



ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 11 за 2013 г.



Кириллов Д.С., Короб О.В.,
Митин Н.А., Орлов Ю.Н.,
Плешаков Р.В.

Распределения показателя
Херста нестационарного
маркированного временного
ряда

Рекомендуемая форма библиографической ссылки: Распределения показателя Херста нестационарного маркированного временного ряда / Д.С.Кириллов [и др.] // Препринты ИПМ им. М.В.Келдыша. 2013. № 11. 16 с. URL: <http://library.keldysh.ru/preprint.asp?id=2013-11>

**Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Российской академии наук**

**Д.С. Кириллов, О.В. Короб, Н.А. Митин,
Ю.Н. Орлов, Р.В. Плешаков**

**Распределения показателя Херста
нестационарного маркированного
временного ряда**

Москва — 2013

Кириллов Д.С., Короб О.В., Митин Н.А., Орлов Ю.Н., Плешаков Р.В.

Распределения показателя Херста нестационарного маркированного временного ряда

На примере тикового ряда цен на фьючерсы WTI исследована зависимость показателя Херста этого ряда от длины выборки и времени. Предложено несколько вариантов обобщения этого показателя на неэквидистантные временные ряды. Построены распределения показателя Херста для выборок разных длин и показано, что они являются стационарными в пределах точности, с которой определены соответствующие эмпирические вероятности, и близки к нормальным распределениям.

Ключевые слова: показатель Херста, маркированный случайный процесс, стационарные распределения

Kirillov D.S., Korob O.V., Mitin N.A., Orlov Yu.N., Pleshakov R.V.

On the stationary distributions of the Hurst indicator for the non-stationary marked time series

The tick time-series of WTI futures is considered. The dependence of Hurst indicator on the set length and moment of time is examined. Several variants of generalization of this factor on the marked time series are suggested. The distributions of Hurst indicator are constructed. These distributions are stationary and normal with the accuracy, according to the empirical probabilities definition.

Key words: Hurst indicator, marked time series, stationary distributions

Работа выполнена при поддержке грантов РФФИ, проекты

№ 11-06-00471, № 13-01-00617

Оглавление

Введение	3
1. Определение выборочного показателя Херста	4
2. Нестационарный поток событий	7
3. Показатель Херста для нестационарного маркированного процесса	9
4. Уровень нестационарности тикового ряда и поток событий	10
5. Распределение показателя Херста	13
Литература	16

Введение

В настоящей работе исследуется возможность обобщения классического показателя Херста на неэквиливантные временные ряды. Процессом Херста или процессом фрактальной диффузии (см., напр. [1, 2]) называется эквиливантный временной ряд, в котором накопленный размах отклонений от скользящего среднего значения, нормированный на скользящее стандартное отклонение, зависит от длины выборки накопления показательным образом. Соответствующий показатель и называется показателем Херста. Такой процесс является удобной идеализацией некоторых реально происходящих нестационарных случайных процессов, хотя в чистом виде на практике он не встречается (как, впрочем, и другая популярная идеализация – белый шум). Для каждого временного ряда можно вычислить его показатель Херста как коэффициент регрессии логарифма нормированного накопленного размаха к логарифму длины выборки. Детерминация такой регрессии покажет, с какой точностью исследуемый процесс можно аппроксимировать процессом Херста.

Традиционная трактовка показателя Херста H заключается в том, что при превышении значения $H = 0,5$ накопленный размах растет быстрее, чем для случайного блуждания, т.е. ряд с большей вероятностью сохраняет тенденцию своего изменения по выборке той длины, для которой подсчитывался этот показатель, а если $H < 0,5$, то с большей вероятностью тенденция переменится. Поэтому показатель Херста часто используется при анализе финансовых рынков для оценки длительности трендов или для оценки длины выборки, по которой следует вычислять скользящие средние [3-4].

Имея целью приложение данного исследования к биржевым рядам, отметим, что основным временным рядом, который изучается при проведении статистического анализа того или иного финансового инструмента, является тиковый ряд, т.е. ряд цен сделок, маркированный моментом времени, в который сделка состоялась. Такие ряды не эквиливантны по времени, поэтому номер события не является естественным моментом времени. В результате возникает проблема анализа нестационарного временного ряда, каковым является просто последовательный ряд цен на финансовый инструмент, но еще и маркированного случайным и тоже нестационарным рядом моментов времени, в который произошло наблюдаемое событие. Необходимость такого анализа состоит в том, что статистические методы исследования оперируют с выборками данных, т.е. со значениями случайных величин безотносительно к маркерам, тогда как выводы из проведенного анализа часто требуется делать не в терминах номера события, а в терминах реального времени, что обусловлено спецификой функционирования изучаемых систем. Требуется, например, составить прогноз биржевых цен на тот или иной финансовый инструмент на момент открытия (закрытия) торгов. Если же изучать временной ряд по выборкам, содержащимся в определенных временных промежутках, то получаемые оценки априори не будут статистически эквивалентными,

поскольку в этих промежутках будет содержаться различное, вообще говоря, число событий.

Кроме того, в задачах такого типа часто наблюдается корреляция между значениями ряда $x(n)$ и интенсивностью потока событий, т.е. числом событий в единицу времени. Это означает, что выборки, взятые из промежутков времени с существенно различающейся интенсивностью, могут оказаться зависимыми. Следовательно, при проведении статистического анализа таких рядов необходимо, во-первых, сравнивать статистические свойства распределений, которые построены в условиях близких интенсивностей потока событий, и, во-вторых, уметь переводить те или иные выборочные статистики, отнесенные к номеру события, в функции времени. Такой статистикой, изучаемой в настоящей работе, и является выборочный показатель Херста.

1. Определение выборочного показателя Херста

Рассматривается одномерный временной ряд $x(n)$, где номер n некоторого элементарного события отвечает случайному моменту времени $t(n)$, что параметрически определяет функцию действительного переменного $x(t)$ на некотором дискретном множестве точек – теоретически счетном, но на практике конечном. В качестве ряда $x(n)$ обычно берется ряд первых разностей изучаемого процесса $x(n) = c(n+1) - c(n)$ или, как вариант, ряд логарифмических приростов $x(n) = \ln(c(n+1)/c(n))$.

