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

Препринты ИПМ • Препринт № 17 за 2013 г.

Лебедев А.В., Орлов Ю.Н., Шагов Д.О.

Индикатор оптимальной длины сканирования нестационарного временного ряда

Рекомендуемая форма библиографической ссылки: Лебедев А.В., Орлов Ю.Н., Шагов Д.О. Индикатор оптимальной длины сканирования нестационарного временного ряда // Препринты ИПМ им. М.В.Келдыша. 2013. № 17. 16 с. URL: http://library.keldysh.ru/preprint.asp?id=2013-17

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.В. Лебедев, Ю.Н. Орлов, Д.О. Шагов

Индикатор оптимальной длины сканирования нестационарного временного ряда

Москва — 2013

Лебедев А.В., Орлов Ю.Н., Шагов Д.О.

Индикатор оптимальной длины сканирования нестационарного временного ряда

Исследованы свойства распределения расстояний между двумя выборочными плотностями распределений по выборкам, взятым с непустым пересечением элементов временного ряда. Выборочные плотности оцениваются по гистограмме с оптимальным равномерным разбиением, согласованным с точностью оценки эмпирических вероятностей. Проведено обобщение индикатора оптимальной длины выборки, по которой следует вычислять скользящие статистики нестационарного временного ряда.

Ключевые слова: индикатор статистической добротности, оптимальная длина выборки, нестационарные распределения, временные ряды

Lebedev A.V., OrlovYu.N., Shagov D.O.

Optimal set length indicator for non-stationary time-series

The properties of distributions of the distances between two empirical distribution function densities for non-stationary time-series are investigated. The optimal choice of histogram interval is suggested on the basis of self-agreement accuracy of empirical probability. The generalization of optimal set length indicator with the use of statistical good quality factor is given.

Key words: statistical good quality indicator, optimal histogram class interval, time series, non-stationary distributions

Работа выполнена при поддержке гранта РФФИ, проект № 11-01-00444

Содержание

1. Введение

В настоящей работе изучается двумерный индикатор нестационарности временных рядов, основанный на распределении расстояний между плотностями выборочных распределений, построенных по выборкам длины N данных, сдвинутых одна относительно другой на определенное число $\tau \leq N$ шагов. Данная работа продолжает исследование [1], где был предложен индикатор статистической добротности нестационарного временного ряда для решения задачи о разладке, т.е. о наиболее точном определении момента, когда следует считать, что распределение значимо изменилось. В работе [1] был проанализирован частный вид этого индикатора, когда сдвиг между выборками был равен длине выборки. Именно, определялась оптимальная длина окна, когда из сравнения текущего распределения, построенного в этом окне, с соседним распределением, т.е. из сравнения распределений двух «встык-выборок», делался вывод о состоявшейся разладке с минимальной ошибкой.

Исследование частично перекрывающихся выборок представляет интерес в связи с тем, что введение второго параметра (величины сдвига между выборками) может привести к обнаружению новых оптимальных длин выборок, на которых следует определять разладку. Параметр сдвига является условием, при котором ищется условная вероятность того, что превышение определенного индикативного расстояния между выборочными распределениями свидетельствует о рассогласовании между ними. Для конкретного ряда может оказаться так, что ошибка определения разладки минимальна на определенном окне сдвига. Поскольку желательно промежуток запаздывания между состоявшейся разладкой и моментом ее обнаружения с помощью того или иного индикатора сделать как можно меньше, то анализ расстояний между распределениями «внахлест-выборок» вместо «встык-выборок» является естественным обобщающим шагом индикатора статистической добротности.

Практическое применение индикатора предполагает сравнение текущего расстояния между двумя надлежащим образом выбранными распределениями с некоторым эталонным расстоянием, характерным для такого рода пар выборок. Если наблюдаемое расстояние больше эталонного, то перед нами разладка, а если меньше, то статистическое состояние изучаемой системы не изменилось. Вопрос состоит в том, какое значение признать эталонным и каким способом выбрать для создания этой эталонной величины пару распределений.

Следует иметь в виду, что статистический индикатор будет иногда (или часто) срабатывать неверно. Во-первых, он может не заметить разладку, а также ложно сработать при отсутствии разладки. Суммарная ошибка может быть весьма значительной, если величина стационарного шума, на фоне которого ищется разладка, достаточно велика. Таким образом, следует идентифицировать сигнал о разладке на фоне шума.

Идея такой идентификации для нестационарного ряда состоит в следующем. Если генеральное распределение стационарно, то выборочное

распределение будет отличаться от него на величину, оцениваемую на принятом уровне значимости по дисперсии эмпирических частот в соответствии с классическими критериями. Уровень значимости корректно выбрать таким, чтобы он совпадал с долей ошибочно отвергнутых выборок как не принадлежащих данной стационарной генеральной совокупности, которая (доля) практически наблюдалась бы в серии равнозначных с точки зрения экспериментатора вычислительных процедур по сравнению распределений между собой. Такой уровень будем в дальнейшем называть согласованным. Необходимость использовать согласованный уровень значимости вызвана тем, что бессмысленно требовать уровня значимости, который меньше, чем точность, с которой определяются эмпирические вероятности по выборке конечной длины, а последняя точность в отсутствие теоретического распределения оценивается как раз практическими наблюдениями за флуктуациями эмпирических частот. Согласованный уровень значимости определяет стационарный уровень шума как соответствующий квантиль стационарного распределения.

Аналогично можно построить и согласованный уровень стационарности выборочных распределений по набору текущих расстояний между выборками определенных длин. Этот уровень является характерным расстоянием между выборками, вероятность превышения которого равна самому этому расстоянию. Длина выборки и величина нахлеста выбираются из решения оптимизационной задачи: отношение сигнала (согласованного уровня стационарности) и шума (согласованного уровня значимости) должно быть максимально. В результате оптимизации находится сдвиг и отвечающая ему длина выборки, при которых ошибка идентификации разладки минимальна. Этот подход и реализован в данной работе.

