Программы WBRCM и WBRCM_mult

оценки эволюции робастной вейвлетной меры когерентности многомерного временного ряда в скользящем временном окне.

А.А. Любушин, доктор физ.-мат. наук
Институт физики Земли РАН им. О.Ю.Шмидта,
123995, Москва, Большая Грузинская, 10; факс: +007-499-2556040;
e-mail: lyubushin@yandex.ru
http://AlexeyLyubushin.narod.ru/Index.htm

Метод, излагаемый ниже [*Любушин, 2000, 2002 Любушин, 2007*], строит оценку масштабно-зависимой меры когерентного поведения многомерного временного ряда в скользящем временном окне заданной длины, аналогичную спектральной мере, вычисляемой программой **SpectCohMes**, но основанную на разложении сигналов по ортогональным финитным базисным функциям – вейвлетам [*Press et al., 1996; Mallat, 1998*].

Строятся 2 робастные вейвлетные меры когерентности: одна зависит от выбора ортогонального базиса финитных функций (программа **WBRCM**), во второй мере (программа **WBRCM_mult**) производится усреднение мер по 10 используемым базисам, в результате чего один свободный параметр пропадает. Аббревиатура **WBRCM** означает “Wavelet-Based Robust Coherence Measure”, приставка “mult” означает “multi-bases” (многобазисность).

Используется словарь из 17 вейвлетов: 10 обычных ортогональных вейвлетов Добеши с порядками от 2 до 20 и 7 т.н. «симлетов» - модификаций вейвлетов Добеши, в которых форма базисных функций является более симметричной, чем для обычных вейвлетов [*Mallat, 1998*]. Порядок вейвлета равен числу коэффициентов в дискретных зеркальных фильтрах, реализующих расщепление уровня детальности на 2 частотных диапазона. Число обнуляемых моментов (или степень гладкости) для материнской базисной функции вейвлета равно половине порядка вейвлета. Использование более высоких порядков сопряжено с численной неустойчивостью. Симлеты обладают теми же свойствами компактности, ортогональности, полноты и гладкости, что и вейвлеты Добеши, но для порядков от 2-го до 6-го они совпадают с обычным ортогональным базисом Добеши, а затем, для порядков от 8-го до 20-го, появляются различия в форме базисной функции. Вследствие этого, общее число возможных используемых вариантов ортогональных компактных базисных функций в программе **WBRCM** равно 17.

Описание методов, лежащих в основе программ **WBRCM** и **WBRCM_mult**, фактически является в значительной мере также описание методов оценки вейвлет-агрегированных сигналов, реализованных в программах **AggW** и **AggWR**.

Напомним, что для сигнала длиной $N = 2^m$ отсчетов существует ровно m уровней детальности вейвлет-разложения. Если N не равно 2^m**,** то дополним сигнал нулями до длины, которая будет равна 2^m, где m – минимальное целое число, для которого $N \le 2^m$. Самым мелкомасштабным (высокочастотным) уровнем детальности является 1-й. Каждый уровень детальности с номером $\beta = 1,...,m$ содержит $N \cdot 2^{-\beta}$ вейвлет-коэффициентов, каждый из которых фактически является сверткой исходного сигнала с конечным дискретным фильтром, сосредоточенным в окрестности равномерной сетки узлов с шагом $2^{\beta} \cdot \Delta s$, где Δs - шаг дискретизации по времени. Ширина временной «зоны ответственности»

вейвлет-коэффициента на уровне детальности β примерно равна $\Delta T^{(\beta)} = \Delta s \cdot 2^{\beta}$, а частотный диапазон примерно лежит в интервале $[\Omega_{\min}^{(\beta)}, \Omega_{\max}^{(\beta)}] = [1/(2^{(\beta+1)}\Delta s), 1/(2^{\beta}\Delta s)]$ шириной $\Delta\Omega^{(\beta)} = 1/(2^{\beta+1} \cdot \Delta s)$. Таким образом, произведение временной неопределенности вейвлет-коэффициента $\Delta T^{(\beta)}$ на частотную неопределенность $\Delta\Omega^{(\beta)}$ не зависит от уровня детальности и равно $\frac{1}{2}$, что является выражением принципа неопределенности Гейзенберга.

Описание программ.

Обрабатываемые временные ряды должны представлять собой результаты синхронных наблюдений в виде числовых текстовых файлов, имеющих структуру «одна запись – один отсчет» (или «длинная колонка чисел»). Каждому временному ряду должен соответствовать свой файл. Если исходное данные представляют собой таблицы, то временные метки и прочая служебная информация не должны находится в первой колонке таблицы. Небольшие пропуски данных должны быть восполнены какими-то «правдоподобными» значениями. Перед запуском программы в рабочей директории должен быть создан вспомогательный файл со стандартным именем ”list”, содержащий имена анализируемых файлов, перечисленных «в столбик» (см. программу **MakeList**).

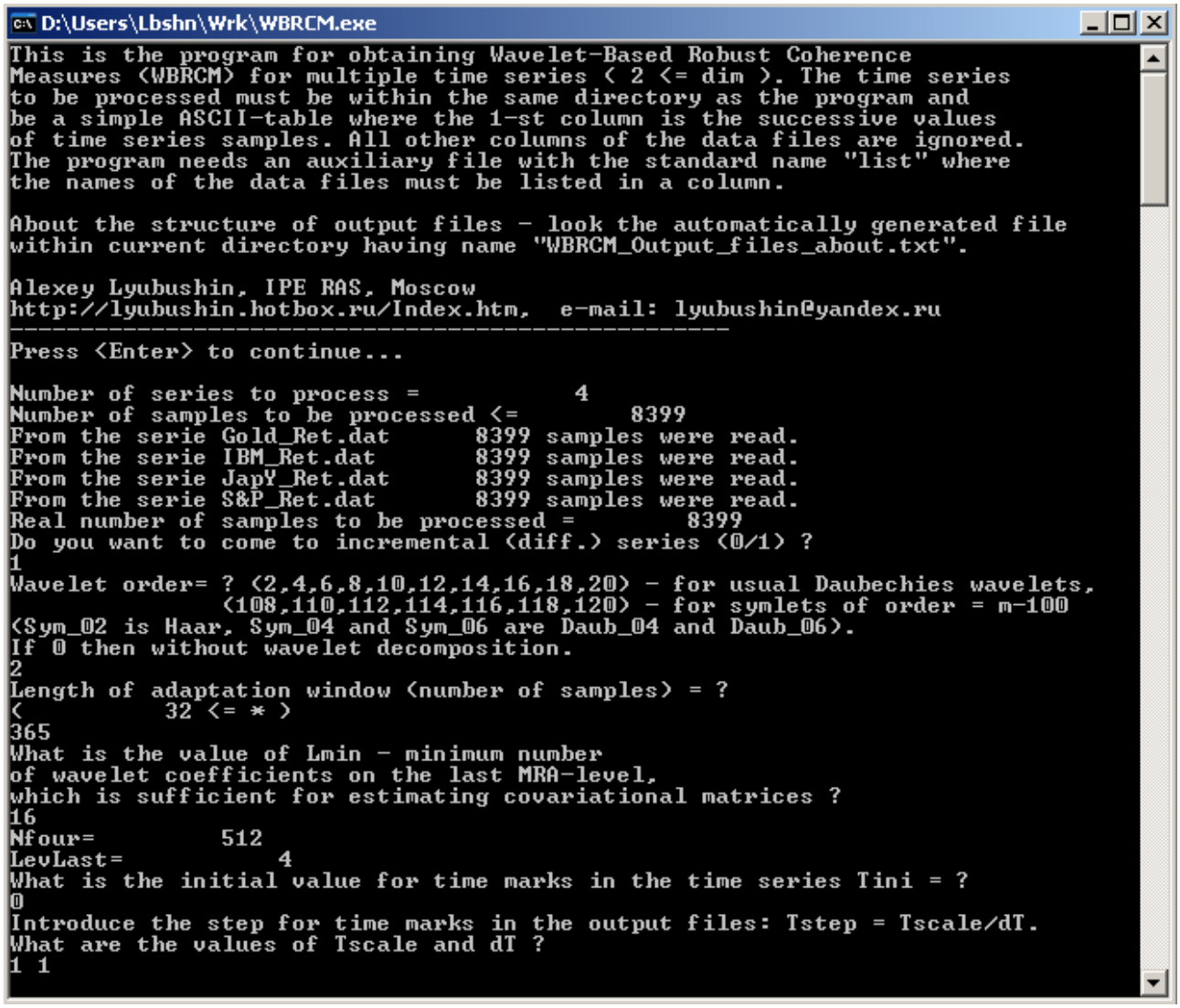
Программы считывают из файла ”list” имена файлов, содержащих анализируемые временные ряды, открывает их, последовательно считывает из них временные отсчеты и закрывает. Если файл ”list” отсутствует, то программы останавливаются с соответствующим сообщением. Если исходные файлы содержат разное число отсчетов, то обработка будет производиться по выборке длины, равной минимальной длине временных рядов.