Показатель Херста для ряда $c(n)$ определяется следующим образом [5]. Вводится скользящее среднее приростов $x(n)$ на шаге n по выборке длины k

$$\bar{x}(n,k) = \frac{1}{k} \sum_{i=n-k+1}^n x(i) \quad (1)$$

и вычисляется накопленное отклонение от среднего для выборки длины k :

$$R(n,k) = \max_{j \leq n} \left(\sum_{i=n-k+1}^j (x(i) - \bar{x}(n,k)) \right) - \min_{j \leq n} \left(\sum_{i=n-k+1}^j (x(i) - \bar{x}(n,k)) \right). \quad (2)$$

Вычисляется также скользящая дисперсия рассматриваемого временного ряда по выборке длины k :

$$\sigma_x^2(n,k) = \frac{1}{k} \sum_{i=n-k+1}^n (x(i) - \bar{x}(n,k))^2, \quad (3)$$

после чего вычисляется логарифм отношения размаха к шуму:

$$\xi(n,k) = \ln \left(\frac{R(n,k)}{\sigma_x(n,k)} \right). \quad (4)$$

Фиксируется некоторая минимальная длина выборки k_0 и максимальная длина N , которая отвечает временному горизонту анализа. Описанная процедура (1-4) делается для выборок длин k таких, что $1 < k_0 \leq k \leq N$.

Показатель Херста $H_N(n)$ по выборке длины N на шаге n определяется как коэффициент регрессии величины $\xi(n, k)$ на логарифм длины выборки:

$$\xi(n, k) - \bar{\xi}_N(n) = H_N(n) \cdot \left(\ln k - \frac{\ln(N!/(k_0 - 1)!)}{N - k_0 + 1} \right) + \varepsilon(k), \quad (5)$$

$$\bar{\xi}_N(n) = \frac{1}{N - k_0 + 1} \sum_{k=k_0}^N \xi(n, k),$$

где $\varepsilon(k)$ представляет собой остаток регрессии. Угол наклона прямой регрессии и равен показателю Херста ряда. Таким образом, полагая для краткости $k_0 = 2$ (минимальная длина выборки), имеем по формулам регрессионного анализа

$$H_N(n) = \frac{\sum_{k=2}^N (\xi(n, k) - \bar{\xi}_N(n)) \left(\ln k - \frac{\ln N!}{N - 1} \right)}{\sum_{m=2}^N \left(\ln m - \frac{\ln N!}{N - 1} \right)^2}. \quad (6)$$

Поскольку длина выборки $N \gg 1$, то в формуле (6) можно использовать формулу Стирлинга для аппроксимации факториала, и в пренебрежении отличием $N - 1$ от N получить более компактное выражение

$$H_N(n) = \frac{1}{N} \sum_{k=1}^N (\xi(n, k) - \bar{\xi}_N(n)) (1 + \ln(k/N)). \quad (7)$$

В результате при фиксированной длине выборки N получаем набор показателей Херста на последовательности шагов с номерами n . Для маркированного же временного ряда можно ввести несколько аналогов показателя Херста, совпадающих в случае эквидистантного ряда, но имеющих различное вероятностное содержание в общем случае.

Во-первых, показатель Херста, зависящий от продолжительности τ времени накопления размаха, можно определить непосредственно по выборке, заключенной внутри соответствующего промежутка. Для этого зададим промежуток времени T , внутри которого вычисляется показатель Херста, и снабдим номер события n указанием на момент времени t , когда это событие произошло. Время T естественно измерять в единицах времени агрегирования данных, поскольку наблюдаемый процесс не является непрерывным в строгом смысле этого понятия. Предполагается, что единичный шаг по времени выбран таким, чтобы эмпирическая вероятность наличия в нем минимального числа событий была бы отлична от нуля. Пусть, например, единицей агрегирования по времени будет 1 минута. Тогда величина τ , как и в случае с длиной выборки, будет целочисленной величиной, изменяющейся от $\tau_{\min} = 1$ до T . Формально определения всех величин (1-4) остаются прежними, но в этих

формулах следует положить $n = n(t)$, а $k = k(\tau)$ есть число событий за промежутки τ , отсчитанный назад от момента t . Логарифм накопленного размаха в новых переменных обозначим $\eta(t, \tau) = \xi(n(t), k(\tau))$. После этого временной показатель Херста $h_T(t)$ определяется как коэффициент в регрессионном приближении

$$\eta(t, \tau) - \bar{\eta}_T(t) = h_T(t) \cdot \left(\ln \tau - \frac{\ln(T!)}{T} \right), \quad \bar{\eta}_T(t) = \frac{1}{T} \sum_{\tau=1}^T \eta(t, \tau). \quad (8)$$

Недостаток определения (8) состоит в том, что в разные моменты времени t в промежутках одинаковой длительности содержится различное число событий, иногда различающееся на порядки, а тогда коэффициенты $h_T(t)$, строго говоря, не сопоставимы.

