

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

**МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНОЛОГИЧЕСКИЙ
УНИВЕРСИТЕТ «СТАНКИН»**

**ИНСТИТУТ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ
РОССИЙСКОЙ АКАДЕМИИ НАУК**

***ФУНДАМЕНТАЛЬНЫЕ
ФИЗИКО-МАТЕМАТИЧЕСКИЕ
ПРОБЛЕМЫ
И
МОДЕЛИРОВАНИЕ
ТЕХНИКО-ТЕХНОЛОГИЧЕСКИХ
СИСТЕМ***

Ежегодный сборник научных трудов
Выпуск 12

*Материалы Международной научной конференции
«Моделирование нелинейных процессов и систем»*
Том 2

**Москва
Янус-К
2009**

Х. МОДЕЛИРОВАНИЕ ВРЕМЕННЫХ ПРОЦЕССОВ

УДК 330.322+330.4

КОМБИНИРОВАННЫЙ МЕТОД ВЫЯВЛЕНИЯ СКРЫТЫХ АНОМАЛИЙ В ХАОТИЧЕСКИХ ВРЕМЕННЫХ ПРОЦЕССАХ

В.В. Борог, А.В. Крянев, Д.К. Удумян

Московский инженерно-физический институт (государственный университет)

В настоящей работе предлагается обобщенная комбинированная схема выявления скрытых периодичностей и связанных с ними аномалий в нестационарных временных рядах. Схема основана на последовательном применении для выделения компонент исследуемых временных процессов методов сингулярно-спектрального анализа, вейвлет-анализа, Фурье-анализа.

1. Введение

Во многих физических задачах (например, в области радиофизики, астрофизики, космофизики и др.) требуется выделить «сигнал» на фоне подавляющего шума, форма и образ которого априорно не известен. Наблюдательная аппаратура может быть как одноканальной [1] (один временной ряд), так и многоканальной, когда одновременно регистрируется несколько временных рядов из смежных участков [2] исследуемой области (многомерные временные ряды). В последнее время, для мониторинга космической погоды планируется использовать наземные установки, регистрирующие поток космического излучения (мюонные телескопы и годоскопы [2]) по десяткам и тысячам направлений в широком диапазоне пространственных направлений.

Целью работы является разработка методов и схем выявления скрытых периодичностей и связанных с ними аномалий в нестационарных временных рядах, обремененных трендом и шумом, как правило, большой амплитуды и описывающих различные физические, биологические, экономические процессы.

Обычно, для выявления аномалий, по отдельности применяется один из хорошо известных математических методов: фурье-анализ [6], вейвлет-анализ [7] или сингулярно - спектральный анализ (SSA) [3-5]. Каждый из них имеет определенные недостатки, особенно для рядов небольшой

конечной длительности, изменчивости и не стационарности возникающих аномалий.

Ввиду того, что исследуемые ряды не стационарны (имеют тренд, вид которого априори неизвестен и меняется во времени), для выделения тренда был применен сингулярно - спектральный анализ (SSA) [3-5]. SSA позволяет также выявлять периодические сигналы из сильно зашумленных рядов (в которых амплитуда шума превосходит амплитуду искомого сигнала в несколько раз) и обладает преимуществом перед классическим спектральным анализом (методом Фурье). Но т.к. периодические процессы в исследуемых временных рядах сами имеют не стационарности (со временем у них может изменяться частота и амплитуда), то применение одного метода SSA для выделения искомого сигнала оказалось недостаточным. Для выявления нестационарных сигналов в работе представлены комбинированные схемы с применением SSA и вейвлет – анализа [9]. Для этого, например, с помощью SSA выделяется и удаляется тренд, затем фильтруется шум, и к отобранным компонентам применяется вейвлет - преобразование. Следует отметить работу [11], в которой комбинированная схема с применением SSA использовалась для выявления скрытых периодичностей в генетических последовательностях.

2. Математическая модель решения задачи выявления скрытых периодичности и аномалий в хаотических временных процессах

Рассмотрим временной процесс $f(t)$. Задача состоит в представлении процесса в виде

$$f(t) = f_T(t) + f_n(t) + f_\gamma(t) + \varepsilon(t), t \in [0, T], \quad (2.1)$$

где $f_T(t)$ - медленная регулярная составляющая, называемая трендом; $f_n(t)$ - периодическая или сумма периодических составляющих; $f_\gamma(t)$ - быстрые нерегулярные малые вариации, в которые обычно включают все, что не укладывается в периодические составляющие; $\varepsilon(t)$ - случайная составляющая, описываемая случайным процессом определенного типа.