Далее пользователю необходимо ответить на следующие вопросы:

1) Надо ли переходить внутри каждого скользящего окна к рядам в приращениях («дифференцировать ряды»). Если ряды носят существенно низкочастотный характер, то такой переход необходим. В противном случае, если доминируют низкие частоты, всегда будет наблюдаться высокая когерентность на самых старших для данной длины уровнях детальности вейвлет-разложения.
2) Необходимо ввести идентификатор используемого вейвлета, который для обычных вейвлетов Добеши равен их порядку, то есть числам 2, 4, 6, 8, 10, 12, 14, 16, 18, 20. Для того, чтобы ввести симлет, надо ввести число, равное порядку симлета плюс 100, то есть возможные ответы для симлета следующие: 108, 110, 112, 114, 116, 118, 120. Возможным ответом является также введение «нулевого порядка вейвлета», то есть 0 – это будет означать, что робастная мера когерентности будет оценена без вейвлет-разложений, а непосредственно для исходных данных. Этот пункт диалога есть только для программы **WBRCM**. Что же касается программы **WBRCM_mult**, то, поскольку в ней мера когерентности вычисляется для 10 обычных вейвлетов Добеши, а потом усредняется, там этот пункт отсутствует.
3) Следует ввести длину скользящего временного окна в числе отсчетов (взаимное смещение временных окон не вводится, поскольку оно равно 1).
4) Следует ввести значение порога представительности $L_{\min}$ – положительное целое число, смысл которого заключается в том, что оценки вычисляются лишь для тех уровней детальности вейвлет-разложения, для которых число вейвлет-коэффициентов на уровне при длине выборки, определяемой длиной временного окна, больше или равно этому пороговому значению.
5) Следует ввести значение T_{ini} начальной временной метки в выводных файлах.

6) Следует определить шаг по времени T_{step} во временных метках, введя значения T_{scale} и dT, после чего шаг по времени вычисляется по формуле $T_{step} = T_{scale} / dT$. Это удобно, например, в ситуации, когда временные ряды определены с шагом 1 сутки от начала 1900-го года, а временные метки в выводных файлах необходимо задать в годах. Тогда можно ввести $T_{ini} = 1900$, $T_{scale} = 1$, $dT = 365.25$, где параметр dT равен среднему числу суток в годе с учетом високосных лет.

Пример экранной копии диалога с программой **WBRCM** приведен ниже:

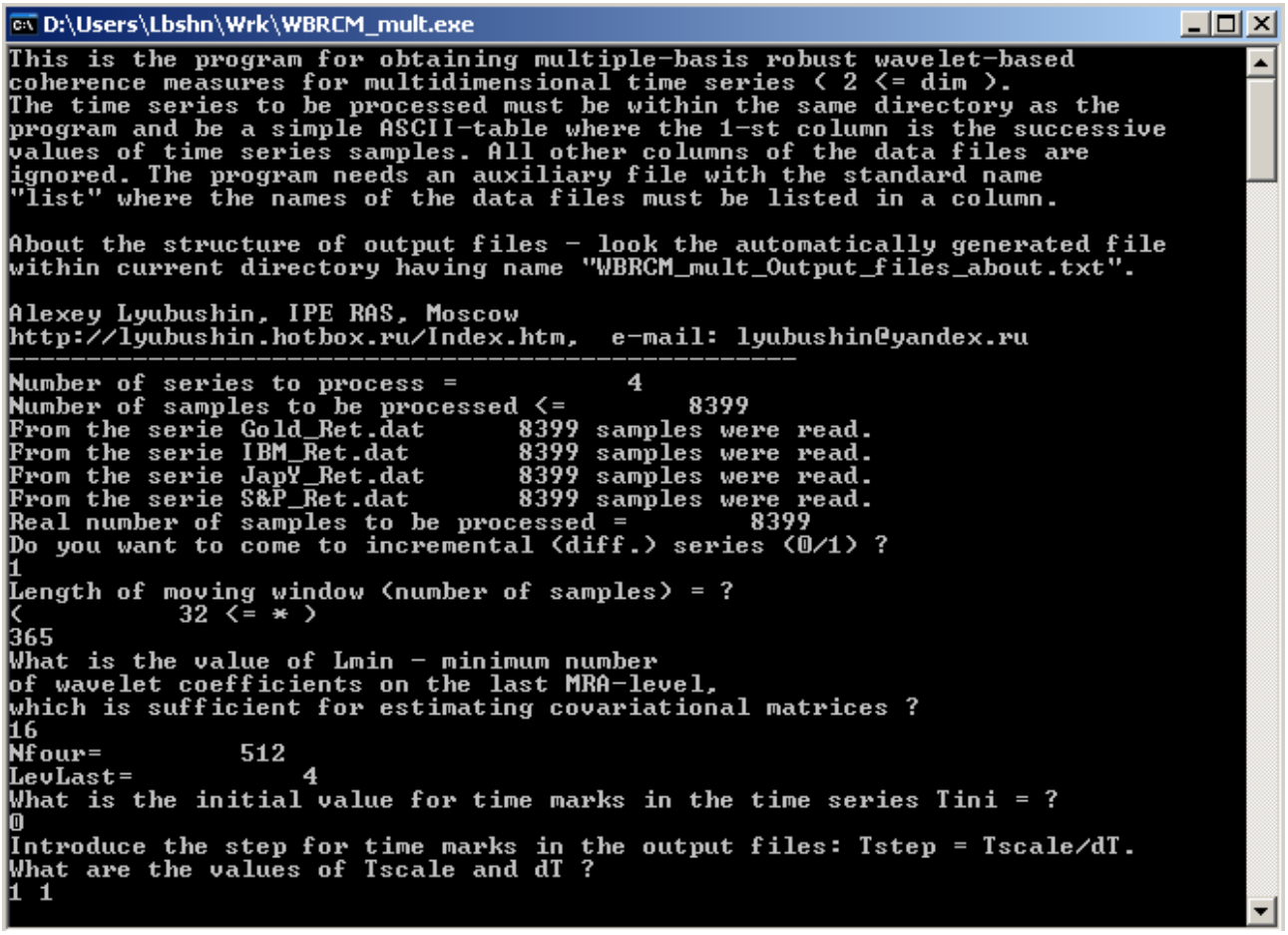


После завершения счета программа создаст в рабочей директории выводные файлы, структура которых описана также в автоматически создаваемом файле с именем “WBRCM_Output_files_about.txt”:

- "Nju_0.dat" – если порядок вейвлета задан «0», что означает что робастные канонические корреляции вычисляются в скользящем временном окне без проведения вейвелет-разложения, непосредственно для исходных сигналов;
- "Nju_01.dat", "Nju_02.dat", "Nju_03.dat" и так далее, если задан реальный порядок вейвлета. Числа 01, 02, 03, … означают номера уровней детальности вейвлет-разложения. Максимальный уровень детальности зависит от длины временного окна и значения параметра $L_{\min}$.