Второй вариант анализа накопленного размаха как функции времени состоит в усреднении регрессии (5) по интенсивности событий. Поскольку промежутки времен наступления k -ых событий случайны, естественно ввести вероятность того, что в промежутке времени τ произошло ровно k событий. Например, для стационарного пуассоновского потока событий с интенсивностью μ эта вероятность дается формулой

$$p_k(\tau) = \frac{(\mu\tau)^k}{k!} \exp(-\mu\tau). \quad (9)$$

Введем среднюю по распределению (9) величину накопленного размаха за время τ с учетом перенормировки на число событий, начинающееся с двух,

$$S(t, \tau) = \frac{\sum_{k=2}^{\infty} p_k(\tau) \xi(n(t), k)}{1 - (1 - \mu\tau) \exp(-\mu\tau)}, \quad (10)$$

и усредним регрессионное соотношение (5) по распределению (9). Тогда получаем

$$S(t, \tau) - \bar{S}_T(t) = \tilde{h}_T(t) \cdot (a(\tau) - \bar{a}_T), \quad (11)$$

где

$$\bar{S}_T(t) = \frac{1}{T} \sum_{\tau=1}^T S(t, \tau), \quad a(\tau) = \frac{\sum_{k=2}^{\infty} p_k(\tau) \ln(k)}{1 - (1 - \mu\tau) \exp(-\mu\tau)}, \quad \bar{a}_T = \frac{1}{T} \sum_{\tau=1}^T a(\tau). \quad (12)$$

Величина $\tilde{h}_T(t)$ является коэффициентом регрессии на модифицированное логарифмическое время $a(\tau)$, связанное со средневзвешенным логарифмом числа событий. Рассматривая вместо величины $a(\tau)$ величину $b(\tau) = \exp(a(\tau))$, получаем, что аналогом длины выборки в терминах времени является величина $b(\tau)$, т.е. рассматривается регрессия среднего логарифма нормированного размаха S на $\ln b$. Такой подход отвечает рассмотрению процесса во времени, которое течет с разной скоростью в

зависимости от интенсивности потока событий. Однако это может быть не очень удобно, если требуется анализировать динамику в реальном времени.

Третий вариант определения аналога показателя Херста состоит в пересчете показателя $H_N(n)$ с использованием вероятности числа событий, произошедших за время T . В соответствии с (9) полагаем, что

$$\hat{h}_T(t) = \frac{\sum_{k=N_{\min}}^{N_{\max}} p_k(T) H_k(n(t))}{\sum_{m=N_{\min}}^{N_{\max}} p_m(T)}, \quad (13)$$

где N_{\min}, N_{\max} представляют диапазон, в котором заключено число событий за интересующее нас время T .

2. Нестационарный поток событий

Как следует из (9), интенсивность потока событий μ – это среднее число событий в единицу времени (например, тик/мин). Для описания нестационарного потока вводится величина $p_k(t - \tau, t)$ как вероятность того, что за промежуток времени $\Delta_t(\tau) = [t - \tau; t]$ произошло ровно k событий. Тогда интенсивность потока $\mu(t - \tau, t)$ на промежутке $\Delta_t(\tau)$ определяется формулой

$$\mu(t - \tau, t) = \frac{1}{\tau} \sum_{k=1}^{\infty} k p_k(t - \tau, t). \quad (14)$$

Если события независимы, то вероятность $p_k(t - \tau, t)$ дается нестационарным распределением Пуассона (Хинчин, [6]):

$$p_k(t - \tau, t) = \frac{(\Lambda(t, \tau))^k}{k!} \exp(-\Lambda(t, \tau)), \quad \Lambda(t, \tau) = \tau \mu(t - \tau, t). \quad (15)$$

Величина $\Lambda(t, \tau)$ называется параметром потока. В стационарном случае $\mu = const$ и $\Lambda(t, \tau) = \mu \tau$ – это среднее число событий за время τ , отсчитанное от момента t назад. Следовательно, параметр потока равен среднему числу событий на промежутке $\Delta_t(\tau)$.

Далее считаем, что параметр потока нам известен. На практике соответствующая зависимость $\Lambda(t, \tau)$ строится непосредственно по наблюдениям. Конкретный пример мы приведем для среднего параметра потока, который определяется следующим образом. В условиях естественной суточной периодичности, характерной для многих рядов, связанных с торговыми операциями, оценка параметра потока получается усреднением суточных профилей активности. Для определенности будем считать единицу агрегирования минутой. Подчеркнем, что мы рассматриваем поток событий, принципиально не являющийся непрерывным во времени, поэтому на практике обязательно приходится вводить некоторую единицу агрегирования данных, отличную от единицы измерения времени. В нашем случае единица измерения – секунда, а единица агрегирования – минута. Введем следующие обозначения.

Пусть $n_m(j)$ есть число событий в m -ю минуту j -ых суток. Тогда число событий за эти сутки есть

$$N(j) = \sum_{m=1}^{1440} n_m(j), \quad (16)$$

а суточный нормированный профиль интенсивности определяется по формуле

$$y_m(j) = \frac{n_m(j)}{N(j)}. \quad (17)$$

После этого среднее число событий (т.е. собственно параметр потока) на промежутке $\Delta_t(\tau)$ может быть введено через построение средневзвешенного профиля суточной активности за некоторый период в T суток. Для этого вводится средневзвешенная интенсивность $w(m)$ в m -ю минуту и среднее число тиков за сутки W :

$$w(m) = \frac{1}{T} \sum_{j=1}^T n_m(j), \quad W = \sum_{m=1}^{1440} w(m) = \frac{1}{T} \sum_{j=1}^T N(j). \quad (18)$$

Затем определяется средневзвешенный нормированный профиль активности:

$$z(m) = \frac{w(m)}{W}. \quad (19)$$

Пусть, таким образом, ориентировочное число событий за сутки определено и равно W . Тогда среднее число тиков за промежутки τ (мин) до момента времени t (мин) есть

$$\Lambda^{(w)}(t, \tau) = \sum_{m=1}^{\tau} w(t-m+1) = W \sum_{m=1}^{\tau} z(t-m+1). \quad (20)$$

Поскольку профиль $z(m)$ нормирован на единицу, то его можно рассматривать как вероятность интенсивности по минутам в сутках. Его функция распределения есть

$$Z(m) = \sum_{k=1}^m z(k), \quad (21)$$

так что

$$\Lambda^{(w)}(t, \tau) = W \cdot (Z(t) - Z(t - \tau)) = \sum_{k=0}^{\tau-1} \Lambda^{(w)}(t - k, 1). \quad (22)$$

Тем самым определен модельный параметр потока событий для любых моментов времени t и промежутков τ .

Описанный подход к формированию потока событий не является единственно возможным. Можно, например, брать средневзвешенный поток или прогнозировать его по авторегрессионной модели или иным путем. Для наших целей важно считать, что параметр потока откуда-либо известен.