2. Оптимальное разбиение гистограммы

Прежде чем сравнивать между собой уровни сигнала и шума, надо определиться с кластеризацией событий по классовым интервалам, что, как и задача о разладке, представляет собой определенную задачу фильтрации. Требуется выбрать число классовых интервалов таким, чтобы частоты попадания в них были оценены с достаточной точностью, а также и форма плотности распределения была бы приближена аккуратно.

Рассмотрим последовательность из N значений случайной величины ξ, которые попали в определенные классовые интервалы, число которых n. Элементы этой последовательности обозначим x_j, $j = 1, 2, ..., N$. Пусть из них значение x_i встретилось k_i раз. Тогда выборочной плотностью функции распределения (далее ВПФР), оцениваемой по гистограмме для заданной выборки длины N, называется совокупность $f_N(i,t)$ величин

$$f_N(i,t)=\frac{k_i}{N},\quad i=1,2,...,n. \tag{1}$$

Время t в аргументе $f_N(i,t)$ указывает на текущий момент времени, от которого назад отсчитывается выборка длины N.

Очевидно, выводы о стационарности или нестационарности плотности сильно зависят от того, как построены классовые интервалы принадлежности случайной величины. Мы будем рассматривать равномерное разбиение на число интервалов, согласованных с точностью оценки эмпирических вероятностей. Суть этого подхода, предложенного в [2], состоит в том, чтобы шаг гистограммы был равен точности оценки вероятностей, которая определяется по классическому критерию. Точность $\varepsilon(N,n)$ оценки ВПФР по выборке длины N по гистограмме, разбитой на n классовых интервалов, дается формулой [1, 2]

$$\frac{t_{1-\varepsilon/2}}{\varepsilon}=\frac{\sqrt{N}}{\Sigma_N(n)}, \tag{2}$$

где t_γ есть γ-квантиль распределения Стьюдента для бесконечного числа степеней свободы (считаем, что $N\ge 100$), а сумма, определяющая влияние мелкости разбиения гистограммы на точность оценки эмпирических вероятностей, равна

$$\Sigma_N(n)=\sum_{i=1}^{n}s_N(j)=\sum_{i=1}^{n}\sqrt{f_N(i)\bigl(1-f_N(i)\bigr)}. \tag{3}$$

Функция $t_{1-\varepsilon}/\varepsilon$ монотонно убывает с ростом ε, поэтому к ней существует обратная, значение которой и дает верхнюю оценку точности определения эмпирических вероятностей по заданному разбиению гистограммы. Обозначим для краткости

$$\varphi(\varepsilon)=\frac{t_{1-\varepsilon}}{\varepsilon},\quad \psi=\varphi^{-1},\quad z\equiv z\bigl(N,n\bigr)=\frac{\sqrt{N}}{\Sigma_N(n)}. \tag{4}$$

Тогда точность оценки ВПФР определяется формулой

$$\varepsilon=2\psi(2z). \tag{5}$$

Сумма (3) выражает качество приближения плотности гистограммой, поскольку, чем меньше сумма, тем выше точность оценки ВПФР, т.е. тем меньше число ε. С увеличением числа интервалов сумма (3) возрастает безотносительно вида распределения, а аргумент z функции ψ убывает.

Заметим теперь, что если вероятности оценены с точностью ε, то с той же точностью оценивается и среднее значение случайной величины. Поэтому знание фактических величин с большей точностью будет излишним (не вообще, а в смысле превышения точности при данном разбиении). С другой стороны, если выбрано разбиение на n классовых интервалов, то среднее значение по выборке отличается от среднего значения по гистограмме на величину $1/n$.

Исходя из этих качественных соображений, можно сделать вывод, что на практике не имеет смысла выбирать мелкость разбиения гистограммы, превышающую точность оценки вероятностей (по этой самой гистограмме). Тогда оптимальным равномерным разбиением считается ближайшее натуральное число n (классовых интервалов) к решению уравнения, согласующего уровень значимости (5) и мелкость разбиения:

$$\frac{1}{n} = 2\psi\left(\frac{2\sqrt{N}}{\Sigma_N(n)}\right). \qquad (6)$$

Уравнение (6) относится к одной конкретной выборке длины N, оканчивающейся в текущий момент времени t. Следовательно, чтобы не перестраивать разбиение гистограммы на каждом шаге по времени, надо определить характерную для данного ряда точность оценки эмпирических вероятностей и промежуток времени, на котором следует ее пересчитывать. Обе эти задачи решаются методом анализа соответствующих плотностей распределений.

Пусть в результате произвольно выбранного разбиения гистограммы на n классовых интервалов получен временной ряд точностей оценки ВПФР в соответствии с (5). Строится плотность $\nu_M(\varepsilon, n)$ распределения этих точностей по некоторому массиву $M > N$ данных и определяется согласованный уровень точности $\varepsilon_M^*(n)$, определяемый как квантиль распределения $\nu_M(\varepsilon, n)$, равный собственному уровню значимости [2], т.е. (индекс *M* для краткости опущен)

$$\int_0^{\varepsilon^*} \nu(\varepsilon, n)\, d\varepsilon = 1 - \varepsilon^*. \qquad (7)$$

Затем определяется отвечающее этому согласованному уровню точности число промежутков разбиения $n^* = [1/\varepsilon^*(n)]$. Если оказалось, что $n^* > n$, число интервалов увеличивается и процедура повторяется, строится новая плотность $\nu(\varepsilon, n)$ и т.д. Если же $n^* < n$, число интервалов аналогично уменьшается. Таким образом, процесс нахождения усредненно-оптимального разбиения будет иметь итерационный характер. Наилучшее целочисленное приближение к решению уравнения (6) и даст оптимальную мелкость равномерного разбиения гистограммы. Если количество данных столь велико, что точность (5) совпала с точностью измерения значений случайной величины, то достигнута естественная мелкость разбиения и дальнейшего увеличения числа интервалов не проводится.

