М. С. Абрамович, М.Н. Мицкевич

0.1. ОБНАРУЖЕНИЕ СКАЧКООБРАЗНЫХ ИЗМЕНЕНИЙ СРЕДНЕГО С ИСПОЛЬЗОВАНИЕМ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ ХААРА

Для временных рядов, наблюдаемых с шумом, разрабатываются критерии обнаружения скачкообразных изменений среднего, основанные на вейвлет-преобразовании. Относительно шума предполагается, что он имеет более "тяжелые хвосты", чем у нормального распределения. Методом статистического моделирования исследуется эффективность критериев.

0.1.1. Введение

Обнаружение скачкообразных изменений среднего временных рядов, которые описывают процессы в экономике, технике и других приложениях, является актуальной прикладной задачей [1]. В последние годы для выявления локальных особенностей, в том числе и скачкообразных изменений среднего, широко применяется вейвлет-анализ [2,3]. Использование вейвлет-преобразования для обнаружения скачкообразных изменений основано на том, что в момент скачка абсолютные значения вейвлет-коэффициентов имеют максимальные значения. Альтернативный подход предполагает рассмотрение разностей соседних вейвлет-коэффициентов на уровнях разрешения и определение величины сдвига, для которого разность является максимальной. Эта величина сдвига и определяет момент скачка [4].

Критерии, основанные на использовании вейвлет-коэффициентов, их максимальных значений или статистик от них, позволяют обнаружить скачкообразные изменения даже при наличии шума. В параметрических критериях обнаружения разладок обычно предполагается нормальное распределение для шума [2]. Но на практике распределение шума часто отлично от нормального.

В настоящей работе относительно распределения шума предполагается, что он имеет более "тяжелые хвосты", чем у нормального распределения (например, как у распределения Стьюдента). Для этого случая в работе построены критерии обнаружения скачкообразных изменений временных рядов и методом статистического моделирования исследована их эффективность.

0.1.2. Математическая модель скачкообразных изменений

Пусть $T = 2^M$, $t = 0, \ldots, T-1$ и $x_t \in R$ – временной ряд, который описывается следующей моделью:

$$x_t = f(t) + \xi z_t, f(t) = \begin{cases} \mu, & 0 \leqslant t \leqslant t_0 - 1; \\ \mu + \tau, & t \geqslant t_0. \end{cases} \tag{0.1}$$

Здесь $z_t \in R, \quad t = 0, ..., T-1$ – независимые одинаково распределенные случайные величины с нулевым математическим ожиданием и конечной дисперсией σ^2, имеющие распределение Стьюдента $t(n)$, где n – число степеней свободы; ξ – уровень шума; t_0 – момент времени, в который происходит скачкообразное изменение; τ – величина скачка. Заметим, что среднее временного ряда $E\{x_t\} = f(t)$ имеет скачок τ в момент времени t_0.

Определим нулевую гипотезу H_0, состоящую в том, что $\tau = 0$, т. е. анализируемый временной ряд не имеет скачкообразных изменений среднего, и альтернативную гипотезу $H_1 = \overline{H_0}$, состоящую в том, что $\tau \neq 0$, т. е. в момент времени t_0 временной ряд имеет скачкообразное изменение.

0.1.3. Статистические свойства вейвлет-коэффициентов

Дискретное вейвлет-преобразование временного ряда (0.1) определяется выражением [2]

$$d_{j,k}^{(\psi)} = \sum_{t=0}^{T-1} x_t \psi_{j,k}(t), \ j = 1, ..., M, \ k = 0, ..., 2^{M-j} - 1, \tag{0.2}$$