3. Показатель Херста для нестационарного маркированного процесса

Мы будем считать, что единица агрегирования (в нашем случае 1 минута) выбрана так, что в любой момент времени в течение этой единицы произошло хотя бы одно событие. Тогда при любом фиксированном начальном моменте t функция $\Lambda(t, \tau)$ является строго возрастающей по τ , так что к ней существует обратная также строго возрастающая функция $\theta(\Lambda, t)$, равная промежутку времени τ до момента времени t , в течение которого произошло Λ событий.

Поскольку наша цель состоит в том, чтобы связать накопленный нормированный размах за некоторый промежуток времени с самим этим промежутком времени, то формально в формулах (1-3) следует заменить длину выборки k на $\Lambda(t, \tau)$, а суммирование в них проводить до $n(t)$, где $n(t)$ есть номер события, ближайшего слева к текущему моменту времени t . Но в формуле (5) регрессию следует проводить не на логарифм длины выборки, а на логарифм промежутка времени, т.е. на $\ln \tau$, где τ меняется от 1 до T в единицах агрегирования. В нестационарном случае величина T зависит от момента времени t , поскольку для сопоставимости результатов вычисления показателя Херста в разные моменты времени число данных, участвующих в вычислении, должно быть одинаковым. Следовательно,

$$T(t) = \Lambda^{-1}(N, t) \equiv \theta(N, t). \quad (23)$$

Тогда показатель Херста (6) может быть переопределен следующим образом:

$$h'_T(t) = \frac{\sum_{\tau=1}^T (\xi'(t, \tau) - \bar{\xi}'_T(t)) \left(\ln \tau - \frac{\ln T!}{T} \right)}{\sum_{\tau=1}^T \left(\ln \tau - \frac{\ln T!}{T} \right)^2}, \quad T = T(t), \quad (24)$$

а новый нормированный размах ξ' вычисляется по формулам (1-3), в которых сделаны соответствующие пересчеты от номеров событий к моментам времени. Например, для вычисления нового скользящего среднего применяется формула

$$\bar{x}'(t, \tau) = \frac{1}{\Lambda} \sum_{i=n-\Lambda+1}^n x(i), \quad n = n(t), \quad \Lambda = \Lambda(t, \tau), \quad (25)$$

и аналогичные замены делаются в формуле (2) для дисперсии и в формуле (3) для размаха. Средний же нормированный размах в (24) вычисляется по формуле

$$\bar{\xi}'_T(t) = \frac{1}{T} \sum_{\tau=1}^T \xi'(t, \tau), \quad T = T(t). \quad (26)$$

Формула (24) корректирует формулу (8) в том плане, что теперь показатели Херста $h'_T(t)$ в разные моменты времени статистически сопоставимы. Обобщением же формулы (13) на нестационарный случай является формула

$$\hat{h}_T(t) = \frac{\sum_{k=N_{\min}}^{N_{\max}} p_k(t-T, t) H_k(n(t))}{\sum_{m=N_{\min}}^{N_{\max}} p_m(t-T, t)}. \quad (27)$$

В отличие от (26), в формуле (27) T не зависит от времени и является свободным параметром задачи. Аналогично обобщается и вариант, даваемый формулами (9-11).

4. Уровень нестационарности тикового ряда и поток событий

Установим сначала, каков уровень нестационарности анализируемого тикового ряда. Рассмотрим расстояние $\rho_N(n)$ между двумя встык-выборками одинаковой длины N в норме $L1$ по выбранному разбиению гистограммы эмпирических частот, где n есть номер последнего события. Выберем естественное разбиение, шаг которого равен единичному приросту цены на 1 цент. Пусть $f_N(i, n)$ – эмпирическая частота попадания прироста цены в i -ую ячейку разбиения для выборки длины N , отсчитанной в прошлое от текущего номера n . Тогда положим

$$\rho_N(n) = \sum_i |f_N(i, n-N) - f_N(i, n)|. \quad (28)$$

Чтобы с достаточной уверенностью говорить о том, что расстояние между выборками слишком большое, и ряд не может считаться стационарным, следует определить точность, с которой данная гистограмма, построенная по конечной выборке, приближала бы распределение генеральной совокупности в предположении стационарности ряда. Если считать, что уровень значимости ε такого приближения равен расстоянию между гистограммами – теоретической и эмпирической, то допустимый уровень $\varepsilon_N(n)$ флуктуаций эмпирических частот в предположении стационарности ряда находится из соотношения [7]

$$\frac{t_{1-\varepsilon}(N-1)}{\varepsilon} = \frac{\sqrt{N-1}}{\sum_i \sqrt{f_N(i, n) \cdot (1 - f_N(i, n))}}, \quad (29)$$

где $t_\gamma(N)$ есть γ -квантиль распределения Стьюдента с N степенями свободы. Если оказалось, что $\rho_N(n) > 2\varepsilon_N(n)$, то две выборки следует признать нестационарными. Чтобы сделать аналогичный вывод относительно фрагмента изучаемого ряда в целом, надо рассмотреть эмпирическое распределение $g_N(\rho)$ расстояний (28) и найти такое $\rho^*(N)$, для которого доля расстояний, меньших $\rho^*(N)$, равна дополнению доли отрезка длины $\rho^*(N)$ в носителе распределения $g_N(\rho)$, т.е.

$$\int_0^{\rho^*} g_N(\rho) d\rho = 1 - \rho^* / 2. \quad (30)$$

Величина $\rho^*(N)$, называемая согласованным уровнем стационарности (СУС) ряда, является основной характеристикой нестационарности. Построим затем аналогичный уровень для точности определения гистограммы $\varepsilon^*(N)$. Если $\rho^*(N) > 2\varepsilon^*(N)$, то ряд на выборках длины N является нестационарным, а если

$$\rho^*(N) \leq 2\varepsilon^*(N), \quad (31)$$

его следует признать стационарным на уровне значимости, равном точности оценки эмпирических частот.