Время, по истечении которого следует пересчитать распределение $\nu_M(\varepsilon, n)$, определяется из анализа уровня нестационарности этого распределения. Это делается по той же методике, что и для самого ряда (п.3).

Приведем результат применения методики оптимального разбиения по формуле (6) в нестационарном случае к некоторым биржевым рядам. Интерес

представляют тиковые ряды, когда можно анализировать выборки больших длин – порядка миллиона данных. В работе рассматриваются тиковые данные фьючерсов GC на золото и CL на нефть за период 2008-2010 гг. Этот период взят исключительно для иллюстрации методики.

Число интервалов разбиения гистограммы в зависимости от длины выборки приведено на рис. 1. Как видно, разбиения довольно похожи. Отметим, что с достоверностью 0,97 число интервалов в нашем примере растет как $N^{0,34}$ (см. [2]), что близко к результатам Смирнова [3], показавшего, что уклонение гистограммы от графика непрерывной плотности функции распределения обратно пропорционально корню третьей степени из длины выборки. Здесь мы получили качественно тот же результат, но методом, который не требует наличия генеральной совокупности, что позволяет его применить и к нестационарным временным рядам.

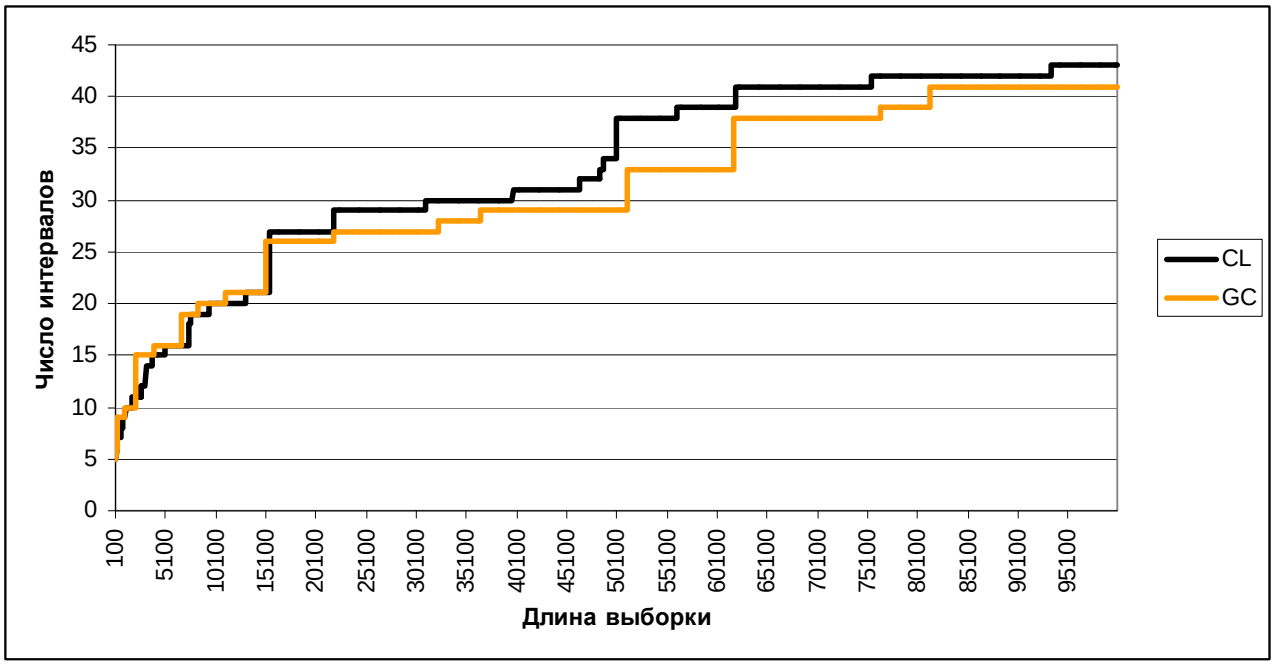

Рис. 1. Оптимальное число интервалов в зависимости от длины выборки для тиковых рядов

При длине выборки свыше 100 тыс. число интервалов становится сравнимо с естественным разбиением ряда приростов цен в соответствии с точностью их измерения (в данном случае цены измеряются в долларах США с точностью до центов). Поэтому задача кластеризации данных актуальна при длинах выборки, меньших 100 тыс., а при превосходящих длинах мы имеем дело с явно дискретным распределением вероятностей. Характерно, что длина порядка 100 тыс. дистинктивных тиков отвечает промежутку времени 1-2 суток, так что более длинные выборки, по нашему мнению, теряют актуальность, поскольку они учитывают данные из слишком удаленного прошлого, не влияющего на текущую ситуацию длительностью обычно в десятки минут.

