодны даже для адекватного количественного отражения процессов, происходящих в информационном пространстве [13].

Количественный анализ динамики информационных потоков, которые генерируются в Интернете, становится сегодня одним из наиболее информативных методов исследования актуальности тех или иных тематических направлений. Эта динамика обусловлена разнообразными качественными факторами, многие из которых не поддаются точному описанию. Однако общий характер временной зависимости количества тематических публикаций в сети Интернет все же допускает построение математических моделей, их исследование, прогнозирование. Наблю-

дения временных зависимостей объемов сетевых информационных потоков убедительно свидетельствуют о том, что механизмы их генерации и распространения, очевидно, связаны со сложными нелинейными процессами. Именно этой теме посвящена данная работа.



Рис. 1. Взаимосвязи между подходами к анализу временных рядов

Для анализа временных рядов, которые отображают зависимость объемов информационных потоков от времени, используют разнообразные методы и подходы. При этом оказывается, что все эти подходы взаимосвязаны и более того, ключевую роль играет понятие корреляции. Изложение статьи построено вокруг схемы, показанной на рис. 1, причем особое внимание уделено взаимосвязям.

Формальное описание информационного потока

Для формального описания информационных потоков введем некоторые общие для всего последующего изложения предположения. Дадим определение информационного потока [26], которое корреспондируется с классическим определением в теории информации.

Рассмотрим отрезок действительной оси времени (t_0, t) , где $t > t_0$. Допустим, что на этом отрезке времени согласно некоторым закономерностям в сети публикуется некоторое количество документов — k . Пусть документы публикуются в моменты $\tau_1, \tau_2, \dots, \tau_k$ ($t_0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_k \leq t$).

Информационным потоком будем называть процесс $N_\alpha(t)$, реализация которого характеризуется количеством точек (документов), которые появились в интервале (t_0, t) , как функцию правого конца отрезка t . Согласно этому определению реализация информационного потока является функцией $N_\alpha(t)$. В дальнейших исследованиях мы рассмотрим временные ряды, значения которых как раз и соответствуют значениям ступеней неубывающей ступенчатой целочисленной функции $N_\alpha(t)$, т.е. набор наблюдаемых значений упорядоченных по времени. Далее рассмотрим дискретные временные ряды, значения которых фиксировались через равные промежутки времени. Обозначим такой временной ряд x_1, x_2, \dots, x_T или коротко $\{x_t\}_{t=1}^T$, подразумевая, что фиксирование значений ряда происходило через равный промежуток времени h : $t_0, t_0 + h, t_0 + 2h, \dots, t_0 + (T - 1)h$.

В рамках рассматриваемой проблематики значения временного ряда можно описывать только в терминах вероятностного распределения, т.е. речь идет о статистическом временном ряде. В дальнейшем, анализируя временные ряды, рассмотрим их как реализацию некоторого стохастического процесса.

В качестве примеров в данной работе будут использоваться три временных ряда, которые были получены с помощью популярного сетевого сервиса Google Trends. Эти временные ряды отображают уровень интереса к Дональду Трампу, Хилари Клинтон и информационным атакам «русских хакеров» с августа 2016 года по апрель 2017 года. Временные ряды, получаемые с помощью Google Trends, отображают динамику популярности поискового запроса. Максимальная точка на графике равна 100 и соответствует дате, когда запрос был наиболее популярен, а остальные точки на графике определяются в процентном соотношении к максимуму. Все три временных ряда показаны на рис. 2. Для простоты ссылок на данные ряды в дальнейшем обозначим их T (Д. Трамп), K (Х. Клинтон), X («русские хакеры»).

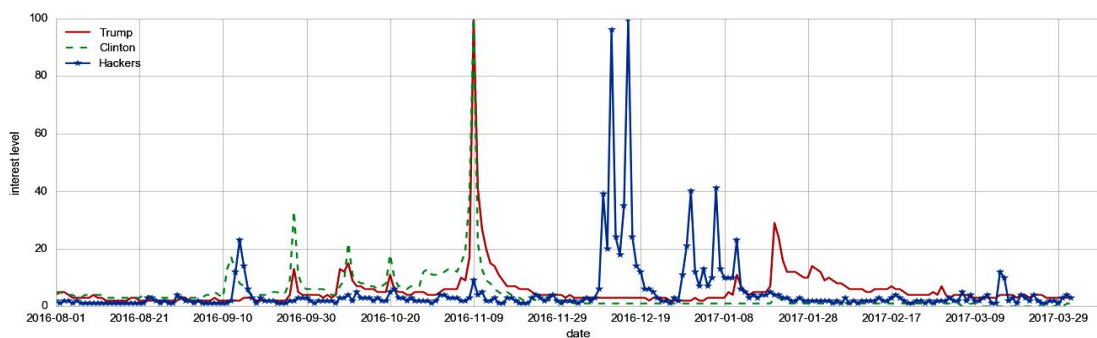


Рис. 2. Временные ряды, отображающие интерес к Дональду Трампу (T), Хилари Клинтон (K) и «русским хакерам» (X) с 1 августа 2016 года по 1 апреля 2017 из Google Trends

Корреляция

Многие методы исследования временных рядов базируются на некотором предположении о статистическом равновесии или постоянстве. Одним из таких полезных предположений является стационарность [4].

Временной ряд называется *строго стационарным* или *стационарным в узком смысле*, если его статистические свойства не изменяются со временем. Формально, если совместное распределение случайных величин $x_t, x_{t+1}, \dots, x_{t+n}$ совпадает с распределением $x_{t+k}, x_{t+k+1}, \dots, x_{t+k+n}$ при любых целых значениях сдвига k , то временной ряд $\{x_t\}_{t=1}^T$ называется строго стационарным. У стационарных временных рядов постоянное математическое ожидание и дисперсия:

$$\mu = Ex_t, \sigma^2 = Var(x_t) = E(x_t - Ex_t)^2.$$

При этом значения μ и σ^2 можно оценить как выборочное среднее и выборочную дисперсию, соответственно:

$$\hat{\mu} = \bar{x} = \frac{1}{T} \sum_{t=1}^T x_t, \quad \hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T (x_t - \bar{x})^2. \quad (1)$$

Свойство стационарности также имеет большое значение при сравнении временных рядов. Линейная зависимость между двумя случайными величинами измеряется ковариацией. Для временных рядов определяют кроссковариацион-

ную функцию. По определению, **кросс-ковариация** с временной задержкой k между случайными процессами $\{x_t\}_{t=1}^T$ и $\{y_t\}_{t=1}^T$ равна:

$$\gamma_{xy}(k, t) = \text{Cov}(x_t, y_{t+k}) = E[(x_t - \mu_x)(y_{t+k} - \mu_y)].$$

Из предположения о стационарности в узком смысле следует, что распределение пар величин x_t, y_{t+k} одинаково для произвольного значения t . Следовательно, ковариация между величинами x_t и y_{t+k} не зависит от t , а зависит только от значения k , то есть $\gamma_{xy}(k, t) = \gamma_{xy}(k), \forall t$. Набор значений $\{\gamma_{xy}(k)\}$ образует кроссковариационную функцию.

Нормировав кроссковариационный коэффициент, получим **кросскорреляционный коэффициент**

$$\rho_{xy}(k) = \frac{\text{Cov}(x_t, y_{t+k})}{\sigma_x \sigma_y} = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y}.$$

Кросскорреляционная функция является мерой подобия между двумя временными рядами.