Далее рассматривается конкретный пример нестационарного (и в смысле ряда, и в смысле потока событий) маркированного временного ряда фьючерсов WTI. Пример распределений для этого ряда двух встык-выборок длины 10 тыс. тиков каждая приведен на рис. 1.

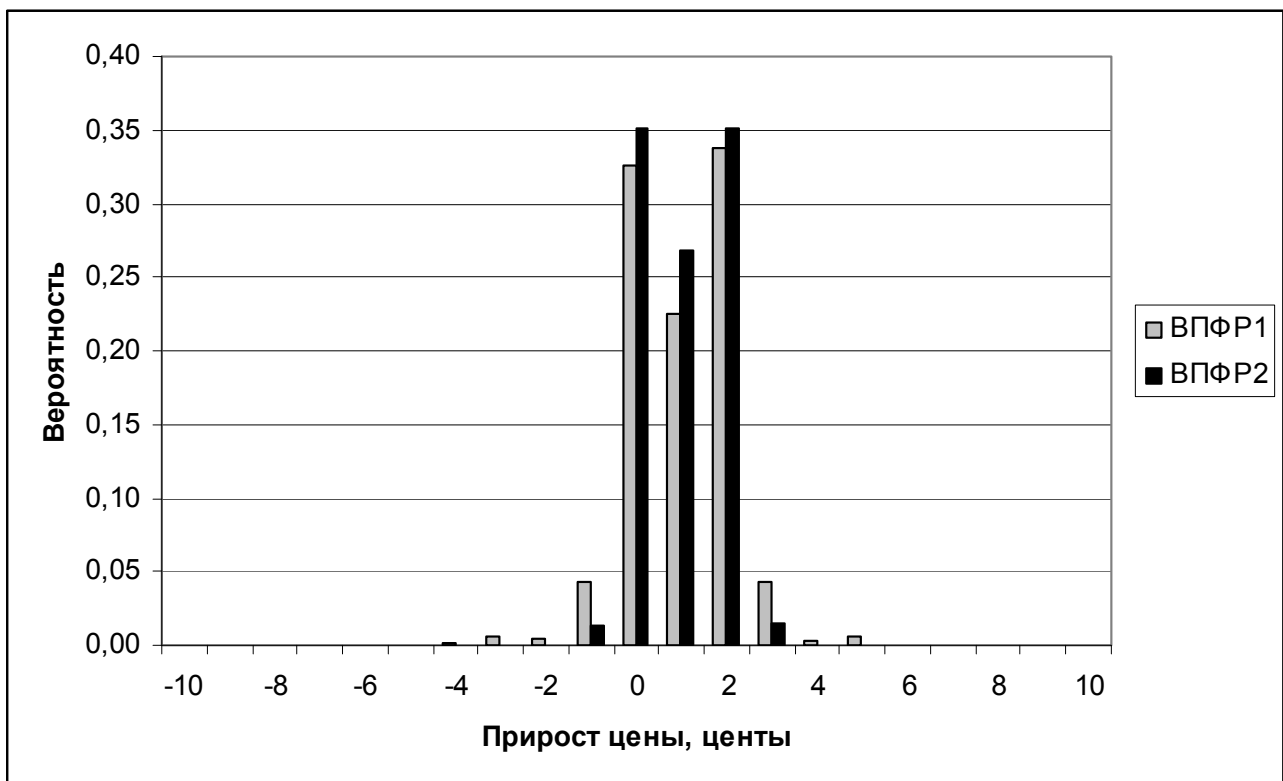


Рис. 1. Выборочные встык-распределения приростов тикового ряда

В данном примере расстояние между распределениями равно 0,16, тогда как точность оценки вероятностей на гистограмме, в соответствии с (29), равна 0,045. В соответствии с (31) выборки следует считать нестационарными на уровне значимости 0,09. На рис. 2 приведен СУС тикового ряда в зависимости от длины выборки и в сравнении с точностью оценки эмпирических частот.