3. Распределение расстояний между нестационарными ВПФР

Определив оптимальное разбиение гистограммы, представляющей ВПФР в текущий момент времени, введем расстояние в пространстве суммируемых функций и определим расстояние между двумя ВПФР, построенными по выборкам длины N, смещенным одна относительно другой на τ шагов:

$$\rho(N,\tau;t) = \left\| f_N(x,t) - f_N(x,t+\tau) \right\| = \sum_{i=1}^{n} \left| f_N(i,t) - f_N(i,t+\tau) \right|. \qquad (8)$$

Величину смещения τ будем варьировать от 1 до N. Возможные значения расстояний между выборками образуют дискретный набор $\tau+1$ чисел: $\rho \in \{0; 2/N; 4/N; ...; 2\tau/N\}$. Обозначим через $g_{N,\tau}(k)$ вероятность того, что расстояние ρ между выборками при сдвиге на τ равно $2k/N$. Величина

$$G_{N,\tau}(\rho) = \sum_{l=0}^{k(\rho)} g_{N,\tau}(l), \quad k(\rho) = [N\rho/2] \qquad (9)$$

представляет эмпирическую вероятность того, что расстояние между распределениями не больше ρ. Наибольшее возможное расстояние между выборками равно двум, когда мера пересечения носителей распределений равна нулю. Определим согласованный уровень стационарности (далее СУС) ρ^* так, что вероятность его превышения равна уровню значимости критерия, т.е.

$$G_{N,\tau}\left(\rho^*\right) = 1 - \rho^* \frac{N}{2\tau}. \qquad (10)$$

Уравнение (10) определяет функцию $\rho^*(N,\tau)$, которая обладает тем свойством, что при проведении достаточно большого числа экспериментов по вычислению расстояний между двумя выборочными распределениями длины N, сдвинутыми на окно τ, в доле ρ^* случаев будет наблюдаться превышение расстояния, равного ρ^*. Тем самым ρ^* можно трактовать как характерное расстояние между распределениями на уровне значимости, не превосходящем этого расстояния. Если оказалось, что $\rho^* \le 2\varepsilon^*$, то на достижимом уровне значимости ряд следует считать стационарным. Если же $\rho^* > 2\varepsilon^*$, то ряд нестационарный, и уровень ρ^* есть характерное расстояние между выборками, превышение которого следует считать разладкой.

Для многих нестационарных рядов характерно наличие локальных минимумов согласованного уровня стационарности в зависимости от длины «встык-выборок», тогда как для стационарного ряда наблюдается монотонно убывающая зависимость. Аргументы этих локальных минимумов можно трактовать как типовые для данного ряда промежутки времени, на которых происходит смена режима работы наблюдаемой системы. Однако важно, чтобы эти минимумы были статистически значимыми, т.е. после их прохождения СУС

становился бы больше статистической ошибки вычисления расстояния между ВПФР. Примеры зависимости СУС $\rho^*(N) \equiv \rho^*(N,N)$ от длины выборки для тиковых рядов CL и GC приведены на рис. 2 в сравнении с точностью $2\varepsilon^*(N)$.

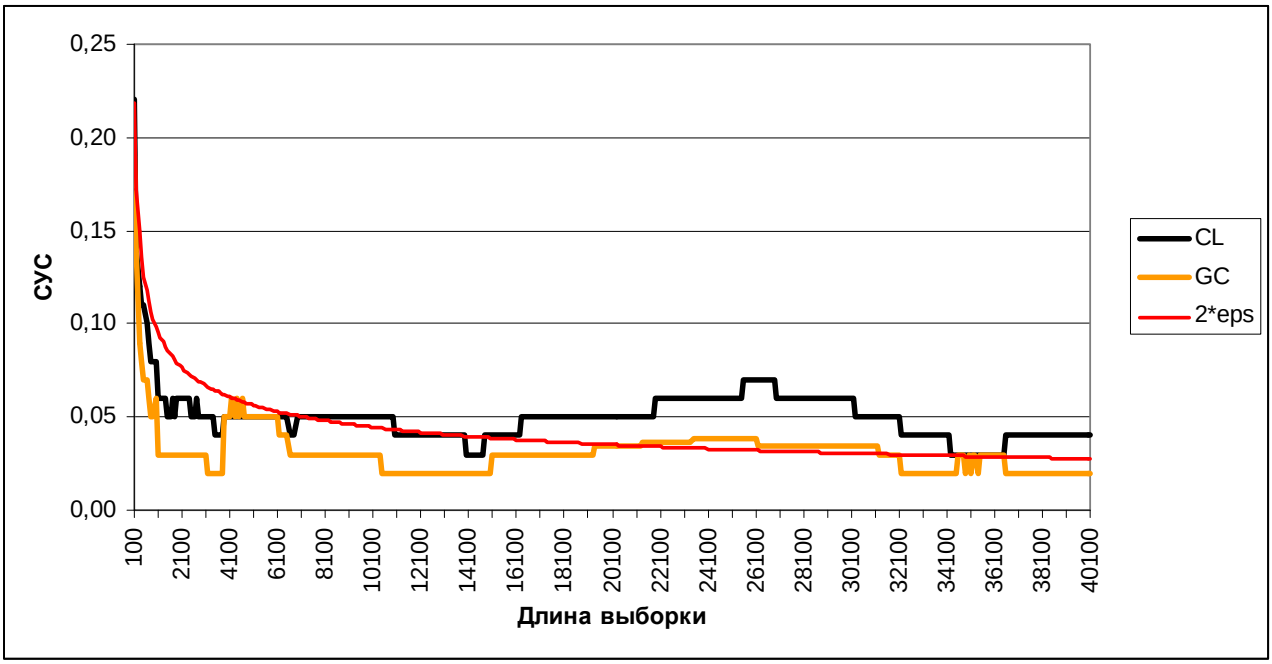

Рис. 2. Зависимость СУС встык-выборок от длины выборки

Из рис. 2 видно, что на малых длинах выборки минимумы СУС не являются значимыми как для CL, так и для GC, поскольку соответствующие графики проходят ниже кривой стационарной точности. Иными словами, минимум в области 4 тыс. тиков для СУС ряда GC интересен, но последующий рост расстояний между ВПФР не может, строго говоря, считаться значимым. Лишь на длинах, превышающих 15 тыс. тиков для ряда CL и 20 тыс. тиков для ряда GC имеет смысл искать доказательств разладки, причем ряд GC сам по себе весьма близок к стационарному. На длинах свыше 50 тыс. тиков СУС ряда CL стабилизируется на относительно низком уровне, слабо превышающим уровень стационарной точности.