Структура выводных файлов "Nju_*.dat" одинакова. Если $q \geq 3$ - общее число одновременно анализируемых временных рядов, то каждый из файлов "Nju_*.dat" представляет собой числовую таблицу из $(q+3)$ колонок. Первая колонка суть временные метки, соответствующие правому концу скользящего временного окна. Вторая колонка представляет собой среднее значение (по числу анализируемых временных рядов) квадратов канонических корреляций (формула (7)) между вейвлет-коэффициентами: $\sum_{k=1}^{q} \bar{\nu}_k^2(\tau,\beta)/q$. Третья колонка является произведением абсолютных значений канонических корреляций (формула (8)) и, по сути, является робастной вейвлетной мерой когерентности, аналогичной спектральной мере когерентности в программе **SpectCohMes**. Колонки с номерами $(3+k), k=1,...,q$ являются значениями канонических корреляций $\bar{\nu}_k(\tau,\beta)$

Пример экранной копии диалога с программой **WBRCM_mult** приведен ниже:



После завершения счета программа создаст в рабочей директории выводные файлы, структура которых описана также в автоматически создаваемом файле с именем “WBRCM_mult_Output_files_about.txt”. Эти фалы имеют те же имена "Nju_01.dat", "Nju_02.dat", "Nju_03.dat" и ту же структуру, что и вышеописанные выводные файлы программы **WBRCM** с заменой величин $\bar{\nu}_k(\tau,\beta)$ на $\tilde{\bar{\nu}}_k(\tau,\beta)$ (формула (21)).

Описание метода.

Пусть $q \geq 3$ - общее число одновременно анализируемых временных рядов, $Z(t) = (Z_1(t), ..., Z_q(t))^T$, а τ - положение правого конца скользящего временного окна длиной N отсчетов. Для краткости будем употреблять словосочетание «временное окно τ» для обозначения выборок векторного временного ряда для моментов времени, удовлетворяющих неравенствам: $\tau - N + 1 \leq t \geq \tau$. Для каждого положения временного окна (смещаемого вправо на один отсчет) анализ производится независимо от анализа в других окнах. Перед вейвлет-разложением фрагментов анализируемых временных рядов, попавших в текущее временное окно, каждый из них подвергается последовательности следующих операций:

(i) устраняется общий линейный тренд в пределах временного окна;
(ii) осуществляется переход от исходных значений к приращениям между соседними значениями времени;
(iii) находится робастная выборочная оценка стандартного отклонения и осуществляется деление каждого значения на эту оценку;
(iv) производилась операция сглаживания на концах окна косинусной весовой функцией;
(v) выборка внутри окна дополняется нулями до полной длины $M = \min\{2^m : 2^m \geq N\}$ отсчетов.

Операция (i) устраняет самые сильные низкочастотные вариации сигналов, которые в пределах окна не могут быть статистически представительными. Операция (ii) перехода к приращениям является стандартным шагом в анализе временных рядов для увеличения стационарности выборок на коротких временных окнах при доминировании низких частот и поэтому она зависит от характера одновременно обрабатываемых сигналов. Если временные ряды высокочастотные, то операция (ii) опускается. Ее негативным свойством является увеличение амплитуды высокочастотного шума. Однако многомерный анализ позволяет избавиться от индивидуальных шумов, присущих только тому или иному временному ряду.

Операция (iii) строит оценку с помощью винзоризации: [*Huber, 1981*] или итеративного устранения больших выбросов: вычисление среднего значения $\bar{x}$, стандартного отклонения σ, операции срезки значений временного ряда внутри текущего окна, выпадающих за уровни $\bar{x} \pm 3\sigma$, и повторения этой последовательности 3-х операций до тех пор, пока значения $\bar{x}$ и σ не перестанут меняться. Деление каждого из сигналов внутри окна на его стандартное отклонение уравнивает различные временные ряды путем приведения энергии их вариаций к одному и тому же значению. Хотя формально последующий канонический анализ коэффициентов вейвлет-разложений инвариантен относительно преобразований масштаба временных рядов, такая операция является полезной для уменьшения ошибок округления. Операция оконного сглаживания (iv) необходима для уменьшения негативного влияния циклического эффекта дискретного вейвлет-преобразования [*Press et al., 1996*]. Наконец, последняя операция (v) необходима для последующего применения быстрого дискретного вейвлет-преобразования.

Затем осуществляется дискретное вейвлет-преобразование с использованием одного из ортогональных финитных базисов от фрагмента каждого из анализируемых временных рядов внутри текущего окна после предварительных операций (i)-(iv). После осуществления вейвлет-преобразований для каждого временного окна τ получается множество таблиц вейвлет-коэффициентов размером $q \times M_\beta$:

$$c_{k\,r}^{(\tau,\beta)}, \quad k=1,...,q; \quad r=1,...,M_\beta=2^{(m-\beta)}; \quad \beta=1,...,m \tag{1}$$

Здесь β - номер уровня детальности вейвлет-разложения. Число m - степень двойки в представлении $M=2^m$ для минимального целого числа такого вида, не меньшего, чем длина временного окна N. На каждом уровне детальности общее число коэффициентов равно $M_\beta=2^{(m-\beta)}$. Индекс k в формуле (1) соответствует положению временного интервала внутри скользящего окна, значения на котором влияют на значение коэффициента на уровне детальности β. Однако часть этих коэффициентов может соответствовать нулевому дополнению выборки в пункте (iv) предварительных преобразований. Поэтому реальное число коэффициентов на уровне β, отражающее поведение сигнала внутри окна, равно $L_\beta(N)=2^{(m-\beta)}(N/M)=2^{-\beta}N$.

Обозначим через $Q^{(\tau,\beta)}$ множество из $L_\beta(N)$ q-мерных векторов вейвлет-коэффициентов, образующих столбцы таблицы (1), отбрасывая столбцы, соответствующие дополнению нулями в операции (v). Поскольку число вейвлет-коэффициентов убывает с ростом номера уровня детальности в геометрической прогрессии с показателем 2, то $L_\beta(N)$ убывает с той же скоростью и может так случится, что, начиная с некоторого уровня β все $L_\beta(N)$ будут равны нулю, т.е. множества $Q^{(\tau,\beta)}$ будут пустыми. Для того чтобы все последующие оценки, выполненные для коэффициентов уровня β в текущем временном окне, были статистически значимыми, введем *порог представительности* $L_{\min}$ – положительное целое число, смысл которого заключается в том, что оценки вычисляются лишь для тех уровней детальности β, для которых

$$L_\beta(N)\geq L_{\min} \tag{2}$$

Таким образом, длина окна N и порог представительности $L_{\min}$ вместе задают максимально возможный уровень детальности $\beta_{\max}$, вейвлет-коэффициенты которого могут быть включены в анализ.