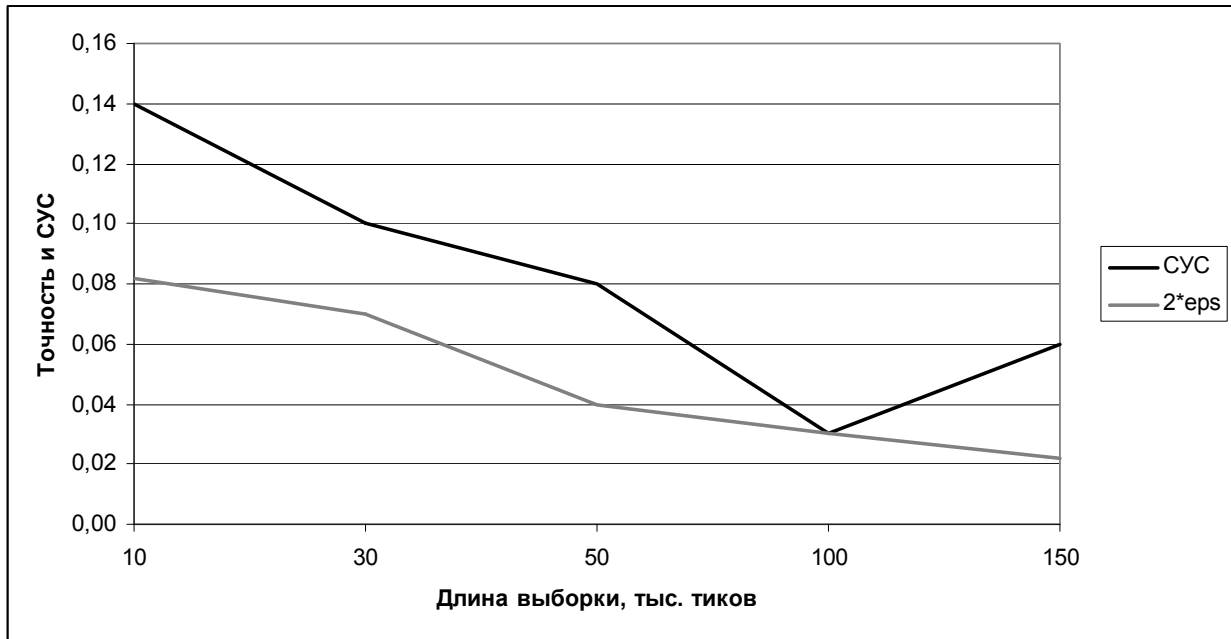


Рис. 2. Согласованные уровни стационарности и точности

Из графиков рис. 2 следует важный вывод о свойствах данного тикового ряда. На коротких выборках этот ряд нестационарный. С увеличением длины выборки он становится все более стационарным, и на длине 100 тыс. его СУС сравнивается с точностью оценки расстояния между эмпирическими распределениями. Подчеркнем, что эта точность является естественным пределом уровня значимости, на котором имеет смысл задавать вопрос о стационарности для данного ряда, поэтому ряд на длине 100 тыс. следует считать стационарным. Уровень значимости для такого утверждения равен 0,03. Однако при дальнейшем увеличении длины выборки СУС начинает возрастать, что отличает поведение данного ряда от действительно стационарного, когда с увеличением длины выборки СУС становится меньше, чем точность оценки эмпирических вероятностей.

Что касается потока событий, то он имеет четкую суточную периодичность. Средневзвешенная минутная интенсивность $w(m)$, определяемая формулой (18), показана на рис. 3. Этот профиль фактически является параметром потока $\Lambda(m,1)$ в течение единицы агрегирования (1 мин). Для рассматриваемого примера число тиков в сутках варьируется от 38 тыс. до 93 тыс., среднесуточное значение – примерно 70 тыс. тиков. Из предыдущего анализа тогда следует, что характерный временной промежуток, на котором тиковый ряд ведет себя квазистационарно, составляет около полутора суток. Поэтому на промежутке менее суток можно строить торговые системы, использующие индикаторы нестационарных разладок, а на промежутке более двух суток выборка будет содержать разнородные данные, что приведет к большому проценту ошибочных выводов статистики.

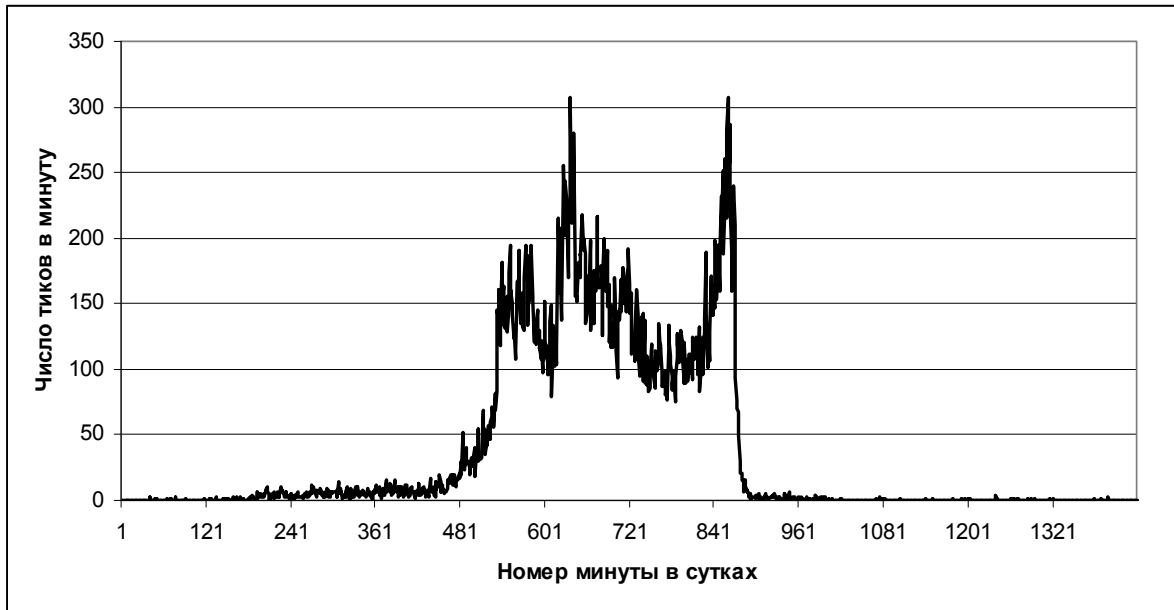


Рис. 3. Средневзвешенный минутный параметр потока за сутки

Из рис. 3 видно, что параметр потока в разные моменты времени различается более чем на два порядка, поэтому различные нестационарные обобщения показателя Херста на маркированные ряды будут давать результаты, сильно отличающиеся от расчетов для эквидистантных рядов.

5. Распределение показателя Херста

Показатель Херста $H_N(n)$ тикового ряда в терминах номеров событий, вычисленный по формулам (1-4) в окне $N = 20$ тыс. тиков, показан на рис. 4. На этом графике отсчет начинается с первого значения после выборки в 20 тыс. тиков.