где $\psi_{j,k}(t) = 2^{-\frac{j}{2}} \psi(2^{-j}t - k)$, $\psi(t)$ – базисный вейвлет, j – параметр масштаба (уровень разрешения), k – параметр сдвига. В настоящей работе будем использовать вейвлет Хаара, который, как отмечается в [2], является наиболее подходящим для обнаружения скачкообразных изменений:

$$\psi_{j,k}(t) = \begin{cases} 2^{-j/2}, & 2^j k \leqslant t \leqslant 2^j (k + 1/2); \\ -2^{-j/2}, & 2^j (k + 1/2) \leqslant t \leqslant 2^j (k + 1); \\ 0, & . \end{cases}$$

Утверждение. *Коэффициенты вейвлет-преобразования (0.2) временного ряда (0.1), построенные с использованием вейвлета Хаара, на каждом уровне разрешения* $j = 1, ..., M$ *независимы.*

Доказательство. Коэффициенты вейвлет-преобразования (0.2), построенные с использованием вейвлета Хаара, можно представить в виде

$$d_{j,k}^{(\psi)} = \sum_{t=0}^{T-1} x_t \psi_{j,t}(t) = 2^{-\frac{j}{2}} \left(\sum_{t=2^j k}^{2^j \left(k+\frac{1}{2}\right)} x_t - \sum_{t=2^j \left(k+\frac{1}{2}\right)}^{2^j (k+1)-1} x_t \right).$$

На каждом уровне jвычисляется 2^{M-j} коэффициентов, каждый из них построен по интервалу временного ряда длиной 2^j, причем они не пересекаются. Так как значения исходного ряда независимы, то отсюда следует независимость вейвлет-коэффициентов $d_{j,k}^{(\psi)}$ и $d_{j,l}^{(\psi)}$, $k \neq l$, на каждом уровне разрешения j.

Заметим, что вейвлет-коэффициенты (0.2) на каждом уровне разрешения одинаково распределены как борелевские функции от одинаково распределенных случайных величин.

0.1.4. Критерии обнаружения скачкообразных изменений

На основании статистических свойств вейвлет-коэффициентов построены три критерия обнаружения скачкообразных изменений временных рядов, наблюдаемых с шумом.

0.1.4.1. Критерий, основанный на превышении вейвлет-коэффициентами порогового значения

Критерий обнаружения скачкообразных изменений основан на предположении о том, что вейвлет-коэффициенты временного ряда (0.1) превышают некоторое пороговое значение в момент скачка t_0.

В работе [3] показано: если $y_i,\ i = 0, ...,n-1$, являются независимыми в совокупности, одинаково распределенными случайными величинами, то

$$P\left((y_i - u)_+ \leqslant x \,|y_i > u\right) = 1 - P\left((y_i - u)_+ \geqslant x \,|y_i > u\right) \approx H\left(x, \rho, \gamma\right)$$

для всех $i = 0, ..., n-1$ и достаточно больших u. Здесь $(x-u)_+ = max\left(x-u, 0\right)$ и

$$H\left(x, \rho, \gamma\right) = \begin{cases} 1 - \left(1 - \gamma x/\rho\right)^{1/\gamma}, & \gamma \neq 0; \\ 1 - e^{-x/\rho}, & \gamma = 0 \end{cases}$$

есть функция распределения обобщенного распределения Парето с параметрами формы $\gamma \in R$ и масштаба $\rho > 0$ [5], где $0 \leqslant x \leqslant \infty$ при $\gamma \leqslant 0$ и $0 \leqslant x \leqslant \rho/\gamma$ при $\gamma > 0$.

Плотность распределения абсолютных значений вейвлет-коэффициентов, превышающих порог u, представляет собой "хвост" плотности распределения Стьюдента, расположенный правее точки u. Будем аппроксимировать эту плотность плотностью обобщенного распределения Парето, график которой при $\gamma < 0,5$ вогнут и убывает [5]. Поэтому можно предположить, что более точная аппроксимация будет достигаться при выборе порога u правее точки перегиба плотности распределения Стьюдента, так как при этом вид плотностей совпадает. Поэтому для выбора промежутка аппроксимации определим точку перегиба плотности распределения Стьюдента

$$p\left(x\right) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Значение первой производной плотности распределения Стьюдента задается выражением

$$p'\left(x\right) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+3}{2}}\left(-\frac{n+\ 1}{2}\right)\frac{2x}{n}.$$

Обозначим $Z\left(n\right) = \frac{2\Gamma\left(\frac{n+1}{2}\right)\left(-\frac{n+1}{2}\right)}{n\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}$.