Обозначим через $R^{(\tau,\beta)}$ матрицы размером $q\times q$, которые представляют собой выборочные оценки ковариационных матриц непустых множеств векторов $Q^{(\tau,\beta)}$:

$$\begin{aligned} R^{(\tau,\beta)} &= \tfrac{1}{L_\beta(N)}\sum_{z\,\in\,Q^{(\tau,\beta)}} z\cdot z^T = \| r_{jk}^{(\tau,\beta)} \|, \\ r_{jk}^{(\tau,\beta)} &= \tfrac{1}{L_\beta(N)}\sum_{r=1}^{L_\beta(N)} c_{jr}^{(\tau,\beta)} c_{kr}^{(\tau,\beta)}, \quad j,k=1,...,q \end{aligned} \tag{3}$$

где z – q-мерные вектор-столбцы вейвлет-коэффицентов уровня детальности β, попавших во временное окно τ. При вычислении (3) из векторов z не вычитается выборочное среднее, поскольку математическое ожидание вейвлет-коэффициентов равно нулю. Разделим компоненты векторов z на две части: скаляр z_1 и $(q-1)$-мерный вектор-столбец остальных компонент $\xi=(z_2,...,z_q)^T$. Далее вычислим канонической корреляцию 1-го временного ряда со всеми прочими. Для этого умножим скалярно каждый вектор ξ на некоторый неизвестный пока вектор φ и получим множество скалярных величин $\varsigma_1=\varphi^T\xi$. Найдем

вектор φ из условия, чтобы квадрат модуля коэффициента корреляции между множествами скалярных величин z_1 и ς_1 был максимален.

Эта задача представляет собой частный случай классической задачи Хотеллинга о канонических корреляциях. Вектор φ находится как собственный вектор, соответствующий максимальному собственному числу ν_1^2, $0 \le \nu_1^2 \le 1$, которое как раз и равно максимуму квадрата модуля коэффициента корреляции между значениями z_1 и ς_1, следующей матрицы размером $(q-1)\times(q-1)$ [*Hotelling, 1936; Rao, 1965*]:

$$S_{\xi\xi}^{-1} S_{\xi z_1} S_{z_1 z_1}^{-1} S_{z_1 \xi} \tag{4}$$

где $S_{z_1 z_1} = M\{z_1^2\}$, $S_{z_1 \xi} = S_{\xi z_1}^T = M\{z_1 \xi^T\}$, $S_{\xi\xi} = M\{\xi \cdot \xi^T\}$. Очевидно, что матрицы в формуле (4) являются подматрицами общей матрицы ковариаций размером $q \times q$ $S_{zz} = M\{z \cdot z^T\}$.

Заменив матрицу S_{zz} в (4) на ее выборочную оценку (3), можно реально вычислить вектор φ, собственное число $\nu_1^2(\tau,\beta)$ и множество из $L_\beta(N)$ скалярных значений $\varsigma_{1r}^{(\tau,\beta)}$, $r = 1,...,L_\beta(N)$, которые назовем значения *каноническими вейвлет-коэффициентами* временного ряда $Z_1(t)$ соответствующими уровню детальности β во временном окне τ. Что же касается величины $\nu_1^2(\tau,\beta)$, то ее естественно назвать *квадратом канонической корреляции* первой скалярной компоненты векторного временного ряда $Z(t)$ со всеми прочими компонентами на уровне детальности β во временном окне τ.

Проделав аналогичные операции со всеми прочими компонентами вектора $z \in Q^{(\tau,\beta)}$, получим таблицу размером $q \times L_\beta(N)$, состоящую из *канонических вейвлет-коэффициентов* для всех компонент векторного временного ряда для рассматриваемого уровня детальности β во временном окне τ:

$$\varsigma_{kr}^{(\tau,\beta)}, \quad k = 1,...,q; \quad r = 1,...,L_\beta(N); \quad \beta = 1,...,\beta_{\max} \tag{5}$$

и q значений квадратов канонических корреляций:

$$\nu_k^2 = \nu_k^2(\tau,\beta), \quad k = 1,...,q \tag{6}$$

Подчеркнем, что в результате этих операций произошел переход от таблицы (1) исходных вейвлет-коэффициентов к таблице (5) канонических коэффициентов. Таблицу (5) можно представить также в виде аналога множества $Q^{(\tau,\beta)}$ - как множество $\Theta^{(\tau,\beta)}$, состоящее из q-мерных векторов-столбцов таблицы (5).

В дальнейшем из величин (6) будем извлекать положительный квадратный корень. Поскольку с ростом номера уровня детальности число вейвлет-коэффициентов, принимающих участие в оценке величины $\nu_k(\tau,\beta)$ экспоненциально уменьшается, то, для уменьшения статистических флуктуаций оценки введем дополнительное усреднение по некоторому числу коэффициентов, полученных на предыдущих окнах:

$$\bar{\nu}_k(\tau,\beta)=\sum_{s=1}^{m_\beta}\nu_k(\tau-s+1,\beta)/m_\beta,\quad m_\beta=2^\beta \tag{7}$$

Чем выше уровень детальности, тем более глубоким является усреднение (7) по прошлым временным окнам, что значительно уменьшает зависимость амплитуды разброса статистических флуктуаций оценки (7) от номера уровня детальности и делает этот разброс примерно одинаковым для разных β. Согласно формуле (7) эффективная длина временного окна становится масштабно-зависимой и равной $N_e^{(\beta)}=N+2^\beta-1$.

Вейвлетную меру когерентности определим формулой:

$$\kappa(\tau,\beta)=\prod_{k=1}^{q}\bar{\nu}_k(\tau,\beta) \tag{8}$$

Значение меры (8) могут лежать в пределах от 0 до 1. Чем значение (8) больше, тем сильнее совокупная связь между всеми анализируемыми процессами на масштабах, соответствующих номеру β. Следует подчеркнуть, что величина (8) есть произведение q неотрицательных величин по модулю меньших 1. Поэтому чем больше число q анализируемых рядов, тем меньше абсолютные значения $\kappa(\tau,\beta)$. Вследствие этого сравнение абсолютных значений статистики (8) возможно лишь при одинаковом значении числа рядов q. Наибольший интерес представляют не абсолютные значения (8), а ее относительные величины для различных τ.

Для того, чтобы было возможно сравнивать вариации меры когерентности для разных уровней детальности одновременно условимся, что временной индекс τ в формуле (8) будет стартовать со значения $N_e^{(\beta_{\max})}=N+2^{\beta_{\max}}-1$ (само значение $\kappa(\tau,\beta)$ основано на информации о рядах строго для временных индексов $t:\tau-N_e^{(\beta)}\le t\le\tau$).

При рассмотрении конструкции агрегированных сигналов нам потребуется множество, состоящее из $L_\beta(N)$ скалярных величин $\chi_r^{(\tau,\beta)}$, $r=1,...,L_\beta(N)$, которые являются главными компонентами множества векторов $\Theta^{(\tau,\beta)}$. То есть, совершенно аналогично формуле (3), вычисляется выборочная оценка ковариационной матрицы канонических вейвлет-коэффициентов:

$$\begin{aligned}P^{(\tau,\beta)}&=\tfrac{1}{L_\beta(N)}\sum_{z\in\Theta^{(\tau,\beta)}}z\cdot z^T=\|\,p_{jk}^{(\tau,\beta)}\,\|,\\ p_{jk}^{(\tau,\beta)}&=\tfrac{1}{L_\beta(N)}\sum_{r=1}^{L_\beta(N)}\varsigma_{jr}^{(\tau,\beta)}\varsigma_{kr}^{(\tau,\beta)},\quad j,k=1,...,q\end{aligned} \tag{9}$$

находится собственный q-мерный вектор $(\psi_1^{(\tau,\beta)},...,\psi_q^{(\tau,\beta)})^T$ матрицы (9), соответствующий ее максимальному собственному числу и вычисляются значения:

$$\chi_r^{(\tau,\beta)}=\sum_{k=1}^{q}\psi_k^{(\tau,\beta)}\varsigma_{kr}^{(\tau,\beta)},\quad r=1,...,L_\beta(N);\quad \beta=1,...,\beta_{\max} \tag{10}$$

Значения (10) назовем *агрегированными вейвлет-коэффициентами*. Таким образом, агрегированные коэффициенты являются значениями главной компоненты канонических

коэффициентов (5). В дальнейшем, с помощью обратного вейвлет-преобразования от агрегированных коэффициентов будет строится так называемый «обычный» вейвлет-агрегированный сигнал в программе **AggW**.