Таким образом, как малые, так и большие длины выборок не дают возможности распознать значимые изменения в распределении вероятностей нестационарного ряда. В первом случае недостаточно точности в оценке ВПФР, а во втором случае ряды имеют квазистационарное поведение без резких изменений СУС. Интерес представляют выборки средних длин, каковыми в нашем случае следует считать длины порядка 15-25 тыс. данных.

Картина для «внахлест-выборок» качественно такая же, как и для «встык-выборок». На рис. 3 показано отношение $\rho^*(N,\tau)/\left(2\varepsilon^*(N,\tau)\right)$ для ряда CL с шагом 100 по обоим параметрам N и τ. Видно, что увеличение сдвига приводит к более явному выделению нестационарности на фоне шума.

Превышение СУС хорошо заметно на выборках длин порядка 5-7 тыс. и 20-25 тыс. На выборках длин 14-15 тыс. наблюдается локальный минимум отношений этих двух показателей, так что соответствующие ВПФР можно с высокой точностью считать стационарными.

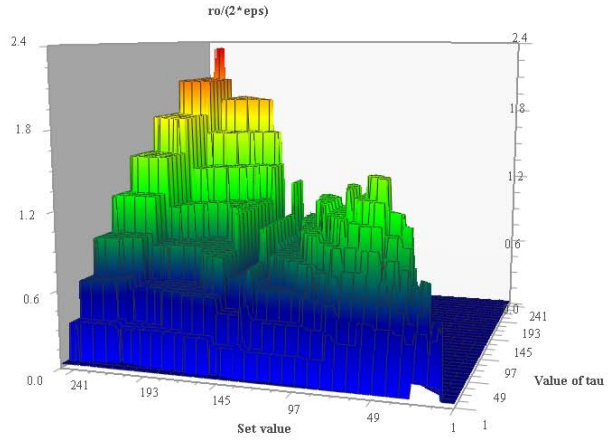

Рис. 3. Отношение СУС и точности оценки ВПФР
в зависимости от длин выборки и сдвига, ряд CL

Когда СУС временного ряда превышает уровень стационарного шума, следует разобраться, сколь значимо это превышение. Например, визуально по рис. 2 кажется, что ряд GC стационарный, поскольку его СУС почти целиком проходит ниже теоретической кривой стационарной точности оценки ВПФР, а превышение уровня $2\varepsilon^*(N)$ вроде бы незначительное. Но реальной величиной, характеризующей это превышение, является доля расстояний, попадающая в промежуток между СУС $\rho^*(N)$ и уровнем $2\varepsilon^*(N)$.

Введем тогда индекс нестационарности $J(N,\tau)$ ряда, положив его равным отношению доли расстояний, не превосходящих фактический СУС, построенный по имеющимся эмпирическим данным, к уровню шума, за который естественно принять долю расстояний между ВПФР, не превосходящих $2\varepsilon^*\tau/N$ при заданной длине выборки N и сдвиге τ:

$$J(N,\tau)=\frac{G_{N,\tau}\left(\rho^*\right)}{G_{N,\tau}\left(2\varepsilon^*\tau/N\right)}. \qquad (11)$$

Если $J(N,\tau)\le 1$, ряд считается стационарным, а если $J(N,\tau)>1$, то ряд нестационарный. Напомним, что с увеличением длины выборки растет и число промежутков разбиения, так что даже незначительное превышения уровня $2\varepsilon^*\tau/N$ для СУС «внахлест-выборок» может привести к заметному отличию

индекса (11) от единицы. Пример индекса нестационарности для «встык-выборок» приведен на рис. 4.