Вычислим значение второй производной плотности:

$$p''(x) = Z(n)\left(\left(1+\frac{x^2}{n}\right)^{-\frac{n+3}{2}} x\right)' = Z(n)\left(1+\frac{x^2}{n}\right)^{-\frac{n+5}{2}}\left(1+\frac{x^2}{n}-\frac{n+3}{n}x^2\right) = 0.$$

Отсюда следует, что значение точки перегиба функции плотности равно

$$x = \pm\sqrt{\frac{n}{n+2}}. \tag{0.3}$$

Для построения оценок параметров распределения шума используем высокочастотный уровень разрешения ($j = 1$), на котором распределение вейвлет-коэффициентов близко к распределению шума [2].

Построим оценки для параметров ξ и n по методу моментов. Обозначим: M_2 и M_4 – второй и четвертый выборочные моменты распределения вейвлет-коэффициентов уровня разрешения $j = 1$, и получаем следующие выражения:

$$M_2 = \frac{1}{2^{M-1}}\sum_{k=0}^{2^{M-1}-1}\left(d_{1,k}^{(\psi)}\right)^2 = \frac{\xi^2 n}{n-2}, \xi^2 = \frac{M_2(n-2)}{n};$$

$$M_4 = \frac{1}{2^{M-1}}\sum_{k=0}^{2^{M-1}-1}\left(d_{1,k}^{(\psi)}\right)^4 = \frac{\xi^4 2n^2}{(n-2)(n-4)} = \frac{2M_2^2(n-2)}{(n-4)}.$$

Значения параметров определяются следующим образом:

$$n = 2\left(\frac{M_4}{M_2^2}-1\right)\Big/\left(\frac{M_4}{2M_2^2}-1\right), \xi^2 = \frac{M_4 M_2}{2(M_4 - M_2^2)}.$$

С учетом (0.1) и (0.3) находим значение порога:

$$u = \xi\sqrt{\frac{n}{n+2}} = \frac{M_2 M_4}{2(M_4 - M_2^2)}\sqrt{\frac{2}{3}+\frac{2}{3\left(3\frac{M_4}{M_2^2}-4\right)}}. \tag{0.4}$$

Выражение (0.4) можно записать в следующем виде:

$$u = \frac{M_2}{2\left(1-\frac{M_2^2}{M_4}\right)}\sqrt{\frac{2}{3}+\frac{2}{3\left(3\frac{M_4}{M_2^2}-4\right)}}.$$

Очевидно, что $\lim\limits_{M_4\to\infty} u = \frac{M_2}{\sqrt{6}}$. Поэтому формулу (0.4) для определения порога u можно применять для всех $n \geqslant 3$, так как в этом случае существует второй момент вейвлет-коэффициентов (в силу существования второго момента распределения Стьюдента с числом степеней свободы $n \geqslant 3$), даже если четвертый момент $M_4 \underset{T\to\infty}{\longrightarrow} \infty$ (для числа степеней свободы $n = 3, 4$).

Пусть $q_{(j,1)} \geqslant q_{(j,2)} \geqslant ... \geqslant q_{(j,2^{M-j}-1)}$ – упорядоченные абсолютные значения вейвлет-коэффициентов $q_{j,k} = \left(d_{j,k}^{(\psi)}\right)$, $k = 0, ..., 2^{M-j}-1$, для каждого уровня разрешения $j = 1, ..., M$. Определим статистики $r_{j,l} = q_{(j,l)} - u$ и такое число L_j, что $r_{j,l} > 0, l = 0, ..., L_j - 1$.