Чаще всего кроссковариационные и кросскорреляционные коэффициенты оценивают по формулам [18]:

$$\hat{\gamma}_{xy}(k) = \frac{1}{T} \sum_{t=1}^{T-k} (x_t - \bar{x})(y_{t+k} - \bar{y}), \quad \hat{\rho}_{xy}(k) = \frac{\hat{\gamma}_{xy}(k)}{\hat{\gamma}_{xy}(0)}.$$

Для примера приведем оценку кросскорреляционной функции для рядов T и K (рис. 3). Максимальное значение (приблизительно равное 0,8) функция достигает при временной задержке 0. То есть два временных ряда, связанные с интересом к Дональду Трампу и Хилари Клинтон, сильно коррелированы.

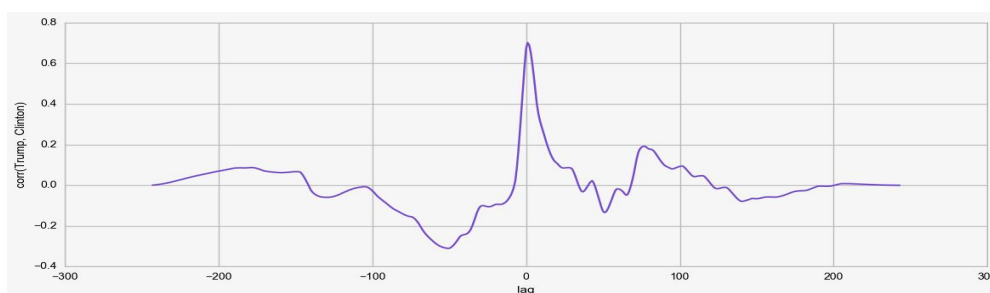


Рис. 3. Корреляционная функция для рядов T и K в зависимости от временной задержки (лага)

Автокорреляция

Можно подсчитать ковариацию не для двух различных рядов, а для одного ряда. Такая ковариация называется **автоковариацией** с временной задержкой или лагом k :

$$\gamma_k = \text{Cov}(x_t, x_{t+k}) = E[(x_t - \mu)(x_{t+k} - \mu)].$$

Набор величин γ_k , $k = 0, 1, 2, \dots$ называется автоковариационной функцией, а их нормированное значение ρ_k , $k = 0, 1, 2, \dots$ — **автокорреляционной функцией**:

$$\rho_k = \frac{E[(x_t - \mu)(x_{t+k} - \mu)]}{\sqrt{E(x_t - \mu)^2 E(x_{t+k} - \mu)^2}} = \frac{Cov(x_t, x_{t+k})}{Var(x_t)} = \frac{\gamma_k}{\gamma_0}.$$

Автокорреляционная функция описывает зависимость между значениями случайного процесса в различные моменты времени [6].

Чаще всего автоковариационные и автокорреляционные коэффициенты оценивают по формулам:

$$\hat{\gamma}_k = \frac{1}{T} \sum_{t=1}^{T-k} (x_t - \bar{x})(x_{t+k} - \bar{x}), \quad \hat{\rho}_{xy}(k) = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}.$$

На рис. 4 показана автокорреляционная функция для ряда T .

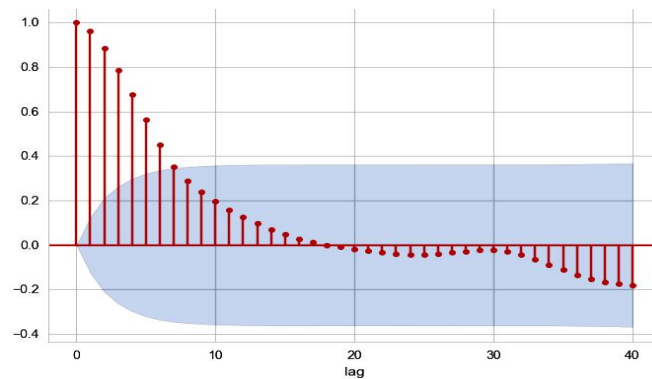


Рис. 4. Автокорреляционная функция временного ряда T

Анализ Фурье

Классический анализ Фурье предоставляет возможность исследовать функцию во временной и частотной областях. Суть перехода в частотную область состоит в том, что функция раскладывается на составляющие, которые являются гармоническими колебаниями с разными частотами. При этом каждой частоте соответствует коэффициент, отображающий амплитуду колебания на данной частоте. Если представить функцию графически во временной области, то получим информацию о том, как функция изменяется со временем. Если изобразить функцию в частотной области, то получим информацию о частотах, колебания на которых она содержит. Для этого используют прямое и обратное преобразование Фурье:

$$\hat{x}(v) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi vt} dt, \quad x(t) = \int_{-\infty}^{\infty} \hat{x}(v)e^{i2\pi vt} dv.$$

На рис. 5,а показан пример функции, которая на самом деле является суммой трех синусоид с разными периодами. Глядя только на график функции во временной области достаточно трудно понять, что она состоит из трех гармонических колебаний и определить их периоды. На рис. 5,б показано преобразование

Фурье для этой функции. Из графика в частотной области наглядно видно, что функция содержит колебания на трех разных частотах.

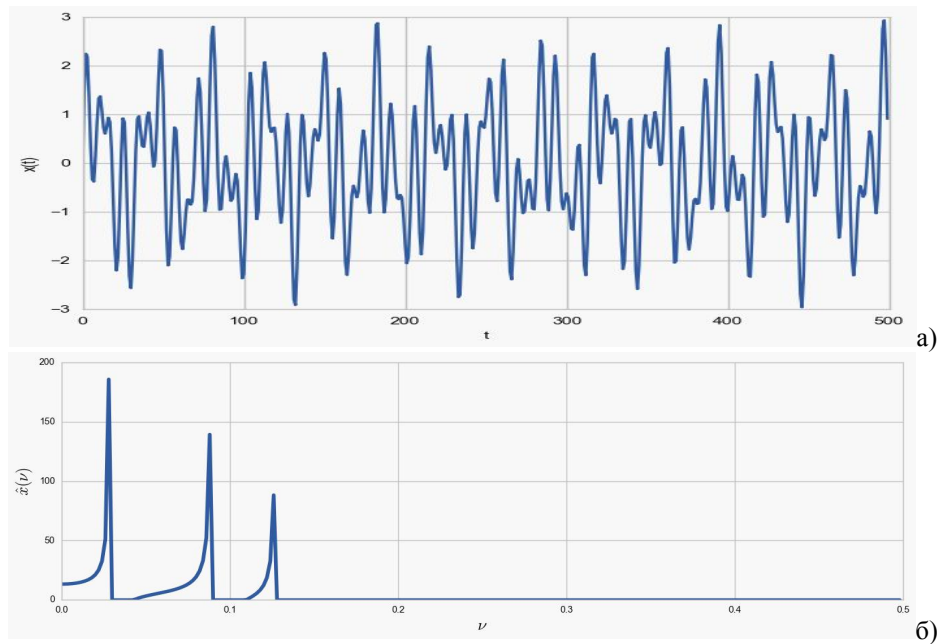


Рис. 5. Функция во временной и частотной областях: а) сумма трех синусоид с различными периодами в зависимости от времени; б) оцененный спектр Фурье

Сегодня преобразование и спектры Фурье находят разнообразные применения в системах машинного обучения. Часто спектры Фурье используются в качестве обучающих параметров. Например, в [21] предложена модель прогнозирования временного ряда, в которой спектр Фурье вместе с некоторыми другими параметрами подается на вход нейронной сети.

Преобразование Фурье можно воспринимать как определение корреляции между исходным сигналом и гармоническими функциями с различными частотами колебания.

