УДК 519.246.8:004.67

Абрамович М. С., Мицкевич М.Н.

ОБНАРУЖЕНИЕ “СКАЧКООБРАЗНЫХ” ИЗМЕНЕНИЙ СРЕДНЕГО ВРЕМЕННЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ ХААРА

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

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. Критерии, основанные на использовании максимальных значений вейвлет-коэффициентов 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.

1. Математическая модель “скачкообразных” изменений

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

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

Здесь $z_t \in R$, $t = 0,...,T-1$ — 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 $\tau \ne 0$, т.е. в момент времени t_0 временной ряд имеет “скачкообразное” изменение.

2. Статистические свойства вейвлет-коэффициентов

Дискретное вейвлет-преобразование временного ряда (1) определяется выражением [2]:

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

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

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

Утверждение 1. Коэффициенты вейвлет-преобразования (2) временного ряда (1), построенные с использованием вейвлета Хаара, на каждом уровне разрешения $j=1,\dots,M$ независимы.

Доказательство. Коэффициенты вейвлет-преобразования (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 j вычисляется 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 (2) на каждом уровне разрешения одинаково распределены как борелевские функции от одинаково распределенных случайных величин.

3. Критерии обнаружения “скачкообразных” изменений

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

Этот критерий обнаружения “скачкообразных” изменений основан на предположении о том, что вейвлет-коэффициенты временного ряда (1) превышают некоторое пороговое значение в момент “скачка” t_0.

В [3] показано, что если y_i, $i=0,\dots,n-1$, являются независимыми в совокупности одинаково распределенными случайными величинами, то:
$P\left((y_i-u)_+\le x \middle| y_i>u\right)=1-P\left((y_i-u)_+\ge x \middle| y_i>u\right)\approx H(x,\rho,\gamma)$ для всех $i=0,\dots,n-1$ и достаточно больших u. Здесь $(x-u)_+=max(x-u,0)$ и

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

— функция распределения обобщенного распределения Парето с параметрами формы $\gamma\in R$ и масштаба $\rho>0$ [4], где $0\le x\le\infty$ при $\gamma\le 0$ и $0\le x\le \rho/\gamma$ 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 γ<0.5 вогнут и убывает [4]. Исходя из этого, 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

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

Значение первой производной плотности распределения Стьюдента определяется выражением:

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

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

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:

$$x=\pm\sqrt{\frac{n}{n+2}}. \qquad (3)$$

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.

Имеем:

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

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

Значения параметров определяются следующим образом:

$$n=\frac{2\left(\frac{M_4}{M_2^2}-1\right)}{\left(\frac{M_4}{2M_2^2}-1\right)},\ \xi^2=\frac{M_4M_2}{2\left(M_4-M_2^2\right)}.$$

С учетом (1) и (3) находим значение порога:

$$u=\xi\sqrt{\frac{n}{n+2}}=\frac{M_2M_4}{2\left(M_4-M_2^2\right)}\sqrt{\frac{2}{3}+\frac{2}{3\left(3\frac{M_4}{M_2^2}-4\right)}}. \qquad (4)$$

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

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

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

$H_{0j} : r_{j,l} \leq C_{j,l}$;

H_{1j} :иначе,

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

Проверка гипотез H_{0j} и H_{1j} на каждом уровне разрешения $j = 1,\ldots,M$ эквивалентна применению набора из M критериев. Частные гипотезы H_{0j} и H_{1j} проверяются следующим образом: принимается гипотеза H_{0j}, если $r_{j,l} \leq C_{j,l}$ для всех $l = 0,\ldots,L_j - 1$, 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 [5], уровень значимости α^* для проверки гипотез H_{0j}, H_{1j} на каждом уровне разрешения должен быть выбран таким, чтобы выполнялось условие $\alpha^* \leq \alpha / M$. Следует также заметить, что условие $\sum_{l=0}^{L_j-1} P\left(r_{j,l} > C_{j,l}\right) = \alpha^*$ 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:

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

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

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

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

Для построения критерия необходимо оценить параметры распределения Парето γ и ρ. Используем метод моментов для обобщенного распределения Парето из [4] и получим оценки: $\hat{\rho} = 0.5\bar{x}(\bar{x}^2 / s^2 + 1)$ и $\hat{\gamma} = 0.5(\bar{x}^2 / s^2 - 1)$. Выборочные среднее $\bar{x}$ и дисперсия s^2 вычисляются с использованием статистик $r_{1,l}$, $l = 0,\ldots,L_1 - 1$, для высокочастотного уровня разрешения (j=1).