Исходные гипотезы H_0 и H_1 заменим набором частных гипотез H_{0j} и H_{1j}:

$$H_{0j}: \ r_{j,l} \leqslant C_{j,l};$$
$$H_{1j}: \ \text{иначе},$$

где $C_{j,l}$, $l = 0, ..., L_j - 1$, – некоторые пороговые значения.

Проверка гипотез H_{0j} и H_{1j} на каждом уровне разрешения $j = 1, ..., M$ эквивалентна применению набора из M критериев. Частные гипотезы H_{0j} и H_{1j} проверяются следующим образом: принимается гипотеза H_{0j}, если $r_{j,l} \leqslant C_{j,l}$ для всех $l = 0, ..., L_j - 1$, и отвергается в противном случае. Гипотеза H_0 принимается в том случае, если принимаются гипотезы H_{0j} для всех $j = 1, \ldots, M$, и отвергается в противном случае.

Пусть α – уровень значимости критерия. Так как вейвлет-коэффициенты на различных уровнях разрешения зависимы, то согласно [6], уровень значимости α^* для проверки гипотез H_{0j}, H_{1j} на каждом уровне разрешения должен быть выбран таким, чтобы выполнялось условие $\alpha^* \leqslant \alpha/M$. Следует также заметить, что условие $\sum\limits_{l=0}^{L_j-1} P\left(r_{j,l} > C_{j,l}\right) = \alpha^*$ выполняется в силу независимости вейвлет-коэффициентов на каждом уровне разрешения.

Вычислим пороговые значения $C_{j,l}$ для проверки частной гипотезы H_{0j} из условия $P\left(r_{j,l} > C_{j,l}\right) = \alpha^*/L_j$.

Последовательно имеем:

$$P\left(r_{j,l} > C_{j,l}\right) = 1 - P\left(r_{j,l} \leqslant C_{j,l}\right) = 1 - P\left(max\left(r_{j,l}, ..., r_{j,L_j-1}\right) \leqslant C_{j,l}\right) =$$

$$= 1 - P\left(r_{j,l} \leqslant C_{j,l}, ..., r_{j,L_j-1} \leqslant C_{j,l}\right) = 1 - P\left(\cap \sum_{k=l}^{L_j-1} \left(r_{j,k} \leqslant C_{j,l}\right)\right) =$$

$$= 1 - \prod_{k=l}^{L_j-1} P\left(r_{j,k} \leqslant C_{j,l}\right) = 1 - \prod_{k=l}^{L_j-1} H\left(C_{j,l}; \rho, \gamma\right) = 1 - \left(H\left(C_{j,l}; \rho, \gamma\right)\right)^{L_j-l} = \alpha^*/L_j.$$

Отсюда $C_{j,l} = H^{-1}\left(\left(1 - \frac{\alpha^*}{L_j}\right)^{\frac{1}{L_j-l}}\right)$, где $H^{-1}\left(x\right) = \begin{cases} \frac{\rho}{\gamma}\left(1 - (1-x)^{\gamma}\right), \gamma \neq 0; \\ -\rho \ ln\left(1-x\right), \gamma = 0. \end{cases}$

Для построения критерия необходимо оценить параметры распределения Парето γ и ρ. Используем метод моментов для обобщенного распределения Парето из [5] и получим оценки: $\hat{\rho} = 0.5\overline{x}(\overline{x}^2/s^2 + 1)$ и $\hat{\gamma} = 0.5(\overline{x}^2/s^2 - 1)$. Выборочные среднее $\overline{x}$ и дисперсия s^2 вычисляются с использованием статистик $r_{1,l}$, $l = 0, ..., L_1 - 1$, для высокочастотного уровня разрешения ($j = 1$).

0.1.4.2. Критерий, основанный на максимуме сумм вейвлет-коэффициентов

Определим статистику

$$V_j = \sqrt{2^j} \sum_{k=0}^{2^{(M-j)}} d_{j,k}^{(\psi)}.$$

Следует заметить, что вейвлет-коэффициенты $d_{j,k}^{(\psi)}$ при фиксированном j и k не зависят от длины временного ряда . Так как вейвлет-коэффициенты на каждом уровне разрешения являются независимыми одинаково распределенными случайными величинами, т. е.