Не смотря на свои преимущества и многочисленные приложения, преобразование Фурье является плохим методом для исследования функций, которые эволюционируют со временем. Для таких функций нужен некоторый способ оценивания спектра не по всей длине временного ряда, а по его различным частям. Примером такого подхода является оконное преобразование Гэбора:

$$G(v, \tau, s) = \int_{-\infty}^{\infty} x(t) e^{-\frac{(t-l)^2}{s^2}} e^{-i2\pi vt} dt.$$

Временное окно $e^{-\frac{(t-l)^2}{s^2}}$ выделяет отрезок временного ряда с центром в точке l и имеет ширину, которая определяется параметром s , что позволяет выделить часть исследуемого ряда.

При использовании преобразования Гэбора возникает проблема выбора ширины окна. Сделать оконную функцию зависящей от частоты так, чтобы для низ-

ких частот окно становилось шире, а высоких — уже, позволяет следующий класс преобразований, а именно вейвлет-преобразование, основное преимущество которого состоит в том, что выделенный из временного ряда кусок анализируется с той степенью детальности, которая соответствует его масштабу.

Вейвлеты

Вейвлет-преобразование имеет корреляционную природу. В данном случае рассматривается корреляция исходной функции с функцией вейвлетом на разных масштабах. Для того чтобы такую процедуру всегда можно было выполнить и корреляционные коэффициенты были информативными, вейвлет должен обладать определенными математическими свойствами. Буквально слово вейвлет переводится как «маленькая волна» или «всплеск», и, как следует из названия, вейвлет хорошо локализован во времени. С математической точки зрения, вейвлет — это функция $\psi(t)$, которая удовлетворяет следующим свойствам:

1) функция $\psi(t)$ квадратично интегрируема ($\psi \in L^2(\mathfrak{R})$) или, другими словами, имеет конечную энергию

$$E = \int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty;$$

2) обозначим $\hat{\psi}(\lambda)$ преобразование Фурье от функции $\psi(t)$, тогда:

$$\int_0^{\infty} \frac{|\hat{\psi}(\lambda)|^2}{\lambda} d\lambda < \infty.$$

На рис. 6 показаны примеры вейвлетов, которые часто используются на практике.

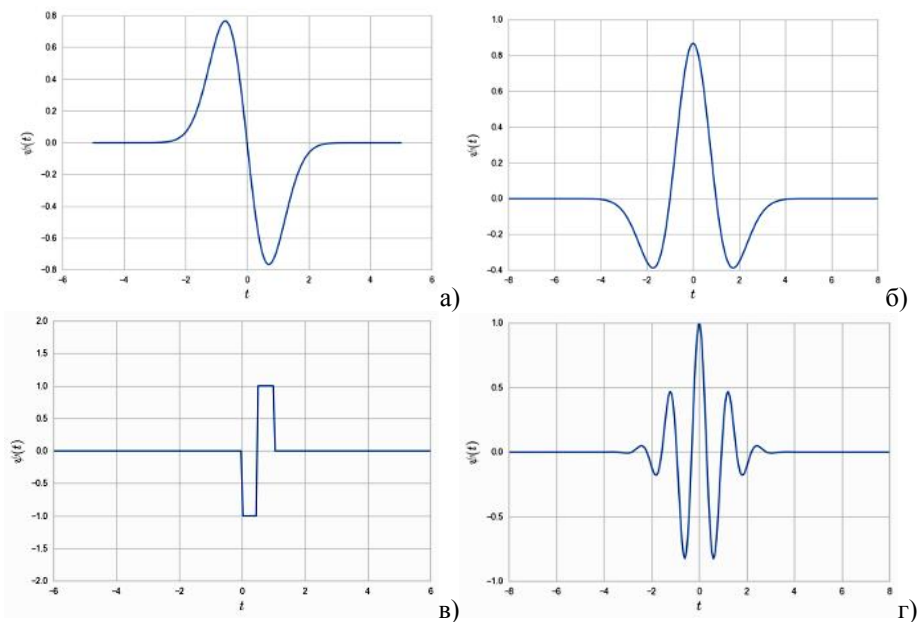


Рис. 6. Примеры вейвлетов, которые часто используются в приложениях: а) гауссова волна; б) мексиканская шляпа; в) вейвлет Хаара; г) вейвлет Морле (действительная часть)

Непрерывное вейвлет-преобразование

Вейвлет $\psi(t)$, свойства которого были описаны выше, часто называют материнским или базовым вейвлетом. На основании материнского вейвлета строят семейство функций с помощью растяжения/сжатия и параллельного переноса. Это необходимо, чтобы исследовать различные области исходного сигнала и с различной степенью детальности.

Введем параметры масштаба s и сдвига l , тогда преобразованная версия материнского вейвлета будет следующей:

$$\psi_{s,l}(t) = \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-l}{s}\right).$$

Непрерывным вейвлет-преобразованием функции $x \in L^2(\mathfrak{R})$ называется выражение

$$W(s,l) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t) \psi^*\left(\frac{t-l}{s}\right) dt = \int_{-\infty}^{\infty} x(t) \psi_{s,l}^*(t) dt,$$

где $l, s \in \mathfrak{R}$, $s \neq 0$; ψ^* — функция комплексно сопряженная с ψ ; величины $\{W(s,l)\}_{l,s \in \mathfrak{R}}$ называются коэффициентами вейвлет-преобразования.

Из формулы в определении непрерывного вейвлет преобразования непосредственно видно, что суть такого преобразование состоит в вычислении корреляционных коэффициентов специального вида.

На рис. 7 показаны коэффициенты вейвлет-преобразования для временных рядов T, K, X .

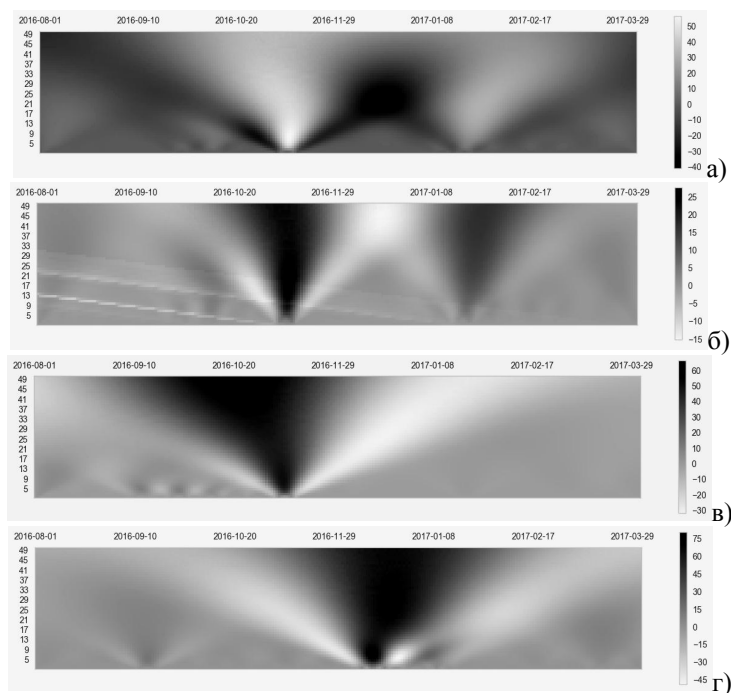


Рис. 7. Вейвлет-коэффициенты: а) ряд T с вейвлетом мексиканская шляпа; б) ряд T с вейвлетом Морле (действительная часть); в) ряд K с вейвлетом мексиканская шляпа; г) ряд X с вейвлетом мексиканская шляпа

Еще раз заметим, что непрерывное вейвлет-преобразование, как и преобразование Фурье можно рассматривать в терминах корреляции. Преобразование Фурье — это корреляция между исходным временным рядом и волной $\varphi(t) = e^{-i2\pi\nu t}$. Волна покрывает всю временную ось и характеризуется только частотой ν , поэтому преобразование Фурье зависит только от частоты. Вейвлет-преобразование — это корреляция между исходным временным рядом и вейвлетом $\psi(t)$. Таким образом, вейвлет-преобразование зависит от положения вейвлета на временной оси и его масштаба, которые определяются параметрами l и s соответственно.