Одним из основных свойств вейвлетов, которое делает их привлекательными для использования в задачах сжатия информации – это то, что они аккумулирует максимум информации в очень небольшом числе (в процентном отношении к общему количеству) вейвлет-коэффициентов [*Press et al., 1996, Mallat, 1998*]. Как следствие этого качества, множества вейвлет-коэффициентов характеризуются наличием больших выбросов, которые нельзя интерпретировать как результат ошибок измерений или сбоев систем регистрации. Наличие этих выбросов ставит ограничения на используемые методы анализа данных, основанные на переходе из временной области в область вейвлет-коэффициентов (что аналогично переходу в частотную область при классическом Фурье-анализе), поскольку множество коэффициентов представляет собой, таким образом, существенно негуассовскую выборку. Поэтому выводы, основанные на использовании классических регрессионных процедур, основанных на методе наименьших квадратов, а также классических методов многомерного анализа, основанных на обычных выборочных оценках ковариационных матриц и их собственных чисел и векторов, при условии их применения к множествам вейвлет-коэффициентов, должны приниматься с известной осторожностью и с осознанием того, что часть полезной информации может быть пропущена именно из-за неполного соответствия применяемых методов природе обрабатываемых данных.

В статистике проблема устойчивости полученных выводов к нарушению предположений о природе данных известна как проблема *робастности* статистических методов и впервые систематически изложена в классической монографии [*Huber, 1981*], хотя неустойчивость методов типа наименьших квадратов к наличию небольшого числа выбросов в данных известна давно и была осознана многими статистиками-практиками (в том числе и геофизиками, например, Г. Джеффрисом) много раньше работ Хьюбера – уже в конце 19-го и начале 20-го веков. В это же время были предложены методы, повышающие устойчивость выводов к наличию выбросов, основное свойство которых состоит в отказе от квадратичной меры оценки качества подгонки и переходу к другим мерам, растущим не так быстро, как квадрат, например, к модулю невязок. Платой за увеличение устойчивости результатов статистической обработки является существенное усложнение вычислительных процедур и увеличение времени счета.

Для того, чтобы построить робастные аналоги величин (5), (8) и (10) заметим следующее. Вернемся к рассмотрению множества $Q^{(\tau,\beta)}$ векторов исходных вейвлет-кэффициентов (1). Рассмотрим задачу регрессии $(q-1)$-мерного случайного вектора $\xi = (z_2,...,z_q)^T$ на скалярную случайную величину z_1, т.е. задачу оценки вектора u регрессионных коэффициентов в линейной формуле:

$$z_1 = \sum_{i=1}^{q-1} u_i z_{i+1} + \varepsilon_1 = u^T \xi + \varepsilon_1 \qquad (11)$$

где ε_1 - регрессионный остаток. Если вектор u определять методом наименьших квадратов:

$$\sum_{z \in Q^{(\tau,\beta)}} \left(\sum_{i=1}^{q-1} u_i z_{i+1} - z_1\right)^2 = \sum_{z \in Q^{(\tau,\beta)}} (u^T \xi - z_1)^2 \to \min_u \qquad (12)$$

то его оценка легко находится:

$$\hat{u} = S_{\xi\xi}^{-1} \cdot S_{\xi z_1} \tag{13}$$

Введем обозначение:

$$\hat{\xi}_1 = \hat{u}^T \xi \tag{14}$$

- оценка регрессионного вклада в формуле (11). Поскольку

$$\mathrm{cov}(z_1, \hat{\xi}_1) = \mathrm{cov}(z_1, S_{\xi\xi}^{-1} S_{\xi z_1} \xi) = S_{z_1\xi} S_{\xi\xi}^{-1} S_{\xi z_1} \tag{15}$$

то нетрудно показать, что квадрат модуля коэффициента корреляции между (14) и z_1 равен значению $S_{z_1\xi} S_{\xi\xi}^{-1} S_{\xi z_1} S_{z_1 z_1}^{-1}$, а это есть ничто иное, как максимальное собственное число матрицы (4) [*Rao, 1965*].

Таким образом, значения первой канонической вейвлет-компоненты могут быть определены как $\hat{\xi}_1 = \hat{u}^T \xi$ из решения регрессионной задачи (11)-(12). Аналогичное утверждение, очевидно, имеет место и для всех прочих канонических вейвлет-компонент. Этот факт открывает возможность определения канонических вейвлет-компонент как решения регрессионной задачи (12) в ее робастной модификации, т.е. вместо (12) решать задачу:

$$\sum_{z \in Q^{(\tau,\beta)}} |\sum_{i=1}^{q-1} u_i z_{i+1} - z_1| = \sum_{z \in Q^{(\tau,\beta)}} |u^T \xi - z_1| \to \min_u \tag{16}$$

Очевидно, что решение задачи (16) много сложнее, чем (12), не может быть выражено простой формулой типа (13) и нуждается в итерационной процедуре. В данной реализации задача (16) решалась следующим образом. Организовывалась последовательность итераций методом градиентного спуска, в котором в качестве градиента брался обобщенный градиент недифференцируемой функции (16) [*Clarke, 1975; Шор, 1979*] по компонентам искомого вектора u. Шаг вдоль обобщенного антиградиента вычислялся путем решения задачи одномерной минимизации методом золотого сечения. В качестве начального приближения для градиентной итерационной процедуры бралось решение метода наименьших квадратов по формуле (13), в которой ковариационные матрицы оценивались по обычным формулам типа (3), но перед оценкой матриц данные подвергались процедуре винзоризации. Следует подчеркнуть, что винзоризизация вейвлет-коэффициентов использовалась лишь для выборочных оценок ковариационных матриц для получения начального приближения – весь прочий анализ производился с исходными коэффициентами (после предварительных операций (i)-(v)).

Условие остановки градиентной процедуры заключалось либо в том, что общее число итераций (т.е. вычислений градиента) достигло 10000, либо в том, что шаг вдоль антиградиента, найденный в методе золотого сечения, становился слишком малым. После остановки градиентной процедуры считалось, что решение задачи (16) приближенно найдено и вычислялась каноническая вейвлет-компонента по формуле (14). Аналогичная процедура повторялась для всех компонент. В результате получается таблица *робастных канонических вейвлет-коэффициентов*, аналогичная (5), элементы которой пометим знаком «тильда», для различения с (5):

$$\tilde{\zeta}_{k\,r}^{(\tau,\beta)}, \quad k = 1,\ldots,q; \quad r = 1,\ldots,L_\beta(N); \quad \beta = 1,\ldots,\beta_{\max} \tag{17}$$

Как только что отмечалось, величина $\nu_1(\tau,\beta)$ являются обычным коэффициентом корреляции между регрессионным вкладом $\hat{\xi}_1$ задачи (12) и выделенной скалярной компонентой z_1. Воспользуемся этим фактом для получения робастного аналога $\nu_1(\tau,\beta)$ и используем не классическую формулу для вычисления выборочного значения коэффициента корреляции, а ее робастную модификацию [*Huber, 1981*]. Согласно ей коэффициент корреляции между выборками $x(r), y(r), r=1,...,n$ можно вычислить по формуле:

$$\rho(x,y)=\frac{S(\breve{z}^2)-S(\hat{z}^2)}{S(\breve{z}^2)+S(\hat{z}^2)} \tag{18}$$

где

$$\begin{aligned}&\hat{z}(r)=a\cdot x(r)+b\cdot y(r),\quad \breve{z}(r)=a\cdot x(r)-b\cdot y(r),\\&a=1/S(x),\quad b=1/S(y),\quad S(x)=med\,|x-med(x)|\end{aligned} \tag{19}$$

$med(x)$ означает медиану выборки x, а $S(x)$ означает абсолютное медианное отклонение выборки x.