удовлетворяют условиям центральной предельной теоремы, то можно показать, что при выполнении нулевой гипотезы H_0 статистика $\frac{V_j}{\hat{\theta}\sqrt{T}}$ имеет асимптотически стандартное нормальное распределение $N(0,1)$, где $\theta = \xi\sigma$.

Пусть $V = max\left(|V_1|, |V_2|, ..., |V_M|\right)$. Найдем пороговое значение Δ для критерия из условия $P\left(\frac{V}{\hat{\theta}\sqrt{T}} > \Delta\right) = \alpha$. Проводя такие же рассуждения, как и для предыдущего критерия, найдем пороговое значение $\Delta = -\Phi^{-1}\left(\left(1-(1-\alpha)^{1/M}\right)/2\right)$, где $\Phi^{-1}(\cdot)$ – квантиль стандартного нормального распределения.

В качестве оценки стандартного отклонения θ с целью исключения влияния аномальных наблюдений будем использовать ее робастный аналог [7]:

$$\hat{\theta} = \underset{k}{median}\left|d_{1,k}^{(\psi)} - \underset{k}{median}\left(d_{1,k}^{(\psi)}\right)\right| / 0,6745, \tag{0.5}$$

где median$(\cdot)$ – символ выборочной медианы.

Тогда решающее правило определяется следующим образом: принимается гипотеза H_0, если $\frac{V}{\theta\sqrt{T}\sqrt{T}} \leqslant \Delta$, и гипотеза H_1 в противном случае.

0.1.4.3. Критерий, основанный на сумме вейвлет-коэффициентов

Рассмотрим статистику $Q = \sum\limits_{j=0}^{M} V_j$. Тогда статистика $Q/\sqrt{MT}$ имеет асимптотически нормальное распределение $N(0,\theta^2)$ как сумма асимптотически нормальных случайных величин с распределением $\frac{V_j}{\sqrt{T}} \sim N\left(0,\theta^2\right)$.

Найдем пороговое значение Δ для данного критерия обнаружения разладки аналогично предыдущему критерию: $\Delta = \Phi^{-1}(1-\alpha/2)$. Как и для критерия, основанного на максимуме сумм вейвлет-коэффициентов, в этом случае в качестве оценки для θ используем робастную оценку (0.5).

Решающее правило состоит в следующем: принимается гипотеза H_0, если $\frac{|Q|}{\hat{\theta}\sqrt{MT}} \leqslant \Delta$, и гипотеза H_1 в противном случае.

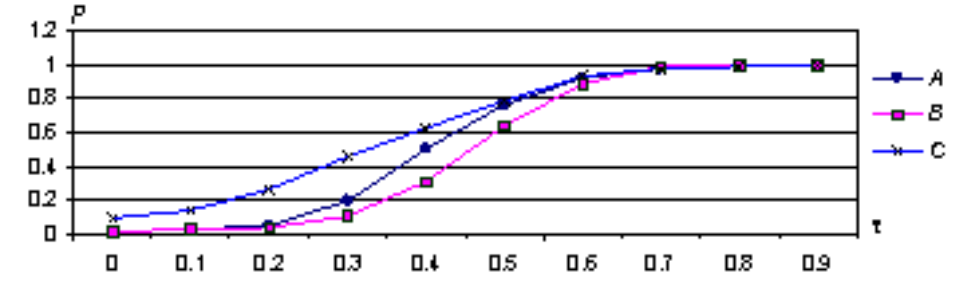
0.1.5. Исследование эффективности критериев обнаружения скачкообразных изменений

Для статистического оценивания вероятностей ошибок первого и второго рода выполнялось моделирование временных рядов с шумом z_t, имеющим распределение Стьюдента с числом степеней свободы $n \in \{3, 7, 15\}$, а также стандартное нормальное распределение $N(0,1)$. Моделируемые временные ряды состояли из двух однородных фрагментов длиной $T_1 = \frac{T}{3}, T_2 = \frac{2T}{3}$, скачок $\tau \in \{0.1, 0.2, ..., 0.9\}$ моделировался в момент времени $t_0 = T/3$. Эксперименты проводились для различных длин временного ряда.