Сравнение временных рядов с помощью вейвлет-преобразования

Рассмотрим некоторые способы сравнения временных рядов с помощью коэффициентов вейвлет-преобразования. Эти способы также можно применять для того, чтобы выявить некоторый тип отношения или взаимосвязи между временными рядами. Метрики сравнения коэффициентов вейвлет преобразования, а также примеры их применения к реальным практическим задачам, подробно описаны в [1].

Рассмотрим два временных ряда x_t и y_t , и обозначим коэффициенты вейвлет-преобразования этих рядов $W_x(s, l)$ и $W_y(s, l)$. Начнем с простейшего способа сравнения — возьмем разность модулей соответствующих коэффициентов:

$$DiffMOD_{x,y}(s, l) = |W_x(s, l)| - |W_y(s, l)|.$$

На рис. 8 показаны значения $DiffMOD_{x,y}(s, l)$ для пар рядов T и K и T и X . Таким простым способом можно выделить области в которых коэффициенты вейвлет-преобразования схожи, а значит и в исходных временных рядах есть похожие участки.

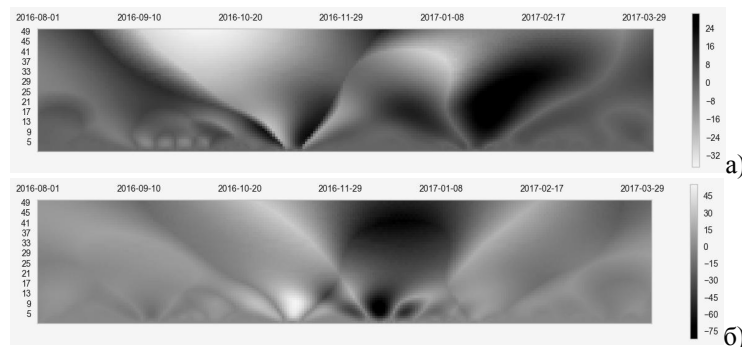


Рис. 8. Значения $DiffMOD_{x,y}(s, l)$ для пары рядов: а) T и K ; б) T и X

Дополнительную информацию можно получить, если использовать комплексный вейвлет (например, вейвлет Морле). Тогда кроме абсолютного значения вейвлет-коэффициентов можно использовать фазу. Комплексный коэффициент всегда можно представить в виде:

$$W(s, l) = |W(s, l)| e^{i\varphi(s, l)}.$$

Следовательно, можно также сравнить фазы коэффициентов:

$$\Delta\varphi_{x,y}(s,l) = \varphi_x(s,l) - \varphi_y(s,l).$$

Кросс-вейвлет-преобразование используется для выделения областей одинаковой энергии между сигналами в области преобразования, а также определения относительной фазы:

$$CrWT_{x,y}(s,l) = W_x^*(s,l)W_y(s,l).$$

На рисунках обычно отображают значение $CrWT_{x,y}(s,l)$, по аналогии со скейлограммой. В таком случае, если временной ряд x идентичен ряду y , то мы получим скейлограмму для ряда x .

Представляет особый интерес вычисление кросс-вейвлет-преобразования в случае, когда используется комплексный вейвлет (например, вейвлет Морле). Тогда

$$\begin{aligned} CrWT_{x,y}(s,l) &= W_x^*(s,l)W_y(s,l) = |W_x(s,l)| e^{i\varphi_x(s,l)} |W_y(s,l)| e^{i\varphi_y(s,l)} = \\ &= |W_x(s,l)||W_y(s,l)| e^{i(\varphi_y(s,l)-\varphi_x(s,l))}. \end{aligned}$$

Таким образом, вычисляя кросс-вейвлет-преобразование, можно извлечь значение разности фаз между коэффициентами вейвлет-преобразования для двух временных рядов.

На рис. 9 показаны значения $CrWT_{x,y}(s,l)$ для рядов T и K (использовался вейвлет мексиканская шляпа). Для рядов T и K выделяется область соответствующая пику интереса во время выборов, что означает, что это область высокой энергии для обоих рядов.

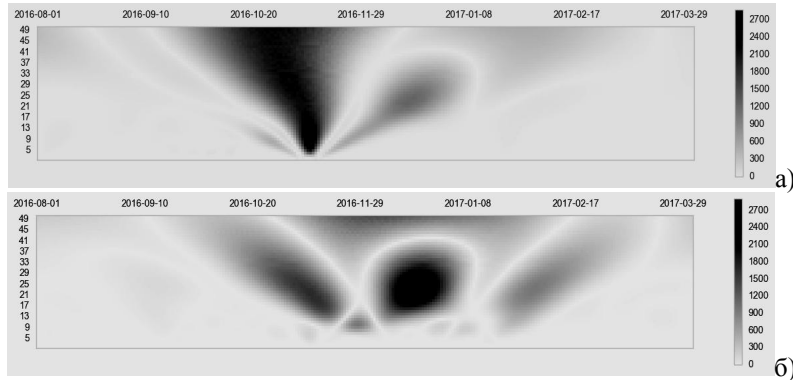


Рис. 9. Кросс-вейвлет-преобразование: а) для рядов T и K ; б) для рядов T и X

Методы кросс-вейвлет-анализа используют при исследовании свойств нескольких временных рядов, зависимых между собой нетривиальным образом. Например, в [2] используются инструменты кросс-вейвлет-анализа, чтобы показать, что связь между переменными денежной политики и макроэкономическими переменными со временем изменилась, причем эти изменения не являются однородными на разных частотах. Данные, полученные с помощью кросс-вейвлет-преобразования, также могут использоваться как исходные данные для алгорит-

мов классификации. В [7] коэффициенты кросс-вейвлет-преобразования подавались на вход искусственной нейронной сети и классификатором Fuzzy.

ΔL -метод

Скейлограммы, полученные с помощью непрерывного вейвлет-преобразования, используют для визуализации особенностей временного ряда. В [14] предложен другой метод визуализации, который также помогает выявить тренды, периодичности и локальные особенности ряда. Предложенный подход значительно проще в реализации, чем вейвлет-анализ.

Метод, который авторы назвали ΔL -метод, базируется на методе DFA (Detrended Fluctuation Analysis), который также будет рассмотрен ниже. Суть подхода состоит в определении и отображении абсолютного отклонения точек ряда накопления значений измерений от соответствующих значений линейной аппроксимации.

Опишем ΔL -метод более подробно. Для начала зафиксируем некоторую ширину окна s (масштаб на котором рассматривается ряд). Рассмотрим точку x_t и выберем для нее окно ширины s так, чтобы точка x_t была в центре этого окна (или смещена на 1, если s четное). Построим линейную аппроксимацию по точкам окна и обозначим $L_{t,j,s}$ значение локальной аппроксимации в точке x_j . Далее вычислим абсолютное отклонение x_t от линии аппроксимации $\Delta_{t,j,s} = |x_j - L_{t,j,s}|$.