Для функций вида $f(t) = f_T(t) + \varepsilon(t)$ хорошо работает теория аппроксимации (при малых $\varepsilon(t)$) или МНК (при больших $\varepsilon(t)$), для функций вида $f(t) = f_n(t)$ хорошо работает теория гармонических рядов Фурье. Однако для функций вида $f(t) = f_n(t) + \varepsilon(t)$ при больших амплитудах $\varepsilon(t)$ теория рядов Фурье для выявления полезного сигнала перестает работать. Для функций вида $f(t) = f_T(t) + f_n(t)$ при отсутствии

априорной информации о характере тренда выявление полезного сигнала с помощью теории рядов Фурье также перестает работать, а попытка представления тренда с помощью полиномов и дальнейшего его исключения может создать искусственные периодичности в ряду остатков или же уничтожить искомые компоненты.

Для выделения трендовых, периодических и шумовых компонент нами используется: SSA и комбинированные схемы с применением SSA и вейвлет - анализа. Для реализации комбинированной схемы с помощью SSA определяется и удаляется тренд, фильтруется шум, а к отобраным составляющим применяется вейвлет - анализ.

Рассмотрим временной ряд $\{x_i\}$, образованный последовательностью N равноотстоящих значений функции $f(t): x_i = f[i] = f(i\Delta t), i = 1, \dots, N$.

Выберем число $M < N$, называемое длиной длинной окна. Введем $K, K = N - M + 1$. Построим K M -мерных векторов $X_i = (x_i, \dots, x_{i+M-1})^T, 1 \leq i \leq K, X_i \in R^M$. Из этих векторов составим матрицу

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_K \\ x_2 & x_3 & x_4 & \dots & x_{K+1} \\ x_3 & x_4 & x_5 & \dots & x_{K+2} \\ \dots & \dots & \dots & \dots & \dots \\ x_M & x_{M+1} & x_{M+2} & \dots & x_K \end{pmatrix}.$$

Замечание: Ограничением на целочисленный параметр M является условие $M < N/2$. Выбор параметра M существенно зависит от решаемой задачи.

Пусть $S = XX^T \in R^{M \times M}$. Так как матрица S неотрицательная и симметричная, то ее собственные значения (с.з.) неотрицательны: $\lambda_1, \dots, \lambda_M \geq 0$. Расположим их в порядке не возрастания: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M \geq 0$. Обозначим через U_1, \dots, U_M ортонормированную систему собственных векторов матрицы S , соответствующих с.з. Назовем $d = \max\{i : \lambda_i > 0\}$ порядком сингулярного разложения.

Введем матрицы

$$V_i = \frac{1}{\sqrt{\lambda_i}} X^T U_i, i = 1, \dots, d.$$

Тогда сингулярное разложение матрицы X имеет вид

$$X = X_1 + X_2 + \dots + X_d, \quad X_i = \sqrt{\lambda_i} U_i V_i^T. \quad (2.2)$$

Вклад каждой компоненты выражается вкладом соответствующих с.з. в сумму всех собственных значений.

Существуют различные методы для автоматической идентификации следующих компонент:

1. Выделение тренда;
2. Выделение периодических компонент с определенными частотами.

Поскольку для нестационарных процессов периодические компоненты имеют изменяющиеся частоты и амплитуды, то была разработана схема совместного использования SSA и вейвлет - анализа для идентификации этих компонент. Только в результате применения этих схем удалось выявить периодические аномалии в различных нестационарных рядах при наличии шума большой амплитуды.

В результирующей мощности учитывались как периодичности с относительно постоянными характеристиками, так и периодичности с изменяющимися частотными и амплитудными характеристиками.

Перейдем теперь к выделению периодических компонент методом SSA и сравним результат с обычным Фурье - анализом, т.е. определим - при каком уровне зашумленности удается отделить сигнал от шума.

Выделения периодических сигналов методом SSA обладает рядом преимуществ. Во-первых, метод позволяет выделить трендовую составляющую, в то время как выделение тренда полиномами является некорректной операцией для нестационарных рядов и обработка ряда после такой процедуры может уничтожить некоторые периодические компоненты или выявить несуществующие. Во-вторых, при тех уровнях зашумленности, при которых не работают такие методы как Фурье-анализ и вейвлет - анализ, SSA позволяет выявлять периодические сигналы, что было сделано для вышеприведенного примера описанным выше способом. Более того, при наличии нескольких периодических компонент преимущество SSA становится более значимым.

Метод SSA эффективно и удобно применять в сочетании с вейвлет-анализом [9]. Это связано с тем, что сигнал может иметь изменяющуюся частоту. Тогда с помощью SSA можно убрать трендовую составляющую и составляющие, относящиеся к шуму (которые отвечают нескольким последним сингулярным числам), и после восстановления ряда по выделенным компонентам применять вейвлет-анализ.