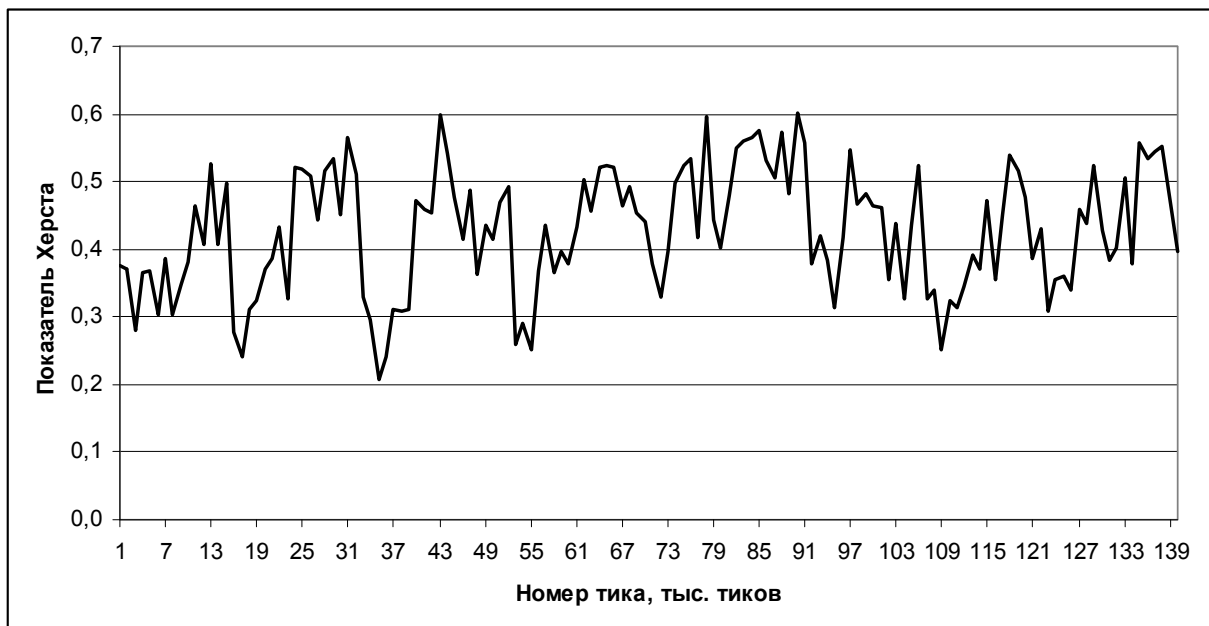


Рис. 4. Динамика показателя Херста $H_N(n)$ для $N = 20$ тыс. тиков

Из рис. 4 видно, что в разные моменты времени, трактуемые как номера событий, показатели Херста существенно отличаются. В общем-то, для нестационарного, как мы показали выше, тикового ряда в этом нет ничего удивительного. Интересно, однако, то, что временной ряд показателей Херста оказывается стационарным в узком смысле, начиная с выборок длины 300 точек, т.е. две встык-выборки показателей Херста длиной 300 отличаются одна от другой в смысле формулы (30) менее, чем на удвоенную точность оценки эмпирических частот соответствующего распределения. С увеличением окна N для вычисления $H_N(n)$ этот ряд становится все более стационарным. Более того, если рассмотреть ряд показателей Херста длиной более 1 тыс. точек, то с высокой точностью (на уровне достоверности более 0,99) его распределение оказывается нормальным. Центр этого распределения приблизительно линейно смещается вправо с увеличением окна N от 0,41 для $N = 10$ тыс. тиков до 0,53 для $N = 100$ тыс. тиков и далее не меняет своего значения. Стандартное отклонение практически не зависит от N и составляет 0,09. Таким образом, это распределение имеет вид

$$f_N(H) = \frac{1}{\sqrt{2\pi}\sigma_H} \exp\left(-\frac{(H - \bar{H}(N))^2}{2\sigma_H^2}\right), \quad \sigma_H = 0,09. \quad (32)$$

Распределение показателей Херста для окна $N = 100$ тыс. тиков (в этом случае и сам тиковый ряд стационарный) по выборке 20 тыс. значений приведено на рис. 5.

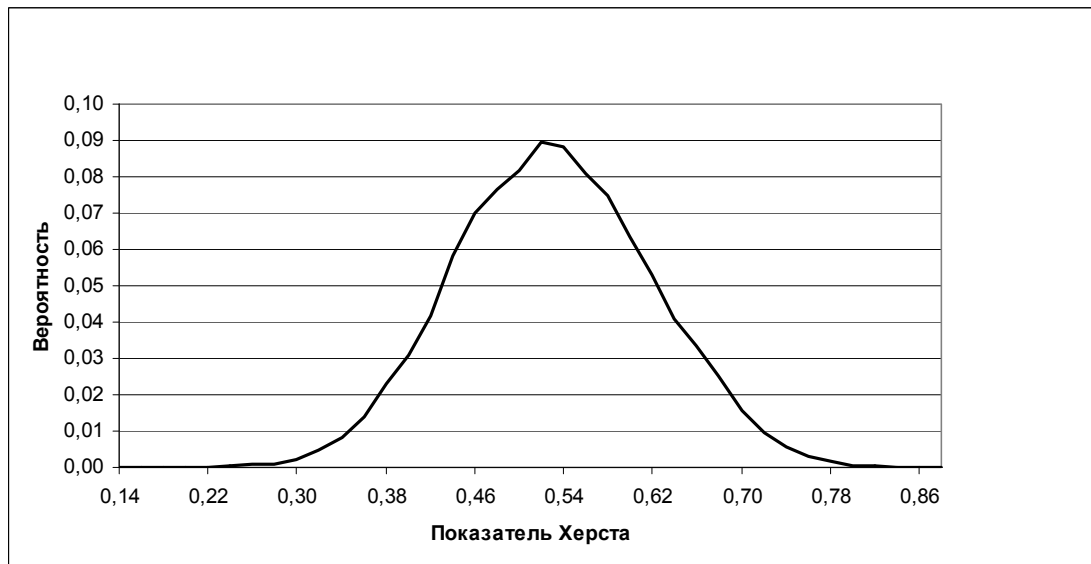


Рис. 5. Распределение показателей Херста $H_N(n)$ для $N = 100$ тыс. тиков

Следовательно, обнаруженное свойство тикового ряда можно интерпретировать как стандартное броуновское блуждание самого показателя Херста, тогда как цена испытывает случайное блуждание в виде процесса с толстым хвостом (типа Леви), что видно из распределений на рис. 1.

Что касается детерминации показателя Херста как коэффициента в регрессионном соотношении (7), то в целом она оказывается не очень высокой для окна малой длины N и возрастает с его увеличением. Для окна в 10 тыс. тиков детерминация в среднем составляет 0,35, но для окна в 100 тыс. тиков она становится равной 0,85. Распределение значений детерминации унимодально, но не нормально, и имеет вид, более похожий на гамма-распределение.