Метод предполагает вычисление значений $\Delta_{t,j,s}$ для всех точек $j = 1, \dots, T$ и окон шириной $s = 1, \dots, [T/4]$. Для фиксированной ширины окна вычисляется среднеквадратичное отклонение:

$$E(j, s) = \sqrt{\frac{1}{s} \sum_{t=1}^T |x_j - L_{t,j,s}|^2} = \sqrt{\frac{1}{s} \sum_{t=1}^T \Delta_{t,j,s}^2}.$$

Для полученных значений $E(j, s)$ берется среднее значение по точкам ряда:

$$F(s) = \frac{1}{j} \sum_{j=1}^T E(j, s).$$

Далее полученные значения демонстрируются на диаграмме, похожей на скейлограмму. Пример такой диаграммы показаны на рис. 10.

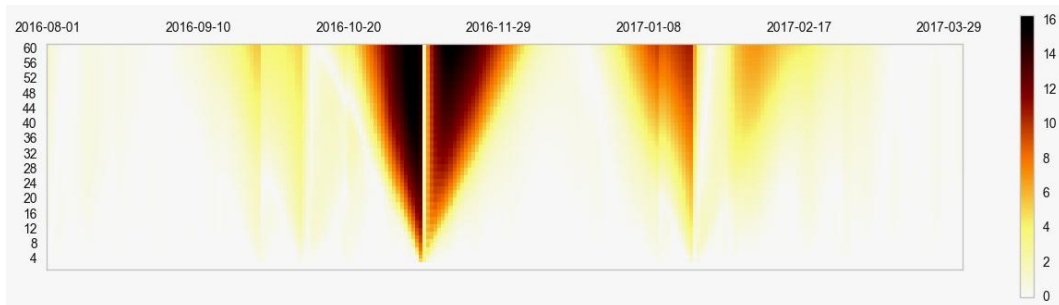


Рис. 10. Коэффициенты полученные с помощью ΔL -метода для ряда T

Предложенный метод визуализации абсолютных отклонений ΔL , как и метод вейвлет-преобразований, позволяет выявлять единичные и нерегулярные «всплески», резкие изменения значений количественных показателей в разные периоды времени, а также гармонические составляющие в ряде.

Фрактальный анализ

Многие объекты в окружающем нас мире статистически самоподобны (классический пример — береговые линии), это означает, что части таких объектов имеют одинаковые статистические характеристики при изменении масштаба. При изучении эволюции информационных потоков, структуры массивов документов в Интернет и исследовании процессов в информационном пространстве часто возникают самоподобные структуры, и в частности временные ряды.

Дадим определение самоподобного процесса.

Действительнозначный процесс $\{x(t)\}_{t \in \mathbb{R}}$ является самоподобным с показателем Херста $H > 0$, если для всех $\alpha > 0$ конечномерные распределения $\{x(\alpha t)\}_{t \in \mathbb{R}}$ идентичны конечномерным распределениям $\{\alpha^H x(t)\}_{t \in \mathbb{R}}$, что можно кратко записать как

$$\{x(\alpha t)\}_{t \in \mathbb{R}} \stackrel{d}{=} \{\alpha^H x(t)\}_{t \in \mathbb{R}}.$$

По определению, для самоподобного процесса изменение временного масштаба эквивалентно изменению масштабу значений процесса. Это означает, что реализации такого процесса выглядят одинаково на разных масштабах. При этом, естественно, что процесс не является точной копией себя на разных масштабах, сохраняются только статистические свойства.

Показатель Херста представляет собой меру персистентности — склонности процесса к трендам. Значение $H = 0,5$ соответствует некоррелированному поведению значений ряда, как у броуновского движения. Значения в диапазоне $0,5 < H < 1$ означают, что направленная в определенную сторону динамика процесса в прошлом, вероятнее всего, повлечет продолжение движения в том же направлении. Если же $H > 0,5$, то прогнозируется, что процесс изменит направленность [5].

Опишем некоторые свойства самоподобных процессов, которые важны для приложений. Во-первых, у таких процессов автоковариационная функция гиперболически затухает и имеет вид

$$\rho_k \approx k^{(2H-2)} L(t) \text{ при } k \rightarrow \infty,$$

где $L(t)$ — медленно меняющаяся на бесконечности функция. Следовательно, для самоподобных процессов ряд из ковариационных коэффициентов расходится:

$$\sum_{k=1}^{\infty} \rho_k = \infty.$$

Такая бесконечная сумма говорит о долговременной зависимости в ряде.

Во-вторых, дисперсия выборочного среднего убывает медленнее, чем величина, обратная размеру выборки $\sigma^2(x_t^{(m)}) \sim m^{2H-2}$, где последовательность $\{x_t^{(m)}\}$

получили, разбив исходную последовательность $\{x_t\}$ на непересекающиеся блоки длины m и взяв среднее в каждом из блоков.

Методы оценивания показателя Херста

Метод оценивания показателя Херста, предложенный им самим, называется методом нормированного размаха или R/S -анализом. Для временного ряда $\{x_t\}_{t=1}^T$ стандартное отклонение S определяется по формуле

$$S = \sqrt{\frac{1}{T} \sum_{t=1}^T (x_t - \bar{x})^2}, \text{ где } \bar{x} = \frac{1}{T} \sum_{t=1}^T x_t,$$

а величина размаха ряда будет:

$$R = \max_{1 \leq t \leq T} x^{(t)} - \min_{1 \leq t \leq T} x^{(t)}, \text{ где } x^{(t)} = \sum_{i=1}^t (x_i - \hat{x}).$$

Отношение R/S и есть нормированным размахом. Для многих наблюдаемых временных рядов нормированный размах хорошо описывается эмпирическим соотношением:

$$\frac{R}{S} = \left(\frac{T}{2}\right)^H.$$

Значения показателя Херста можно оценить, если вычислить значения статистики R/S в зависимости от T и построить график такой зависимости в двойной логарифмической шкале. Оценкой показателя Херста будет оценка наклона прямой, которая наилучшим образом аппроксимирует зависимость $\log R/S$ от $\log t$.

Используем метод R/S для вычисления показателя Херста для рядов T , K и X . Рис. 11 иллюстрирует результаты оценивания для ряда T . Полученные значения показателя Херста — 0,62 для ряда T и 0,68 для ряда K — свидетельствуют о склонности данных процессов к трендам, хотя и не очень высокую.

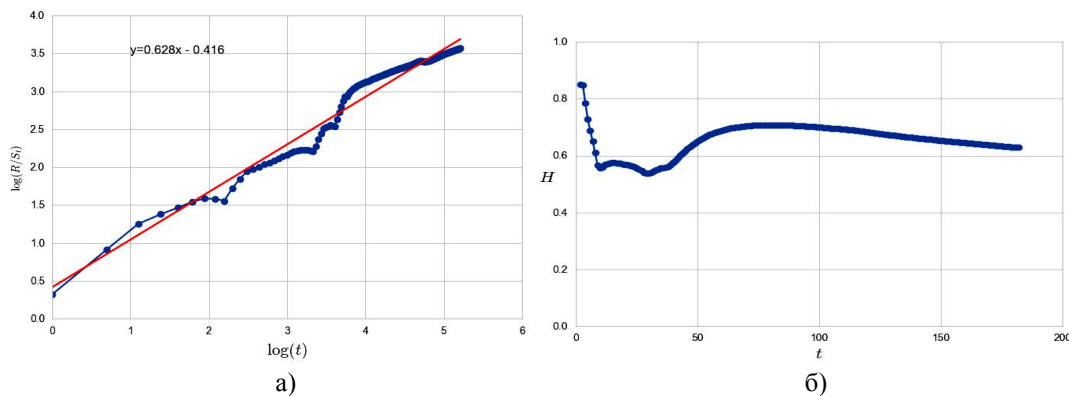


Рис. 11. Оценка показателя Херста для ряда T : а) зависимость статистики R/S от времени в логарифмической шкале; б) зависимость показателя Херста от времени

В случае ряда X на рис. 12,а видно, что зависимость $\log R/S$ от $\log t$ плохо аппроксимируется линейной зависимостью, так как график имеет сильный излом. Если построить зависимость показателя Херста от времени (рис. 12,б), то можно определить момент времени, начиная с которого значение показателя начинает убывать. Отметив этот момент времени на графике временного ряда X (рис. 12,в), становится видно, что это момент резкого возрастания значений ряда, до которого значения ряда имели значительно меньшую дисперсию.