3. Применение SSA совместно с вейвлет-анализом для выделения различных периодов солнечной активности

Основным индексом солнечной активности являются числа Вольфа. Для характеристики пятнообразовательной деятельности Солнца Вольф ввел специальный индекс $W = k(f + 10g)$, где f – количество всех отдельных пятен, замеченных на диске Солнца, а g – число образованных ими групп. Коэффициент k – учитывает качество наблюдений и позволяет различные наблюдения свести в общую систему. Солнечные пятна — это места выхода сильного магнитного поля. Основные периодичности солнечной активности известны и их можно выявить методами вейвлет-анализа [7]. Основной 11-летний цикл легко выявляется обычным Фурье-анализом. Он также легко виден в самом ряду среднегодовых чисел, полученных в период с 1700 по 2004 год.

Применение Фурье-анализа позволяет выявить 11-летний и 100-летние циклы.

Одиннадцатилетний пик для спектра мощности Фурье является сложным, имеет смысл огибающая этого пика. Длительность одиннадцатилетнего цикла зависит от времени.

Применение вейвлет - анализа позволяет выявить 11-летний, 50-летний и 100-летний периоды, а также проследить за эволюцией их основных составляющих (амплитуды и частоты).

Со временем 50-летний цикл ослабевает и почти исчезает (поэтому Фурье-анализ его не выявил).

Для выявления циклов применим теперь SSA. Стоит отметить, что метод Фурье при автоматической идентификации с $d_m = 0.65$ выявил лишь с.в. λ_6 и λ_7 это связано с тем, что периодограммы для остальных с.в. более размыты т.к. у них частота и амплитуда меняются.

Применяя к восстановленному ряду преобразование Фурье можем найти, что ряд имеет период $T = 76$. Вклад этого ряда в общий ряд можно

$$\frac{\lambda_6 + \lambda_7}{\sum_{\forall i} \lambda_i} = 0.036.$$

оценить по вкладу с.з.: Т.к вклад этой компоненты невелик, то её удалось выявить только методом SSA и не удалось выявить, ни преобразованием Фурье, ни вейвлет - анализом.

Далее, применяя вейвлет - анализ к ряду восстановленному по компонентам 2-20, что соответствует снятию тренда (в данном случае тренду соответствует постоянный уровень) и фильтрации шума, можно выявить 11-ти 50-ти и 100- летние периоды.

Изучая дневным данные по числам Вольфа, удалось с помощью SSA выявить наличие 27-дневного периода, который в вейвлет - преобразовании практически не виден (виден лишь в пиках 11-ти летнего периода). В то же время, с помощью SSA его удастся выявить в любых временных интервалах, разница лишь во вкладе во временной ряд. Т.о для каждого года этот период можно легко выявить и оценить его вклад.

Рассмотрим теперь применение описанных методов для исследования и выявления аномалий перед Форбуш-эффектом.

Для этого важно понять, как устроены исследуемые ряды. Исследование различных одномерных рядов длины $N = 180$ (3 часа) интенсивности потока мюонов в различных направлениях показывают, что трендовой составляющей отвечает первая компонента сингулярного разложения.

Явное преобладание первого с.з. над остальными свидетельствует, что трендовой составляющей соответствует ряд, восстановленный по первой компоненте.

$$F_1 = (f_1, \dots, f_N), F_1 = G(X_1), X_1 = \sqrt{\lambda_1} U_1 V_1^T \quad (3.1)$$

где G - оператор диагонального усреднения.

Для выбранного ряда тренд представлен на рис.1.

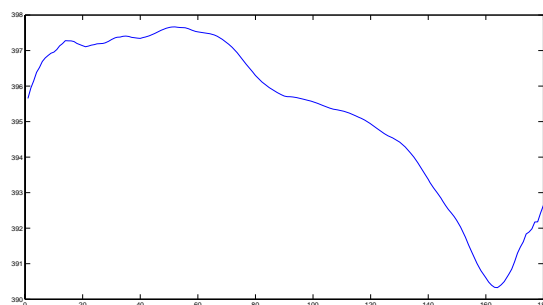


Рис.1. Ряд, восстановленный по первой компоненте

Отбросив несколько последних компонент, вклад которых пренебрежимо мал по сравнению с первыми, мы отфильтруем шум и, применяя вейвлет-анализ к ряду, восстановленному из отобранных компонент, можно оценить их энергетический вклад, подсчитав глобальный спектр энергии (скейлиграмму).

4. Схемы рекуррентного применения SSA и вейвлет – преобразования

На первом этапе с помощью SSA выделяем тренд. Для выделения и дальнейшего удаления тренда можно использовать разложение тренда по системе полиномов, ортогональных на дискретном наборе точек.

Определение степени полинома, который соответствует выделяемому тренду, можно воспользоваться графиком зависимости суммы квадратов отклонений исходного и сглаженного рядов от степени полиномов.

Для дальнейшего выделения всех периодических процессов из исследуемого временного ряда воспользуемся рекуррентным применением SSA.