Заменяя в формулах (18) и (19) $x(r)$ на $\tilde{\varsigma}_{kr}^{(\tau,\beta)}$, $y(r)$ на $c_{kr}^{(\tau,\beta)}$, а n на $L_\beta(N)$, получим робастное значение $\tilde{\nu}_k(\beta,\tau)$ коэффициента корреляции, описывающего силу связанности процесса с номером k со всеми прочими сигналами. Подчеркнем, что робастность метода проявляется в двух местах: при решении задачи минимизации (16), где используется метод наименьших модулей вместо наименьших квадратов, и при вычислении коэффициента корреляции по формуле (18).

Робастную вейвлетную меру когерентности, вычисляемую программой **WBRCM** определим формулой, аналогичной (8):

$$\tilde{\kappa}(\tau,\beta)=\prod_{k=1}^{q}|\,\bar{\tilde{\nu}}_k(\tau,\beta)\,| \tag{20}$$

где

$$\bar{\tilde{\nu}}_k(\tau,\beta)=\sum_{s=1}^{m_\beta}\tilde{\nu}_k(\tau-s+1,\beta)/m_\beta,\quad m_\beta=2^\beta \tag{21}$$

Для того, чтобы вычислить робастные агрегированные вейвлет-коэффициенты, аналогичные, следует вспомнить известный факт, относящийся к методу главных компонент [*Rao, 1965*]: собственный вектор, соответствующий максимальному собственному числу ковариационной матрицы (9), является решением задачи на условный максимум:

$$\sum_{r=1}^{L_\beta(N)}\sum_{k=1}^{q}(\varsigma_{kr}^{(\tau,\beta)}\psi_k^{(\tau,\beta)})^2\to\max_{\psi},\quad \sum_{k=1}^{q}(\psi_k^{(\tau,\beta)})^2=1 \tag{22}$$

Из формул (22) естественным образом вытекает робастная формулировка задачи нахождения главных компонент:

$$\sum_{r=1}^{L_\beta(N)}\sum_{k=1}^{q}|\,\varsigma_{kr}^{(\tau,\beta)}\psi_k^{(\tau,\beta)}\,|\to\max_{\psi},\quad \sum_{k=1}^{q}(\psi_k^{(\tau,\beta)})^2=1 \tag{23}$$

Решение задачи (23) находилось итерационной процедурой. В качестве начального приближения для вектора $\psi^{(\tau,\beta)}$ брался собственный вектор ковариационной матрицы робастных канонических вейвлет-коэффициентов (17), соответствующий максимальному собственному числу. При этом выборочная оценка ковариационных матриц также строилась лишь после винзоризации. Далее задача (23) решалась градиентным (вместо обычного градиента – обобщенный градиент) методом с проекцией градиентного шага на единичную сферу. Метод выбора шага вдоль градиента и условия остановки итераций использовались те же самые, что и при решении задачи (16).

После решения задачи (23), которое обозначим $\tilde{\psi}_k^{(\tau,\beta)}$, вычислялись *робастные агрегированные вейвлет-коэффициенты*:

$$\tilde{\chi}_r^{(\tau,\beta)} = \sum_{k=1}^{q} \tilde{\psi}_k^{(\tau,\beta)} \tilde{\varsigma}_{kr}^{(\tau,\beta)}, \quad r = 1,\ldots,L_\beta(N); \quad \beta = 1,\ldots,\beta_{\max} \tag{24}$$

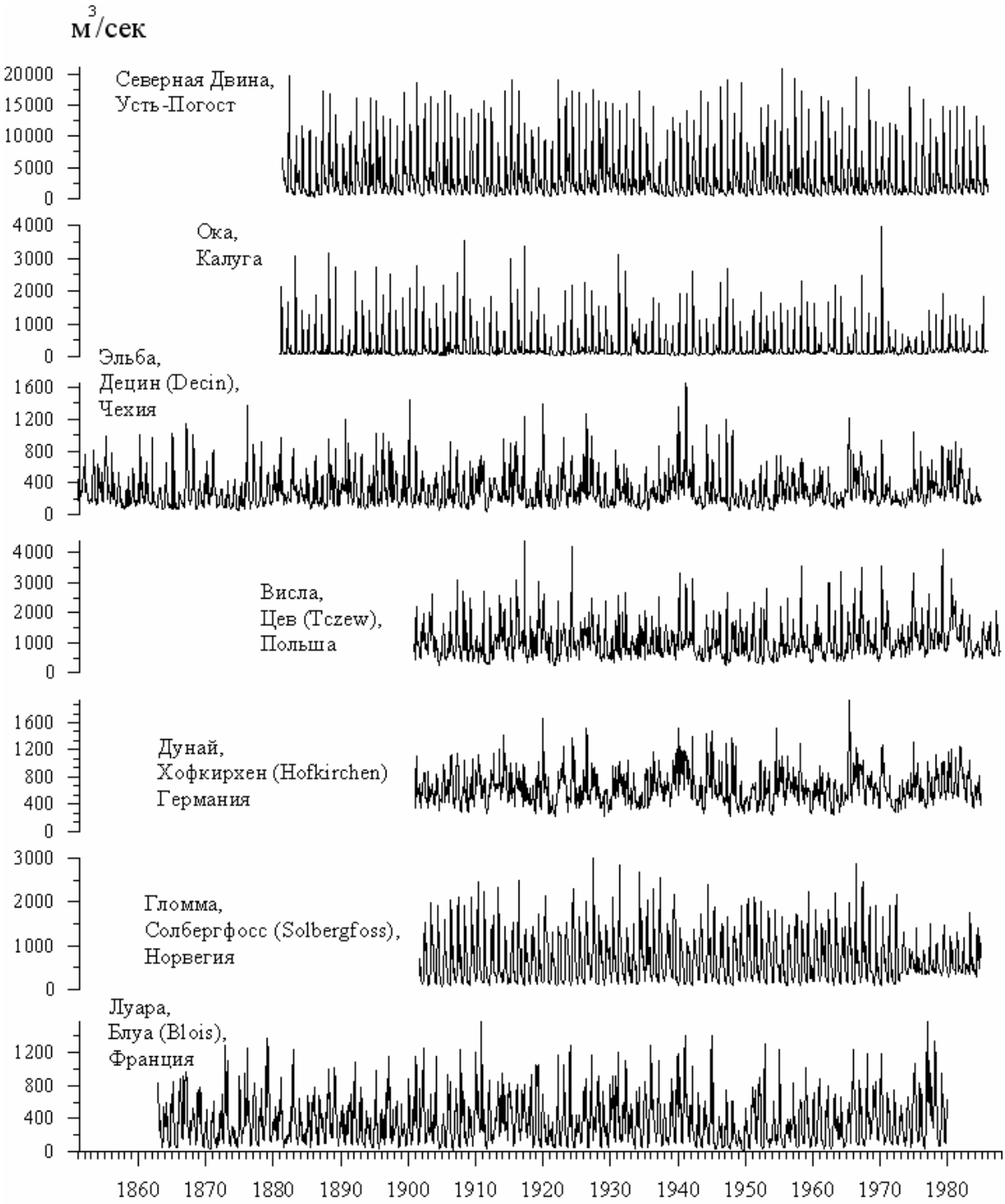
Обратное вейвлет-преобразованием от коэффициентов (24) будет определять робастный вейвлет-агрегированный сигнал в программе **AggWR**.