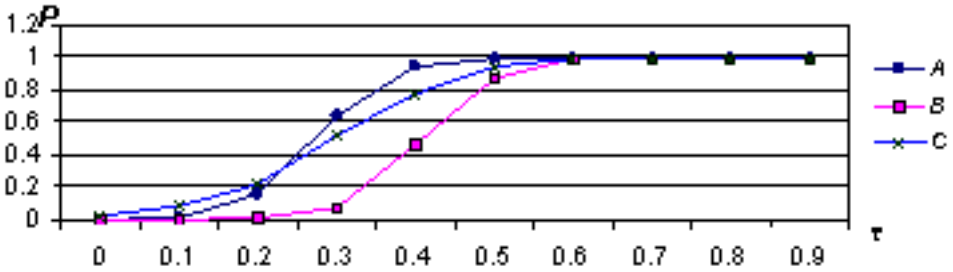
Полученные оценки для параметра γ критерия, основанного на превышении вейвлет-коэффициентами порогового значения, попали в интервал [-0.47, ..., 0.19], что подтвердило предположение относительно значения этого параметра ($\gamma < 0{,}5$).

Для сравнения мощности различных критериев длина временного ряда была выбрана равной $T = 2^{11}$, число экспериментов $K = 1000$, параметр $\xi = 1$. Уровень значимости для критериев полагался равным $\alpha = 0.05$. На рис. 1 представлены зависимости значения мощности критериев от величины скачка τ для различных распределений шума (A

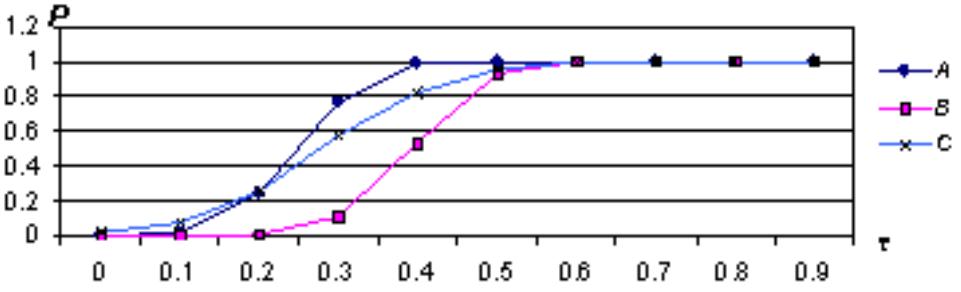
– критерий, основанный на превышении вейвлет-коэффициентами порогового значения; B – критерий, основанный на максимуме сумм вейвлет-коэффициентов; C – критерий, основанный на сумме вейвлет-коэффициентов).



а) распределение Стьюдента с числом степеней свободы n= 3



б) распределение Стьюдента с числом степеней свободы n= 7



в) распределение Стьюдента с числом степеней свободы n= 15

г) стандартное нормальное распределение

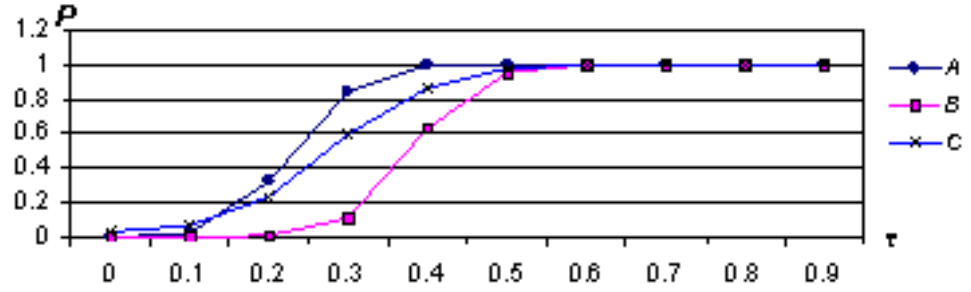


Рисунок 1. Зависимость оценок вероятностей ошибок первого рода и мощности критерия от величины скачка τ