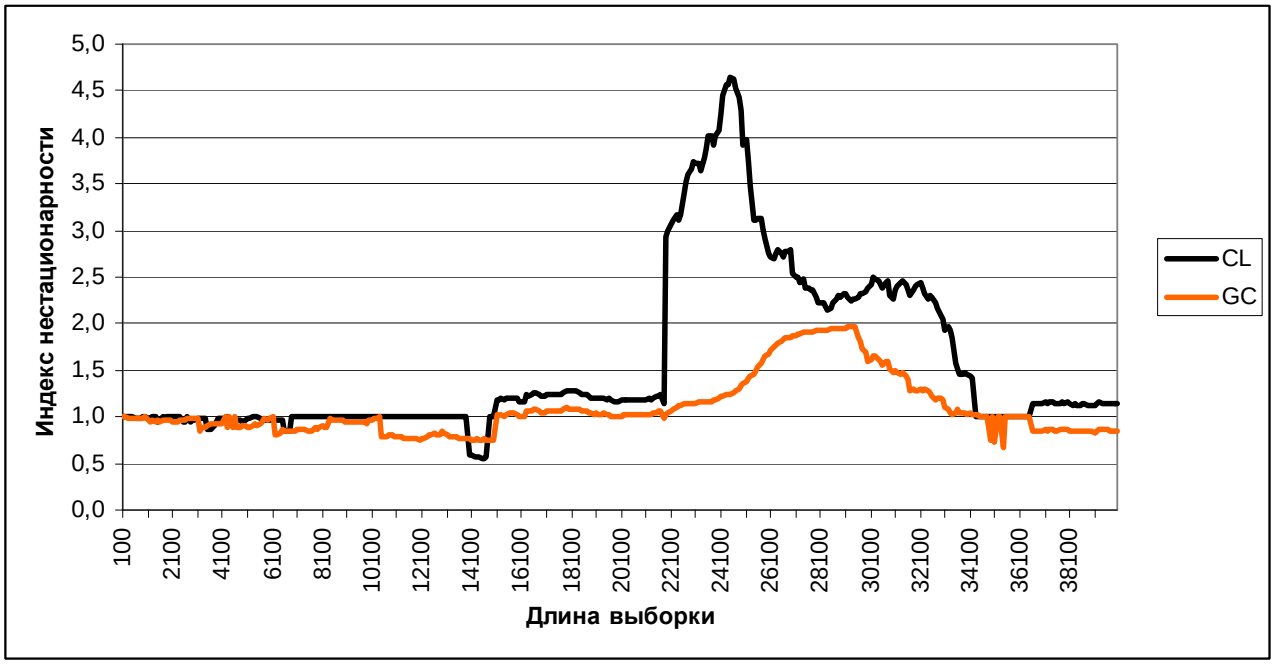

Рис. 4. Зависимость индекса нестационарности от длины выборки

Видно, что на определенных длинах оба ряда являются нестационарными, хотя ряд CL, конечно, гораздо более нестационарен, чем ряд GC. Это, однако, еще не означает, что анализ разладки лучше всего проводить на тех длинах, где индекс нестационарности максимален. Конечно, такой анализ имеет смысл проводить только там, где распределения считаются разнородными на принятом уровне значимости $\rho^*(N,\tau)$ при условии $\rho^* > 2\varepsilon^*$, т.е. там, где индекс нестационарности больше единицы. Но число событий, при которых происходит превышение уровня СУС текущей парой выборок, требует еще некоторой оптимизации. Индикатор для такой оптимизации строится ниже.

4. Индикатор статистической добротности

Важную информацию об анализируемой системе можно извлечь из анализа индикатора статистической добротности $q(N,\tau)$ как функции длины выборки и величины сдвига. Этот индикатор для «встык-выборок» был введен в [1] и представляет собой аналог отношения «сигнал-шум», т.е. отношение доли выборок, расстояние между которыми превосходит СУС, к доле выборок, расстояние между которыми не превосходит стационарного шума. Для «внахлест-выборок» этот индикатор определяется формулой

$$q(N,\tau) = \frac{\rho^* N/(2\tau)}{G_{N,\tau}\left(2\varepsilon^*\tau/N\right)} = \frac{1-G_{N,\tau}\left(\rho^*\right)}{G_{N,\tau}\left(2\varepsilon^*\tau/N\right)}. \qquad (12)$$

Максимум этой функции по N при фиксированном параметре сдвига τ мог бы дать оптимальную длину выборки для прогнозирования свойств системы на горизонте τ, если бы не одно «но»: если ряд сильно нестационарный, то его СУС очень большой, находящийся вблизи максимальных значений расстояний между выборками, а тогда доля событий, когда $\rho \ge \rho^*$, окажется весьма малой, и индикатор будет давать большую ошибку второго рода, т.е. редко срабатывать. Чтобы уменьшить долю несрабатываний, в [1] был введен корректирующий множитель, равный доле меры носителя распределения *G*, для которой расстояния превосходят СУС, так что мы будем далее статистической добротностью называть величину

$$Q(N,\tau) = q(N,\tau)\left(1 - \rho^* N/(2\tau)\right) = \frac{\left(1 - G_{N,\tau}\left(\rho^*\right)\right)\cdot G_{N,\tau}\left(\rho^*\right)}{G_{N,\tau}\left(2\varepsilon^*\tau/N\right)}. \qquad (13)$$

Смысл индикатора (13) в том, что на промежутке $\rho \ge \rho^*$ должно быть достаточно много событий. Если индекс нестационарности (11) большой, но доля идентифицируемых событий мала, то индикатор (13) не обязательно будет иметь максимум на соответствующих значениях N и τ. Поэтому может оказаться, что оптимальным вариантом будет не тот, которому отвечает наибольший индекс нестационарности (11) или максимальное значение $q(N,\tau)$, а тот, где и ряд достаточно нестационарен, и доля превышения уровня нестационарности заметно отлична от нуля. Оптимальной длиной сканирования ряда и длиной сдвига являются такие пары N,τ, при которых $Q \to \max$. Само же по себе численное значение величины Q не играет роли. Оно может быть как больше, так и меньше единицы.

На рис. 5 показана форма поверхности функционала расстояния между двумя «внахлест-выборками» в некоторый момент времени в зависимости от длины выборки и величины сдвига. В целом расстояние между двумя ВПФР с увеличением длины выборки при фиксированном окне сдвига уменьшается, а с увеличением окна сдвига при фиксированной длине выборки растет.

Исходный временной ряд может обладать квазипериодической разладкой, проявляющейся не только локально по времени, но и в среднем на уровне СУС в немонотонном его поведении. Так, на рис. 5 отчетливо виден минимум расстояния на длине выборки около 20 тыс. (200 точек с шагом 100 на оси «Set value») при любом сдвиге, а также локальный максимум на выборках, больших 20 тыс., при сдвиге порядка 5000 (50 точек с шагом 100 на оси «Value of tau»). Последнее свидетельствует о возможности идентифицировать разладку на меньших, чем «встык-выборки», длинах. Главный максимум расстояний приходится на относительно небольшие длины выборки порядка 5 тыс. и величины сдвига на 10 тыс., но этот максимум оказывается статистически незначимым, поскольку находится ниже кривой уровня стационарного шума.