Нетрудно заметить, что метод построения вейвлетных мер когерентности содержит следующие три свободных параметра: N - длина временного окна; $L_{\min}$ - порог представительности для проверки условия (2) и, наконец, тип ортогонального финитного базиса, по которому раскладываются сигналы. Можно избавиться от необходимости выбора вейвлета. Ограничимся, например, 10 обычными вейвлетами Добеши и пусть $\gamma = 1,\ldots,m_\gamma$ - индекс, нумерующий число моментов, обнуляемых материнской базисной функций. Тогда $m_\gamma = 10$. Пусть $\kappa(\tau,\beta\,|\,\gamma)$ - обычная вейвлетная мера когерентности (8), а $\tilde{\kappa}(\tau,\beta\,|\,\gamma)$ - ее робастный аналог (20), построенные с помощью вейвлета, обнуляющего γ моментов. Усреднив эти меры по всем используемым базисам, получим много-базисные вейвлетные меры когерентности, вычисляемые программой **WBRCM_mult**:

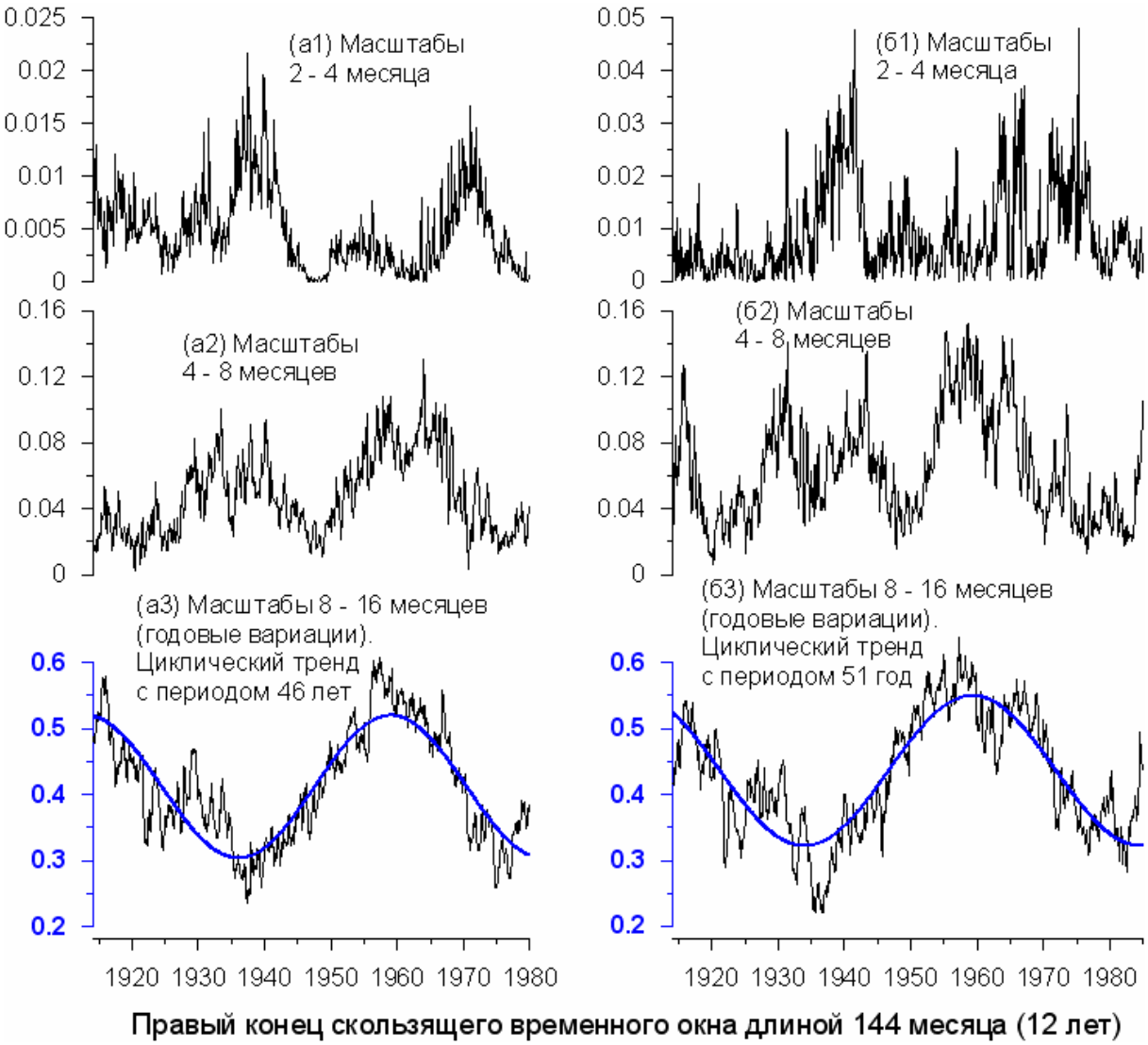
$$\mu(\tau,\beta) = \tfrac{1}{m_\gamma} \sum_{\gamma=1}^{m_\gamma} \kappa(\tau,\beta\,|\,\gamma), \quad \tilde{\mu}(\tau,\beta) = \tfrac{1}{m_\gamma} \sum_{\gamma=1}^{m_\gamma} \tilde{\kappa}(\tau,\beta\,|\,\gamma) \tag{25}$$

Примеры применения.

В качестве первого примера применения вейвлетных мер когерентности рассмотрим эволюцию робастной меры когерентности (20) для среднемесячных расходов воды 7 рек [*Любушин, 2007*]. Графики исходных временных рядов изображены на следующем рисунке. Ввиду сильно негауссова характера исследуемых сигналов, в частности, большого числа выбросов, представляется, что вейвлет-анализ в данном случае будет весьма уместным.



Оценка производилась в окне длиной 144 отсчета (12 лет). Такой выбор обусловлен максимальным округленным значением известного периода солнечной активности. Длина окна и выбранное значение порога представительности $L_{\min}$ =16 определяет номера уровней детальности, возможных для анализа – их три. Третий уровень детальности охватывает временные масштабы от 8 до 16 месяцев и, следовательно, включает в себя период годовых вариаций сток (12 месяцев). Как и следовало ожидать, уровень когерентности на 3-м уровне максимален, причем заметна циклическая модуляция когерентности (см. следующий рисунок, (a1)-(a3)).