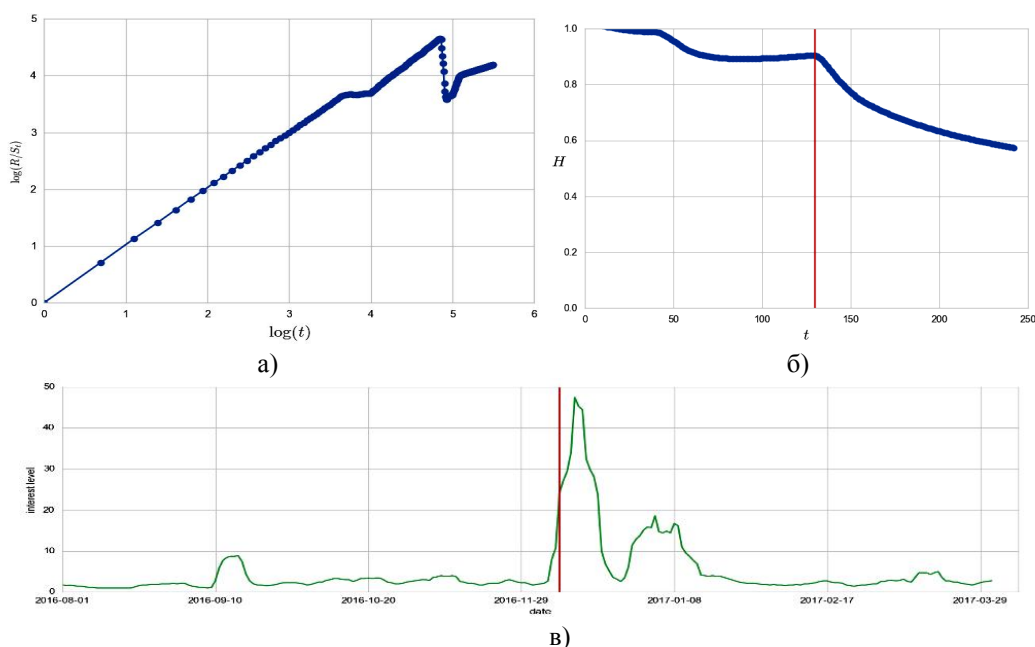


Рис. 12. Оценка показателя Херста для ряда X : а) зависимость статистики R/S от времени в логарифмической шкале; б) зависимость показателя Херста от времени; в) ряд X с отмеченным моментом излома

Поведение ряда X , начиная с начала декабря 2016 (начало наибольшего пика в значениях ряда), можно рассмотреть отдельно. Оценка показателя Херста для второго участка ряда равна 0,7. При этом стоит учесть, что в данном примере временной ряд становится слишком коротким, так как для R/S анализа используют ряды с не менее чем 200 элементами. Тем не менее, резкие изменения в зависимости показателя Херста от времени, которые имеют вид «ступеньки», свидетельствуют о том, что исследуемый процесс состоит из различных процессов, которые имеет смысл рассмотреть отдельно.

Мультифракталы

Для описания самоподобных объектов, которые возникают в природе, часто не хватает одной фрактальной размерности, так как во многих случаях такие объекты не являются однородными. Наиболее общее описание природы таких объектов дает теория мультифракталов, согласно которой объект характеризуется бесконечной иерархией размерностей, что позволяет отличить однородные объекты от неоднородных.

Мультифрактальное множество (сигнал) можно понимать как некое объединение различных однородных фрактальных подмножеств (сигналов), каждое из которых имеет собственное значение фрактальной размерности. Значения таких фрактальных размерностей отображаются в мультифрактальном спектре, формальное определение которого будет введено позже. Важно, что мультифрактальный спектр может использоваться как мера подобия. Такой подход может использоваться, например, для формирования репрезентативных выборок из массивов документов, как дополнение традиционных методов, базирующихся на выявлении содержательного подобия документов. Практические приложения такого подхода: предъявление пользователю обозримых результатов поиска, отражающих весь спектр документального массива или выделение подмножеств документов для дальнейших исследований [14].

Показатель Гельдера и мультифрактальный анализ

Для определения мультифрактального спектра потребуется показатель Гельдера. Он является характеристикой гладкости функции и содержит информацию о поведении функции в окрестности точки. Чем меньше значение показателя Гельдера, тем менее гладкой является функция.

Пусть x — ограниченная функция на \mathfrak{R} и $t_0 \in \mathfrak{R}$, тогда **локальный показатель Гельдера** функции x в точке t_0 определяется как

$$h_x(t_0) = \sup \{ \alpha \geq 0 : |x(t_0 + \Delta t) - x(t_0)| = O(\Delta t^\alpha), \Delta t \rightarrow 0 \}.$$

Другими словами, локальный показатель Гельдера характеризует поведение функции в окрестности точки следующим образом:

$$|x(t + \Delta t) - x(t)| \sim C_t \Delta t^{h_x(t)}.$$

Как было сказано ранее, мультифрактальные объекты не являются однородными, поэтому показатель Гельдера в некоторых точках функции с мультифрактальными свойствами отличается. Значит, имеет смысл рассмотреть множества уровня значений показателя, а именно множества вида

$$E_h = \{t \in \mathfrak{R} : h_x(t) = h\}.$$

Далее можно сравнить размеры множеств E_h при разных значениях h . Во многих практически важных случаях для сравнения таких множеств нужно будет использовать фрактальную размерность [3]. Таким образом, приходим к определению мультифрактального спектра.

Мультифрактальным спектром локально ограниченной функции $x : \mathfrak{R} \rightarrow \mathfrak{R}$ называется отображение

$$d_x(h) = D_H(E_h),$$

то есть с помощью мультифрактального спектра отображается, какие значения показателя Гельдера присутствуют в неоднородном объекте (мере, множестве, сигнале), и в каком соотношении между собой. Каждому значению показателя Гельдера соответствует фрактальная размерность множества точек, в которых значение показателя Гельдера равно данному.

Подход к оцениванию мультифрактального спектра

Выше был описан теоретический подход к определению мультифрактально-го спектра. Для практических целей прямое вычисление показателя Гельдера в каждой точке и вычисление фрактальных размерностей множеств уровня данного показателя не осуществимо. Из определения не сложно вывести следующее утверждение.

Лемма. Пусть $x: \mathfrak{R} \rightarrow \mathfrak{R}$ ограниченная функция, для которой $h_x(t) = H \in [0, 1]$, тогда:

$$h_x(t) = \lim_{j \rightarrow \infty} \left[\frac{\log \left(R_x \left(B(t, 2^{-j}) \right) \right)}{\log 2^{-j}} \right],$$

где $R_x(A) = \max_{t \in A} x(t) - \min_{t \in A} x(t)$ — размах функции x на множестве A , а $B(t, r)$ — одномерный шар с центром в точке t и радиусом r .