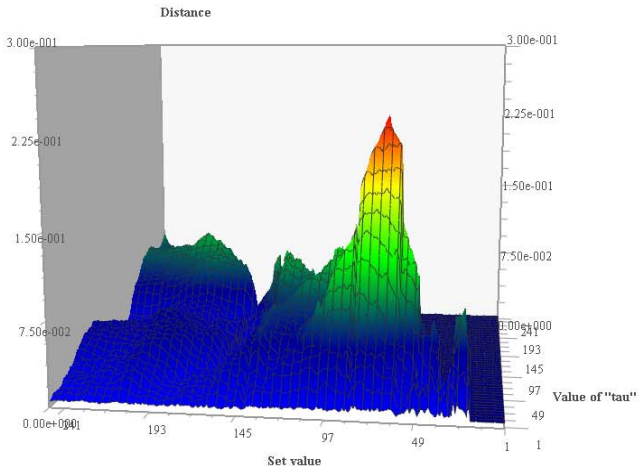

Рис. 5. Функционал расстояния между «внахлест-выборками» в некоторый момент времени в зависимости от длин выборки и сдвига, ряд CL

На рис. 6 показана зависимость СУС (10) от длин выборки и сдвига. Будучи функционалом, полученным в результате усреднения многочисленных расстояний между парами выборок, зависимость СУС от своих аргументов гораздо более монотонна, чем у отдельного расстояния в определенный момент времени (рис.5). Вследствие этого значимые нарушения монотонности представляют практический интерес. При этих аргументах временной ряд в среднем меняет свое поведение, так что разладку имеет смысл искать в таких областях немонотонности.

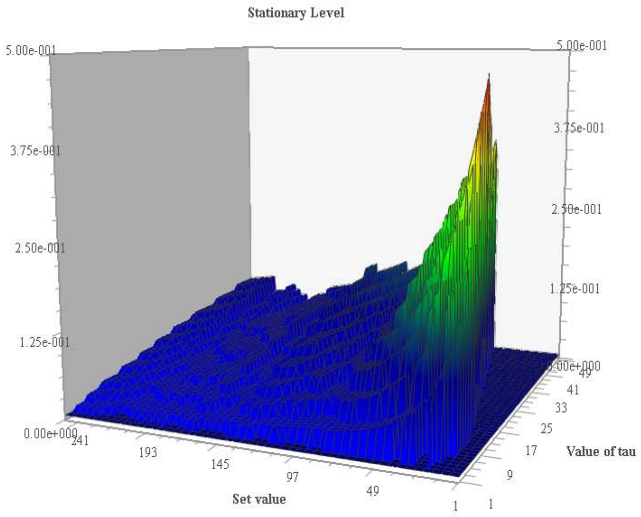

Рис. 6. СУС «внахлест-выборок» в зависимости от длин выборки и сдвига, ряд CL

На рис. 7-8 приведены поверхности индикатора добротности (13) в зависимости от длин выборки и сдвига. Величина сдвига в расчетах не превосходила длину выборки. Оказалось, что для обоих рядов существует область параметров «длина выборки – окно сдвига», где индикатор добротности имеет локальный максимум.

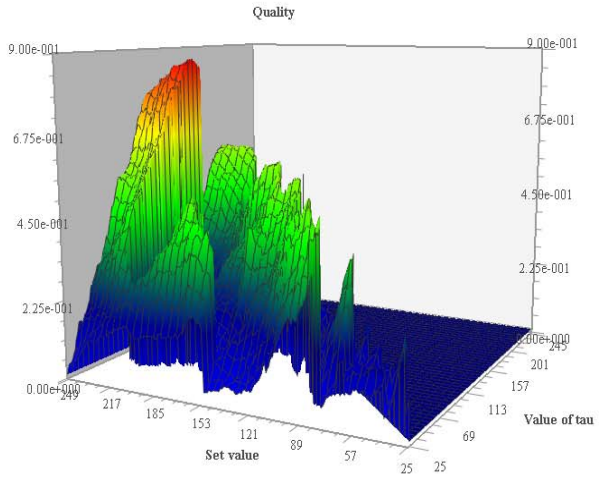

Рис. 7. Индикатор статистической добротности
в зависимости от длин выборки и сдвига, ряд GC

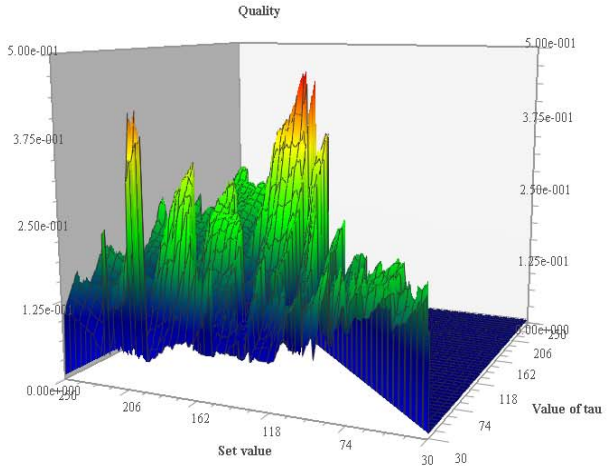

Рис. 8. Индикатор статистической добротности
в зависимости от длин выборки и сдвига, ряд CL

Более наглядно метод отбора оптимальных параметров демонстрируется на графиках рис. 9.