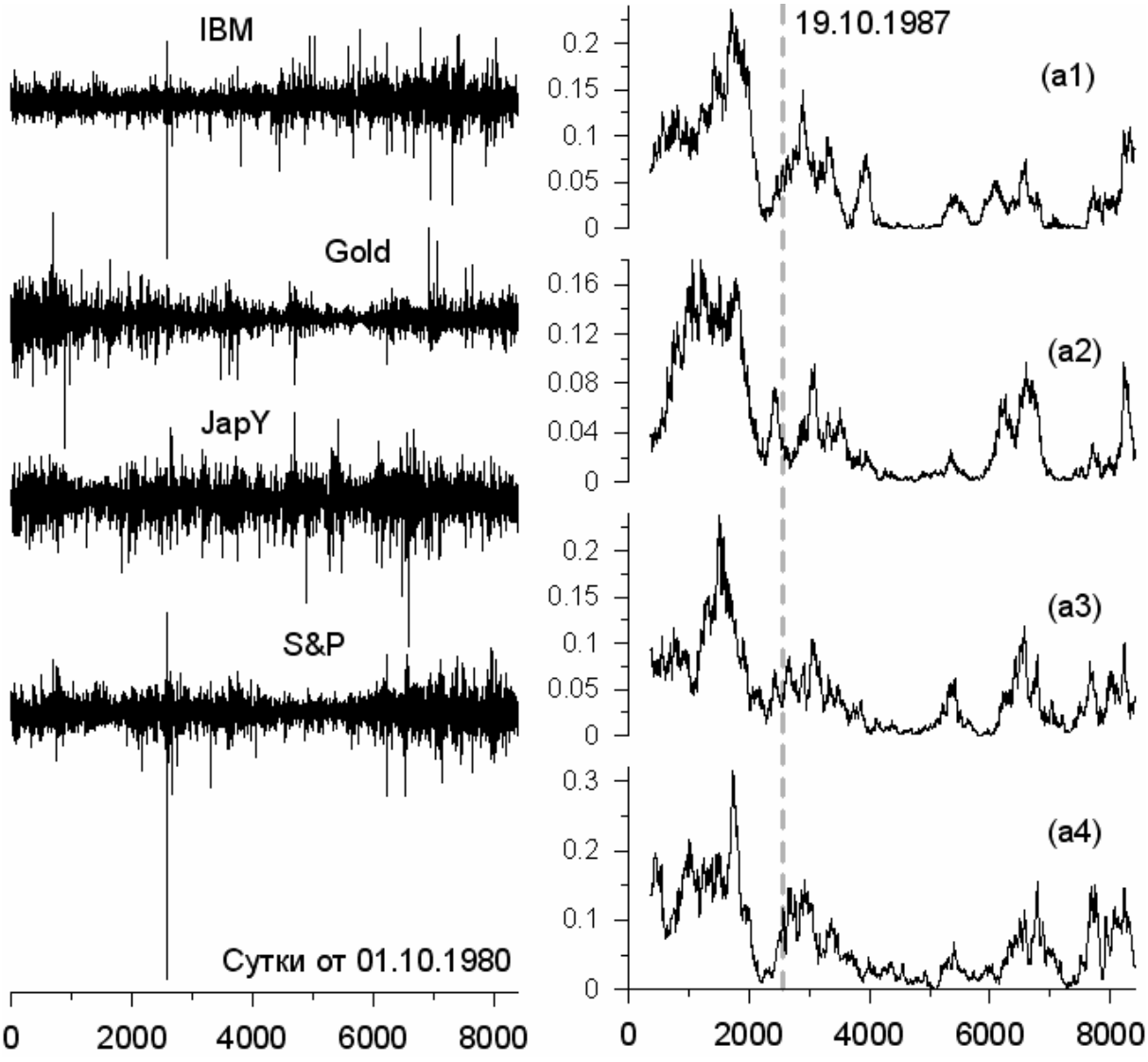
Период циклического тренда был оценен путем подгонки тренда с неизвестным периодом, который выбирался из условия минимума дисперсии остатка после вычитания тренда. Этот период оказался равным 552 месяцам или 46 годам. Учитывая малую длительность наблюдений, это значение можно считать вполне соответствующим независимо найденным климатическим периодичностям. Однако можно рассмотреть вопрос об устойчивости значения этого найденного периода по отношению к составу совместно анализируемых временных рядов речных стоков. С этой целью удалим 2 временных ряда – Гломму (поскольку там явно присутствует антропогенное изменение стока в самом конце ряда) и Луару (временной ряд для которой заканчивается раньше всех).

На рис.(б1)-(б3) представлены оценки меры когерентности для оставшихся 5 временных рядов. На рис.(б3) мы видим, что когерентность на 3-м уровне детальности сохранила циклическое поведение и ее период увеличился – он стал 609 месяцев или примерно 51 год. Таким образом, многомерный анализ позволяет выделить скрытые периодичности климатического происхождения в вариациях стока рек.

Второй пример применения вейвлетных мер когерентности относится к анализу финансовых временных рядов: стоимостей различных акций и сырья к концу дня торгов. На следующем рисунке слева представлены графики ежедневных приращений логарифма стоимости (т.н. log-returns) акций IBM, унции золота, японской йены и интегрированного индекса Standard & Poor`s (сверху вниз) на Нью-Йоркской бирже с 01.10.1980 по 30.09.2003.

Значения цен в выходные и праздники восполнены значениями в предыдущий день торгов. В течение этого промежутка времени произошло событие, которое традиционно является объектом интенсивных исследований в финансовой статистике, своего рода «финансовое землетрясение» или «черный понедельник» 19.10.1987, 2575-й день. Это событие отмечено резким скачком вниз на графике S&P.

Возникает вопрос, были ли какие-нибудь особенности в поведении сигналов перед этим катастрофическим изменением – это своего рода также задача поиска предвестников. Выбранные для анализа временные ряды характерны длительной историей, прочие многочисленные данные такого вида обычно начинаются с конца 1980-х или даже начала 1990-х годов.



Справа на рисунке на графиках (a1)- (a4) представлены эволюции много-базисной робастной меры когерентности $\tilde{\mu}(\tau,\beta)$ из (25) в зависимости от временной метки правого конца скользящего окна для первых 4-х уровней детальности, оцененной в скользящем временном окне длиной 365 суток, параметр $L_{\min}=16$. Таким образом, выбрана годовая длина окна. Как обычно, уровень 1 (a1) соответствует временным масштабам вариаций от 2 до 4 суток. Далее, сверху вниз: 2-й – от 4 до 8 суток, 3-й – от 8 до 16 суток и 4-й – от 16 до 32 суток. Видно, что это событие «перегрева рынка» предвещается почти за 1000 дней всплеском меры синхронизации для всех уровней детальности и последующим спадом к

началу события (типичная «бухтообразная аномалия»). Кроме того, видно, что участники торгов сделали выводы из этой катастрофы и последующая динамика рынка уже не выходит на столь сильное когерентное поведение.

Литература.

Любушин А.А. (2000) Вейвлет-агрегированный сигнал и синхронные всплески в задачах геофизического мониторинга и прогноза землетрясений. – Физика Земли, 2000, N3. С.20-30.

Любушин А.А. (2002) Робастный вейвлет-агрегированный сигнал для задач геофизического мониторинга – Физика Земли. 2002, N9. С.37-48.

Любушин А.А. (2007) «Анализ данных систем геофизического и экологического мониторинга». М.: Наука, 2007, 228с.

Шор Н.З. (1979) Методы минимизации недифференцируемых функций и их приложения. Киев, «Наукова думка», 1979, 200с.

Clarke E. (1975) Generalized gradients and applications //Trans. Amer. Math. Soc. 1975. Vol.205, No.2, pp. 247-262.

Hotelling H. (1936). Relations between two sets of variates. – Biometrika. Vol.28, pp.321-377.

Huber P.J. (1981) Robust statistics. John Wiley and Sons. New York, Chichester, Brisbane, Toronto (Русский перевод: Хьюбер П. (1984) Робастность в статистике. М., Мир, 303с.).

Mallat S. (1998) A wavelet tour of signal processing. Academic Press. San Diego, London, Boston, N.Y., Sydney, Tokyo, Toronto. 577 p. (Русский перевод: Малла С. Вэйвлеты в обработке сигналов. - М.: Мир, 2005, 671с.).

Press W.H., Flannery B.P., Teukolsky S.A. and Vettering W.T. (1996) Numerical Recipes, 2-nd edition, Chapter 13, Wavelet Transforms, Cambridge Univ. Press, Cambridge.

Rao C.R. (1965) Linear statistical inference and its applications. John Wiley & Sons, Inc. N.Y., London, Sydney (Русский перевод: *Рао С.Р.* (1968) Линейные статистические методы и их применение. М., Наука. 548 с.).