В некоторых работах (напр., в [8]) приближение значения показателя Херста к 0,5 с увеличением длины выборки трактуется как свидетельство потери памяти временным рядом, а соответствующее место излома углового коэффициент регрессии (7), найденное при фиксированном моменте времени, как критическую длину, характерную для данного временного ряда. Эта трактовка справедлива для одного конкретного фрагмента траектории. Но, как следует из проведенного анализа, в другие моменты времени место излома будет другим. Более корректной представляется та точка зрения, что исходный ряд характеризуется не показателем Херста, а распределением этого показателя, если только он не самоподобный. Напомним, что случайный процесс $X(t)$ называется статистически самоподобным с показателем самоподобия H , если для любого вещественного $\alpha > 0$ распределения процессов $X(t)$ и $\alpha^{-H} X(\alpha t)$ совпадают. На практике самоподобные процессы в чистом виде не реализуются, поэтому имеет смысл рассматривать не показатель Херста как таковой, а его распределение.

На уровне распределений обобщение показателя Херста на маркированный процесс очевидно: оно дается сверткой распределений (32) и (15, 22) с параметром потока, даваемым эмпирическим распределением, показанным на рис. 3. Оно, естественно, уже не является стационарным. Обозначая это распределение $F_T(H, t)$, где T есть длина временного окна в текущий момент времени t , имеем

$$F_T(H, t) = \frac{1}{\sum_{n=N_{\min}(T)}^{N_{\max}(T)} p_n(t-T, t)} \cdot \sum_{N=N_{\min}(T)}^{N_{\max}(T)} f_N(H) p_N(t-T, t), \quad (33)$$

$$p_N(t-T, t) = \frac{(\Lambda(t, T))^N}{N!} \exp(-\Lambda(t, T)).$$

Примеры распределений $F_T(H, t)$ для $T = 300$ мин. и разных моментов времени ($t = 300, 540$ и 900 мин. в обозначениях рис. 3) приведены на рис. 6. Эти распределения отличны от нормальных. Вероятности в них определены с точностью 0,015 в терминах формулы (31), а расстояния между распределениями в разные моменты времени равны 0,04, 0,35 и 0,38 соответственно для циклически переставляемых пар. Поскольку неравенство (31) в этих примерах не выполняется, распределения следует признать разными, т.е. процесс нестационарный.

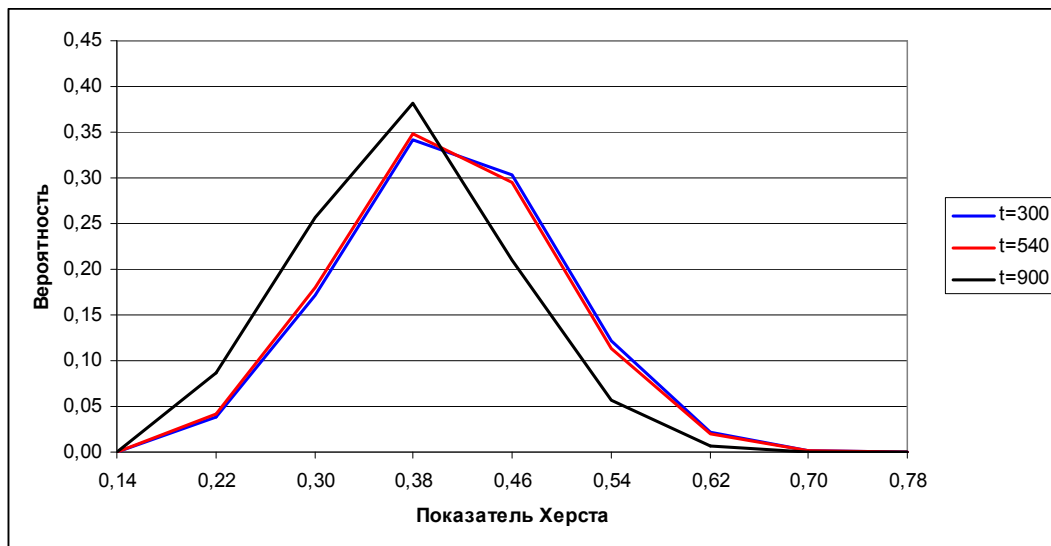


Рис. 6. Распределения показателей Херста для маркированного временного ряда в окне $T = 300$ мин. в разные моменты времени

Заметим, что наиболее вероятное значение нового показателя Херста сдвинулось влево по сравнению с распределением на рис. 5 и стало равным 0,38. Следовательно, как и должно быть при нестационарных немонокотонных потоках событий, тиковый ряд в терминах времени более изменчивый, чем тот же ряд в терминах номера события. Таким образом, анализ именно распределений, а не динамики значений тех или иных индикаторов временных рядов дает более объективное представление о свойствах процесса.

Литература

1. Pracasa Rao B.L.S. Statistical Inference for Fractional Diffusion Processes. – Wiley, NY, 2010.
2. Bassler K.E., Gunaratne G.H., McCauley J. L. Hurst Exponents, Markov Processes, and Nonlinear Diffusion Equations. 2006.
3. Rachev S.T., Weron A., Weron R. CED Models for Asset Returns and Fractal Market Hypothesis // Mathematical and Computer Modeling, 29, 1992.
4. Peters E. Fractal Market Analysis. Applying Chaos Theory to Investment and Economics. – Wiley, NY, 1994.
5. Hurst H., Black R., Simaika Y. Long-term Storage. An Experimental Study. – London: Constable, 1965. – 184 p.
6. Хинчин А.Я. Работы по математической теории массового обслуживания. – М.: Физматлит, 1963.
7. Босов А.Д., Орлов Ю.Н. Эмпирическое уравнение Фоккера-Планка для моделирования нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша РАН. 2013. №3. 30 с.
URL: <http://library.keldysh.ru/preprint.asp?id=2013-3>
8. Бутаков В., Граковский А. Оценка уровня стохастичности временных рядов произвольного происхождения при помощи показателя Херста. // Computer modeling and new technologies, 2005, Vol. 9, № 2, p. 27-32.