Как видно из результатов, показанных на рис. 1, оценки вероятностей ошибок первого рода для критериев и не превосходят уровня значимости. Для критерия оценки вероятностей ошибок первого рода являются максимальными из всех критериев, а в случае самых "тяжелых хвостов" (n=3) превосходит уровень значимости. Это говорит о чувствительности критерия к распределению шума. Значения оценок мощности предложенных критериев увеличиваются с ростом величины скачка τ и числа степеней свободы n распределения Стьюдента. Отметим, что для большой величины скачка ($\tau > 0.6$) наибольшую оценку мощности имеет критерий, основанный на превышении вейвлет-коэффициентами порогового значения.

В качестве примера применения критериев для решения реальной задачи были взяты ежедневные данные об изменении индекса Доу-Джонса за период с 1945 по 1983 г., представленные на рис.2 (8096 значений). Критерий, основанный на превышении вейвлет-коэффициентами порогового значения, обнаружил скачкообразные изменения для отсчетов t=4096 (1960 г.) и t=7680 (1974 г.). Критерии, основанные на максимуме сумм вейвлет-коэффициентов и на сумме вейвлет-коэффициентов, также показали наличие скачкообразных изменений в этом временном ряду.

Рисунок 2. Изменение индекса Доу-Джонса за период с 1945 по 1983 г.

Заключение

Исследованы статистические свойства вейвлет-коэффициентов временного ряда, полученных с использованием вейвлета Хаара. На основании рассмотренных статистик от

вейвлет-коэффициентов построены три критерия обнаружения скачкообразных изменений временных рядов, наблюдаемых с аддитивным шумом, имеющим более “тяжелые хвосты”, чем нормальное распределение. Методом статистического моделирования для случая, когда шум имеет распределение Стьюдента, получены оценки вероятностей ошибок первого рода и мощности критериев. Результаты проведенных экспериментов показывают эффективность предложенных критериев обнаружения скачкообразных изменений среднего временных рядов.

Критерий, основанный на превышении вейвлет-коэффициентами порогового значения может быть использован в разрабатываемом пакете прикладных программ по прогнозированию в разделе предварительного анализа временных рядов (обнаружение скачкообразных изменений).

Список литературы

Kharin Yu.S. Detection of spectral change-point in a two-dimensional time series / Yu.S. Kharin, M.S. Abramovich // Optoelectronics, Instrumentations and Data Processing. – 1999. – N2. – P. 45–53.

Ogden R.T. Testing for abrupt jumps with wavelets / R.T. Ogden, C. Cheng // Proceedings of the 29th Symposium on the Interface, Interface Foundation of North America. – 1997. – P. 138–142.

Raimondo M. Wavelet shrinkage via peaks over threshold / M. Raimondo // Interstat. – 2002. – N5. – P. 1–19.

Raimondo M. Minimax estimation of sharp change points / M. Raimondo // The annals of Statistics. – 1998. – Vol. 26. – No. 4. – P. 1379–1397.

Hosking J.R.M. Parameter and quantile estimation for the generalized Pareto distribution / J.R.M. Hosking, J.R. Wallis // Technometrics. – 1987. – N29. – P. 339–349.

Кокс Д. Теоретическая статистика / Д. Кокс, Д. Хинкли – М.: Наука, 1978. – 560 с.

Хьюбер П. Робастность в статистике / П. Хьюбер – М: Мир, 1984 – 303 с.