Из леммы следует, что мультифрактальный формализм для функций может быть основан на структурной функции

$$Z(q, s) = \frac{1}{s} \sum_t R_x(B(t, s))^q \quad (2)$$

и соответствующей масштабной функции

$$\tau(q) = \lim_{s \rightarrow 0} \left[\frac{\log(Z(q, s))}{\log s} \right], \quad (3)$$

что приводит к определению мультифрактального спектра

$$d_x(h) = \inf_{q \in \mathfrak{R}} (1 - \tau(q) + hq).$$

Выражение для мультифрактального спектра через масштабную функцию уже можно использовать при численном анализе временных рядов. Сначала определяется структурная функция, с помощью нее — масштабная функция, и далее через преобразование Лежандра происходит переход к мультифрактальному спектру [16].

Метод DFA и его применение к оцениванию мультифрактального спектра

В [20] предложен метод Detrended Fluctuation Analysis (DFA) для определения длительных корреляций в зашумленных и нестационарных временных рядах. Ключевая особенность метода DFA состоит в том, что он основан на теории случайных блужданий. Временной ряд не анализируется в исходном виде, вместо этого выполняется центрирование ряда и переход к накопленным суммам:

$$y_t = \sum_{k=1}^t x_k.$$

В таком случае можно рассматривать y_t как положение случайного блуждания после t шагов. Далее метод DFA предполагает анализ среднеквадратического отклонения значений ряда от тренда на различных непересекающихся кусках ряда.

Для метода DFA было предложено множество модификаций, а также вариантов применения для различных практических задач. Обзор таких методов приводится, например в [12]. Важным шагом стала разработка подхода к численному оцениванию мультифрактального спектра на основе метода DFA. Такой метод называется Multifractal Detrended Fluctuation Analysis (MF-DFA) и был предложен в [11]. Эффективность метода MF-DFA была проанализирована для различных модельных временных рядов (броуновское движение, дробное броуновское движение, биномиальные каскады) [19]. Также метод активно используют для анализа реальных временных рядов, часто экономических [22]. Подробное описание алгоритма MF-DFA можно найти в [24].

Использование вейвлетов для оценивания мультифрактального спектра

Для исследования фрактальных характеристик объекта естественным инструментом являются вейвлеты [10] [3]. В первую очередь потому, что точечный показатель Гельдера можно оценить через вейвлет-коэффициенты. Справедливо следующее утверждение, которое связывает вейвлет-преобразования с показателем Гельдера.

Утверждение. Пусть функция x в точке t имеет показатель Гельдера $h_x(t)$. В результате вейвлет-преобразования функции x с вейвлетом ψ получили коэффициенты $W(s,l)$. Предположим также, что в точке t у функции x нет осциллирующей особенности. Тогда

$$W(s,l) \sim s^{h_x(t)} \text{ при } s \rightarrow 0^+, \quad (4)$$

при условии, что у вейвлета ψ первые n моментов равны 0 и $n > h(t_0)$. С другой стороны, если выбрать вейвлет такой, что $n < h(t_0)$, то

$$W(s,l) \sim s^n, \text{ при } s \rightarrow 0^+. \quad (5)$$

Из утверждения следует, что можно охарактеризовать поведение функции x в окрестности точки t следующим образом: чем более гладкой является функция, тем быстрее убывает значение вейвлет-преобразования $W(s,l)$ при уменьшении масштаба s . Например, если функция x является непрерывно дифференцируемой в точке t (это означает, что $h_x(t) = +\infty$), то поведение вейвлет-преобразования при уменьшении масштаба описывается формулой (4). То есть значения $W(s,l)$ зависят только от формы вейвлета. Принципиально другой случай получаем, если функция x в точке t имеет показатель Гельдера в пределах от 0 до 1, что часто встречается на практике. Тогда справедлива формула (5), и такая связь между асимптотическим поведением коэффициентов вейвлет-преобразования и показателям Гельдера лежит в основе методов оценивания мультифрактального спектра с помощью вейвлетов.

Оценивание спектра с помощью линей максимумов. WTMM

Подход к оцениванию мультифрактального спектра с помощью линий максимума активно разрабатывался, начиная с работы [15]. Оказывается, что выражения (4) и (5) будут справедливы, если вместо последовательности вейвлет-преобразований $W(s, l)$ при постоянном значении t_0 и уменьшающемся масштабе s , рассмотреть кривую локальных максимумов модуля $W(s, l)$. Для начала введем необходимые определения.

Точкой максимума модуля (*modulus maximum*) называют точку (s_0, t_0) , такую, что $W(s_0, t) \leq W(s_0, t_0)$, где значение t — это соседние (правое и левое) значения для t_0 , при этом хотя бы для одного значения t (правого или левого) неравенство должно быть строгим $W(s_0, t) < W(s_0, t_0)$. **Линией максимумов** называют связную кривую в пространстве (s, t) , которая состоит из точек максимума модуля. Совокупность линий максимумов, полученных из вейвлет-преобразований функции, называют **скелетом**.

На рис. 13 показано, как выглядит скелетон временного ряда T (с наложением на исходные коэффициенты и без). Такие графики также иногда используют для визуализации.

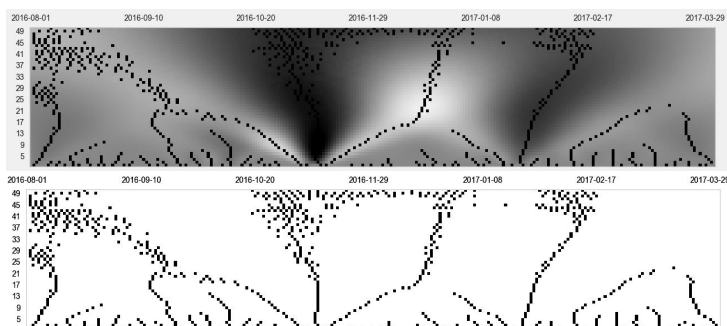


Рис. 13. Скейлтон (линии максимума) для ряда T

Структурная функция через линии максимума имеет вид:

$$Z(q, s) = \sum_{l \in \lambda(s)} \left(\sup_{\substack{(t, s') \in l \\ s' < s}} |W(t, s)| \right)^q, \quad q \in \mathfrak{R}.$$

Далее по формулам (2), (3) можно перейти к масштабной функции и мультифрактальному спектру. На рис. 14 показаны масштабная функция и мультифрактальный спектр для рядов T , K и X . На рис. 14 слева, кроме масштабных функций для исследуемых рядов, показана теоретическая масштабная функция для броуновского движения $\tau(q) = q/2 - 1$.

Мультифрактальный спектр может использоваться для сравнения временных рядов. Мультифрактальные спектры одинаковой формы свидетельствуют о схожести рядов, в то время как различная форма спектра означает, что в природе рядов есть принципиальные различия. Это свойство используется в различных

исследованиях. Например, в [23] статистически анализируется, как варьирование определенных параметров влияет на мультифрактальный спектр для индекса Hang Seng на фондовом рынке Гонконга. Результаты исследований в области экономики, которые основаны на сравнении мультифрактальных спектров, представлены в [25]. Используя пример индекса Dow Jones Industrial Average, выявлены экономические факторы, которые влияют на форму спектра.