3.2. Критерий, основанный на максимуме сумм вейвлет-коэффициентов

Определим следующую 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 T. 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 $V = \max(|V_1|,|V_2|,...,|V_M|)$. 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 [6]:

$$\hat{\theta} = \underset{k}{\mathrm{median}}\left|d_{1,k}^{(\psi)} - \underset{k}{\mathrm{median}}\left(d_{1,k}^{(\psi)}\right)\right| / 0.6745, \quad (5)$$

где median(·) — символ выборочной медианы.

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

3.3. Критерий, основанный на сумме вейвлет-коэффициентов

Рассмотрим статистику: $Q = \sum_{j=0}^{M} V_j$. 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 этого 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 (5).

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

4. 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 (γ<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 рисунках 1-4 представлены зависимости значения мощности критериев от величины скачка τ для различных распределений шума (А – критерий, основанный на превышении вейвлет-коэффициентами порогового значения; В – критерий, основанный на максимуме сумм вейвлет-коэффициентов; С – критерий, основанный на сумме вейвлет-коэффициентов).

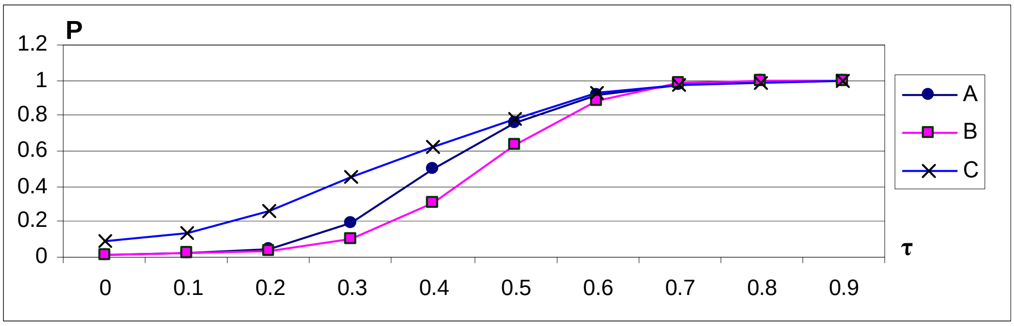

Рисунок 1. Зависимость оценок вероятностей ошибок первого рода и мощности критерия от величины скачка τ для распределения Стьюдента с числом степеней свободы n = 3

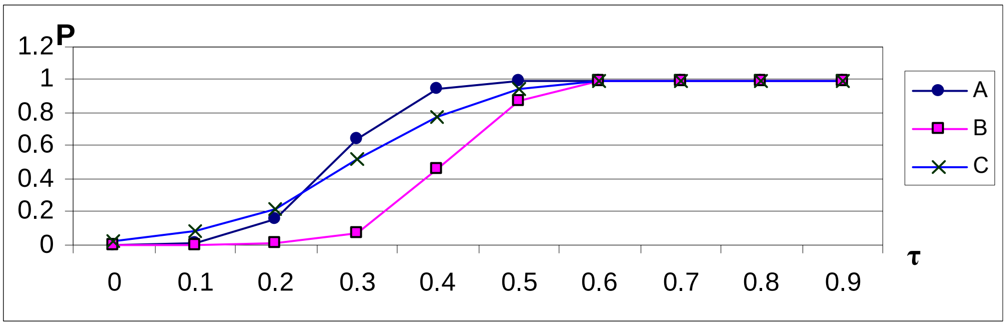

Рисунок 2. Зависимость оценок вероятностей ошибок первого рода и мощности критерия от величины скачка τ для распределения Стьюдента с числом степеней свободы n = 7