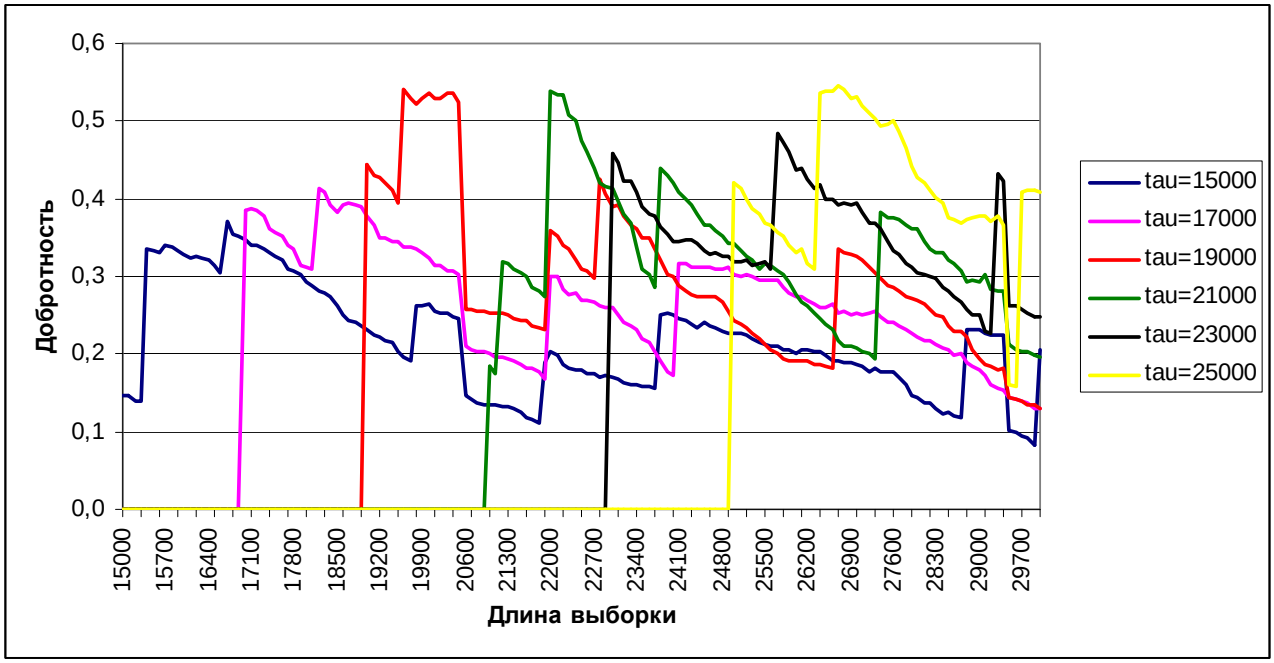

Рис. 9. К определению оптимальных длин сканирования, ряд CL

На графиках рис. 9 хорошо видна характерная особенность индикатора добротности (13): для ряда CL оказалось, что второй локальный максимум всегда выше первого. Это означает, что «встык-выборки» не являются наиболее оптимальным инструментом, а лучше всего разладку искать на выборках, имеющих общую часть длиной 1-2 тыс. тиков, что составляет 5-10 % от длины выборки. Наилучшим представляется вариант, когда выборки длин 19,5-20,5 тыс. сдвинуты на 19 тыс. тиков (красная линия). В этом случае локальный максимум наиболее широк, что свидетельствует о его устойчивости в пределах 1 тыс. тиков. Также представляют интерес выборки длин в 26-28 тыс. тиков, сдвинутые одна относительно другой на 25 тыс. тиков (желтая линия). Аналогично выглядят графики добротности и для ряда GC.

Скачкообразные изменения добротности связаны с изменением числа промежутков разбиения гистограммы для ВПФР с увеличением длины выборки.

Заметим, что индикатор добротности может и не привести к нахождению оптимальной длины сканирования выборки, если нестационарность ряда проявляется на слишком больших длинах, поскольку в таких случаях может не быть локального максимума добротности, а будет наблюдаться его монотонный рост с увеличением длины выборки и длины нахлеста. В таких случаях для выбора оптимального окна рекомендуется использовать локальные максимумы СУС «встык-выборок».

5. Заключение

В работе построены новые индикаторы, классифицирующие типы и текущее состояние нестационарных временных рядов через близость их выборочных плотностей функций распределения.

Подчеркнем, что эти индикаторы не универсальны в плане эффективности своего применения. В большинстве случаев временные ряды, встречающиеся на практике, имеют малые выборки, которые не позволяют построить распределение вероятностей с точностью, различающей нестационарное поведение на фоне шума. Малость выборки в нашем понимании – это длины менее 10 тыс. данных. Лишь тиковые биржевые ряды представляют интерес в плане анализа разладки в терминах ВПФР. Минутные отсчеты или, «хуже» того, часовые не особенно интересны, поскольку для построения выборки большого объема надо погрузиться в неактуальное прошлое. Для таких рядов более эффективен горизонтный анализ, развитый в [4].

Но даже для выборок надлежащих длин можно довольно часто обнаружить лишь тривиальный эффект: на малых длинах распределение неотличимо от стационарного в силу грубости оценки эмпирических вероятностей, а на больших длинах индекс нестационарности и статистическая добротность монотонно возрастают. Этот эффект связан с тем, что СУС многих нестационарных рядов перестает убывать после определенного объема данных, тогда как СУС стационарных рядов (т.е. согласованная значимость) монотонно убывает, поэтому отношение «сигнал-шум» в нестационарном случае будет неограниченно возрастать с увеличением выборки. Интерес представляет именно локальный максимум индикатора добротности, который к тому же имеет еще и устойчивое поведение, т.е. достаточно пологую форму.

Развитая методика позволяет более корректно подойти к проблеме идентификации разладки для нестационарных временных рядов.

Литература

1. Орлов Ю.Н., Шагов Д.О. Индикативные статистики для нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша РАН. 2011. № 53. 20 с. URL: http://library.keldysh.ru/preprint.asp?id=2011-53

2. Орлов Ю.Н. Оптимальное разбиение гистограммы для оценивания выборочной плотности функции распределения нестационарного временного ряда // Препринты ИПМ им. М.В. Келдыша РАН. 2013. № 14. 26 с. URL: http://library.keldysh.ru/preprint.asp?id=2013-14

3. Смирнов Н.В. О построении доверительной области для плотности распределения случайной величины // Доклады АН СССР. 1950. Т. 74. № 2. С. 189-192.

4. Орлов Ю.Н., Осминин К.П. Нестационарные временные ряды: методы прогнозирования с примерами анализа финансовых и сырьевых рынков. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2011. – 384 с.