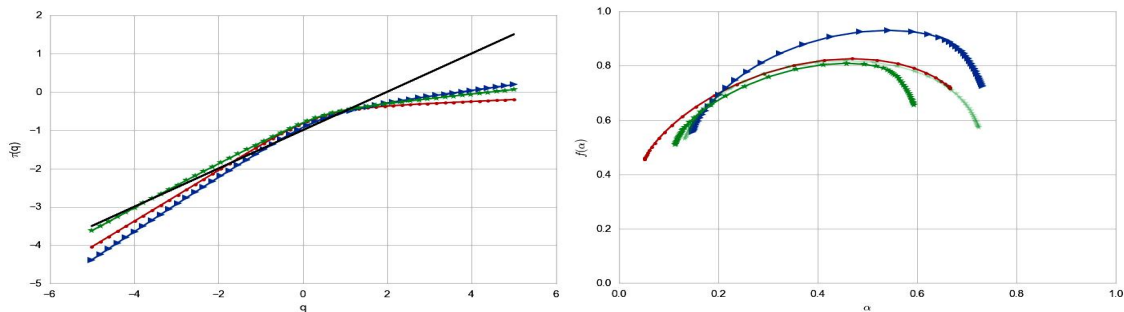


Рис. 14. Масштабная функция для временных рядов T (круглый маркер), K (треугольный маркер) и X (маркер звездочка) и мультифрактальные спектры, оцененные методом WTMM

Выводы

Для эффективного анализа современных информационных процессов на основе мониторинга информационных потоков из глобальных компьютерных сетей должны применяться современные методы, базирующиеся на нелинейном анализе, многие из которых нашли успешное применение в естественных науках. Современные подходы позволяют применять для анализа и моделирования даже общественных и информационных систем методы, апробированные в первую очередь в естественных науках. Анализ информационных потоков выступает фундаментом таких направлений как моделирование, проектирование и прогнозирование. Рассмотренные в работе подходы позволяют описывать информационные процессы, процессы информационного влияния, они пригодны для описания общих тенденций в динамике информационных процессов. При этом продвижение в освоении современного информационного пространства невозможно без общих представлений о структуре и свойствах динамики сетевых информационных процессов, что в свою очередь требует выявления и учета их устойчивых закономерностей. Следует отметить, что часто подходы, которые базируются на применении точных методов и математическом формализме, а также методов компьютерного моделирования, в действительности, могут давать преимущественно качественные выводы, что обуславливается многопараметричностью реальных моделей. Вместе с тем, даже такие результаты могут объяснить реальность во многих случаях лучше, чем традиционные качественные методы.

Методы, алгоритмы, аналитические инструменты, которые рассматриваются в данной работе, выступают не только в качестве демонстрационной основы для объяснения реально происходящих событий и процессов, но и как необходимые компоненты при их планировании и прогнозировании.

Исследование проведено в рамках проекта Ф73/23558 «Разработка методов и средств поддержки принятия решений при выявлении информационных операций». Проект является победителем конкурса Ф73 на грантовую поддержку научно-исследовательских проектов Государственного фонда фундаментальных исследований Украины и Белорусского республиканского фонда фундаментальных исследований.

1. Addison Paul S. The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2016. 446 p.
2. Aguiar-Conraria L., Azevedob L., Soares M.J. Using Wavelets to Decompose the Time-Frequency Effects of Monetary Policy. *Physica A: Statistical Mechanics and its Applications*. 2008. Vol. 387. Issue 12. P. 2863–2878.
3. Aldroubi A., Cabrelli C., Jaffard S., Molter U. New Trends in Applied Harmonic Analysis: Sparse Representations, Compressed Sensing, and Multifractal Analysis. Birkhäuser Basel, 2016. 334 p.
4. Box G.E.P., Jenkins G.M., Reinsel G.C., Ljung G.M. Time Series Analysis, Forecasting and Control. New Jersey: John Wiley & Sons, 2015. 712 p.
5. Braichevsky S., Lande D., Snarskii A. On the fractal nature of mutual relevance sequences in the Internet news message flows. arXiv preprint arXiv: 0710.0228, 2007.
6. Chatfield C. The analysis of time series: an introduction 6th ed. Chapman & Hall/CRC, 2004. 333 p.
7. Dey D., Chatterjee B., Chakravorti S., Munshi S. Cross-wavelet Transform as a New Paradigm for Feature Extraction from Noisy Partial Discharge Pulses. *Transactions on Dielectrics and Electrical Insulation*. 2010. Vol. 17. N 1. P. 157–166.
8. Feder J. Fractals. Springer Science + Business Media, LLC, 1988. 305 p.
9. Harte D. Multifractals. Theory and applications. Chapman and Hall/CRC, 2001. 264 p.
10. Jaffard S. Wavelet Techniques in Multifractal Analysis. Fractal Geometry and Applications: a Jubilee of Benoit Mandelbrot. *Multifractals, Probability and Statistical Mechanincs, Applications*. San Diego. 2004. Vol. 72. Part 2. P. 91–151.
11. Kantelhardt, Jan W., Zschiegner Stephan A., Koscielny-Bunde Eva et al. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications*. 2002. Vol. 316. P. 87–114.
12. Kantelhardt Jan W. Fractal and multifractal time teries. Encyclopedia of Complexity and Systems Science. Springer, 2009. P. 3754–3779.
13. Lande D., Braichevski S. Busch D..Informationsflusse im Internet. IWP — Information Wissenschaft & Praxis, 2007. Heft 5. S. 277–284.
14. Lande D.V., Snarskii A.A. Diagram of measurement series elements deviation from local linear approximations. arXiv preprint arXiv:0903.3328, 2009.
15. Mallat S., Hwang L.W. Singularity Detection and Processing with. *IEEE Transactions on Information Theory*. 1992. Vol. 38. N 2. P. 617–643.
16. Mallat S. A Wavelet Tour of Signal Processing The Sparse Way. Academic Press, 2009. 805 p.
17. Mandelbrot B.B. The fractal geometry of nature. W. H. Freeman and Company, 1982. 468 p.
18. Montgomery Douglas C., Jennings Cheryl L. Introduction to time series analysis and forecasting. New Jersey: John Wiley & Sons, 2008. 441 p.

19. Oswiecimka P., Drozd S., Kwapien J., Gorski A.Z. Effect of detrending on multifractal characteristics. preprint arXiv:1212.0354, 2012.
20. Peng C.-K., Buldyrev S.V., Havlin S. et al. Mosaic organization of DNA nucleotides. *Physical Review E*. 1994. Vol. 49 E. 2.
21. Rodriguez N., Bravo G., Rodriguez N., Barba L. Haar Wavelet Neural Network for Multi-step-ahead. *Anchovy Catches Forecasting. Polibits*. 2014. Issue 50. P. 49–53.
22. Suarez-Garcia P., Gomez-Ullate D. Multifractality and long memory of a financial index. eprint arXiv:1306.0490, 2013.
23. Sun X., Chen H.P., Wu Z.Q., Yuan Y.Z. Multifractal analysis of Hang Seng index in Hong Kong stock market. *A-Statistical Mechanics and its Applications*. 2001. Vol. 291. Issue 1–4. P. 553–562.
24. Thompson James R., Wilson James R. Multifractal Detrended Fluctuation Analysis: Practical Applications to Financial Time Series. *Mathematics and Computers in Simulation*. 2016. Vol. 126. Issue C. P. 63–88.
25. Zhou W.-X. The components of empirical multifractality in financial returns. *EPL*, 2009. Vol. 88. Issue 2. Article Number 28004.
26. Додонов О.Г., Ланде Д.В., Путятін В.Г. Інформаційні потоки в глобальних комп'ютерних мережах. Київ: Наук. думка, 2009. 295 с.
27. Снарский А.А., Ландэ Д.В., Брайчевский С.М., Дармохвал А. Распределение документов по степени релевантности на основе мультифрактальных свойств. Интернет-математика 2007: сборник работ участников конкурса. Екатеринбург: Из-во Уральского университета, 2007. 224 с.
28. Шелухин О.И. Мультифракталы. Инфокоммуникационные приложения. Москва: Горячая линия – Телеком, 2011. 576 с.

Поступила в редакцию 28.08.2017