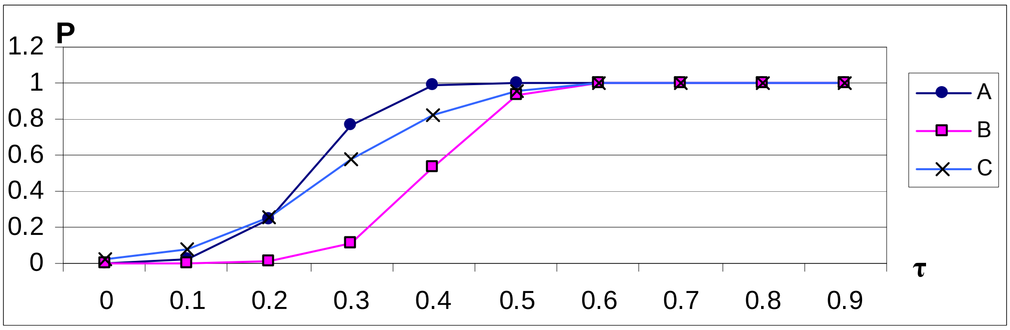

Рисунок 3. Зависимость оценок вероятностей ошибок первого рода и мощности критерия от величины скачка τ для распределения Стьюдента с числом степеней свободы n = 15

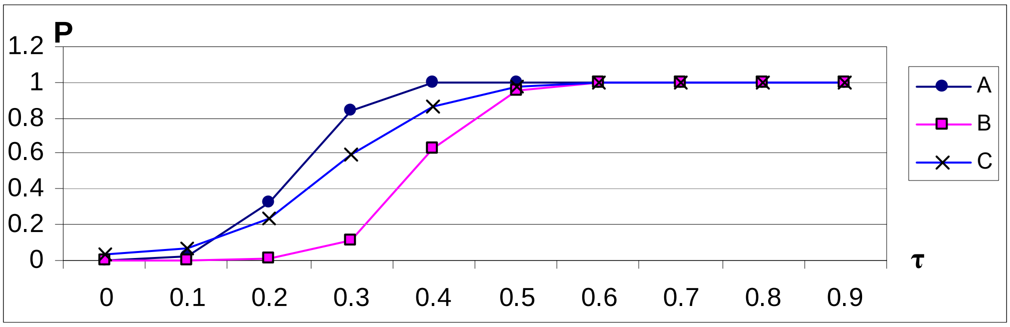

Рисунок 4. Зависимость оценок вероятностей ошибок первого рода и мощности критерия от величины скачка τ стандартного нормального распределения

Как видно из результатов, показанных на графиках 1-4, оценки вероятностей ошибок первого рода для критериев А и В не превосходят уровня значимости. Для критерия С 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 C 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.

Заключение

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

Работа частично поддержана ГКПНИ "Инфотех" (задание Инфотех 15).

Литература

1. Kharin Yu.S., Abramovich M.S. Detection of spectral change-point in a two-dimensional time series // Optoelectronics, Instrumentations and Data Processing. - 1999. - N2. - pp. 45-53.
2. Ogden R.T., Cheng C. Testing for abrupt jumps with wavelets // Proceedings of the 29th Symposium on the Interface, Interface Foundation of North America. - 1997. - pp. 138-142.
3. Raimondo M. Wavelet shrinkage via peaks over threshold // Interstat. - 2002. - May - pp. 1-19.
4. Hosking J.R.M., Wallis J.R. Parameter and quantile estimation for the generalized Pareto distribution // Technometrics. - 1987. - N29. - pp. 339-349.
5. Кокс Д., Хинкли Д. Теоретическая статистика. - М.: Наука, 1978. - 560 с.
6. Хьюбер П. Робастность в статистике. - М: Мир, 1984 - 303 с.

Абрамович М. С., Мицкевич М.Н. **Обнаружение “скачкообразных” изменений среднего временных рядов с использованием вейвлет-преобразования Хаара**

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

Рис. 4. Библиогр. – 6 назв.

M. S. ABRAMOVICH, M. M. MITSKEVICH

JUMP DETECTION IN TIME SERIES MEAN USING HAAR WAVELET TRANSFORMATION

Summary

Wavelet-based tests for time series with noise were developed. Noise is assumed to have more heavy tails than a Gaussian. The efficiency of these tests is analyzed by statistical modeling.

Абрамович Михаил Семенович – заведующий лабораторией статистического анализа и моделирования Национального научно-исследовательского центра прикладных проблем математики и информатики, кандидат физ.-мат. наук, доцент. Раб. тел.: 2095481. Дом. тел.: 2505235. e-mail: abramovichms@bsu.by.

Мицкевич Михаил Николаевич – младший научный сотрудник Национального научно-исследовательского центра прикладных проблем математики и информатики. Раб. тел.: 2095481. Дом. тел.: 2121638. e-mail: mitskevich_m@mail.ru.