Вначале выделим периодические компоненты методом SSA. Периодической компоненте могут соответствовать два соседних собственных вектора с близкими (в идеале одинаковыми) с.з. Периодограммы собственных векторов, отвечающих этим с.з., должны быть близкой формы (в идеале совпадать). Из теории следует, что они имеют следующий вид: $U_i = C \cos(2\pi f_0 n + \phi)$ и $U_{i+1} = C \sin(2\pi f_0 n + \phi)$ - собственные вектора, соответствующие периодической компоненте. Т.к $\|U_i\| = 1$ и $\|U_{i+1}\| = 1$, то

$$2 = C^2 \sum_{n=0}^{M-1} (\cos^2(2\pi f_0 n + \phi) + \sin^2(2\pi f_0 n + \phi)) = C^2 M, \quad C^2 = \frac{2}{M}.$$

и, следовательно,

Нормированная периодограмма собственных векторов имеет вид

$$\begin{aligned} d(f) = MD(f) &= M \frac{1}{4} \frac{2}{M} \frac{\sin^2(M\pi(f - f_0)\Delta t)}{M^2 \sin^2(\pi(f - f_0)\Delta t)} \\ &= \frac{1}{2} \frac{\sin^2(M\pi(f - f_0)\Delta t)}{M^2 \sin^2(\pi(f - f_0)\Delta t)}. \end{aligned} \quad (4.1)$$

Это выражение достигает максимального значения равного 0.5 при $f = f_0$.

Поэтому при $\max(d_i(f) + d_{i+1}(f)) > d_m$ можем утверждать, что пара собственных векторов U_i, U_{i+1} соответствуют периодическому сигналу, здесь $d_i(f), d_{i+1}(f)$ - нормированные периодограммы собственных векторов U_i, U_{i+1} , соответственно, d_m - пороговый уровень обнаружения сигнала.

Примененный таким рекуррентным образом метод SSA позволил выделить все периодические компоненты из зашумленного ряда.

Применив теперь вейвлет - преобразование, подсчитаем мощность выделенных таким образом процессов (в качестве базового вейвлета берем вейвлет Морле).

Подсчитываем глобальный спектр энергии:

$$E_w(a) = \int_{-\infty}^{+\infty} |W(a,b)|^2 db,$$

где $W(a,b) = \frac{1}{|a|^{0.5}} \int_{-\infty}^{+\infty} f(t) \psi^* \left(\frac{t-b}{a} \right) dt$ – интегральное вейвлет - преобразование с базовым вейвлетом Морле.

Рассмотрим теперь применение вейвлет - фильтрации с последующим применением SSA для выделения всех периодических процессов.

Если априори известно, что частоты периодических компонент в исходном ряду в достаточной степени низкие (на самом деле не очень высокие), то можно выявить эти компоненты следующим образом. В начале произведем вейвлет - фильтрацию для снятия высокочастотного шума

$$f(t) = \sum_{j,k \in \mathbb{Z}} d_{jk} \psi(2^j t - k)$$

дискретное вейвлет - преобразование функции $f(t)$. Оно может быть представлено в следующей форме

$$f(t) = \sum_k S_k^j \varphi(2^j - k) + \sum_{j \geq J} \sum_{k \in \mathbb{Z}} d_k^j \psi(2^j - k), \quad (4.2)$$

где $\varphi(t)$ - скаллограмма соответствующая выбранному вейвлету $\psi(t)$.

Первое слагаемое в этом выражении отвечает за регулярную составляющую, ограниченную уровнем J .

Коэффициенты $s_k^j \hat{=} d_k^j$ обычно определяются с помощью пирамидальной схемы быстрого вейвлет - преобразования. Вейвлет - фильтрация подразумевает отбрасывание части коэффициентов, с абсолютными значениями, меньшими заданного порогового уровня λ .

После вейвлет – фильтрации, методом SSA выявляются низкочастотные периодические компоненты. Далее с помощью вейвлета Морле считается суммарная мощность выделенных сигналов.

Для фильтрация шума методом SSA можно использовать статистику ω^2 , которая имеет следующий вид [10]

$$\varpi \omega_n^2 = \int_{-\infty}^{+\infty} [F_n(x) + F_n(-x) - 1]^2 dF_n(x), \quad (4.3)$$

где $F_n(x)$ - эмпирическая функция распределения рассматриваемой случайной величины, описывающей выделенный шум.

Статистику ω^2 можем подсчитать по следующей формуле

$$\omega_n^2 = \sum_{j=1}^n \left[F_n(-\tilde{x}_j) - \frac{n-j+1}{n} \right]^2, \quad (4.4)$$

где $\tilde{x}_1 \leq \tilde{x}_2 \leq \dots \leq \tilde{x}_n$ - наблюдаемые величины, упорядоченные в порядке возрастания.

Рассмотрим в качестве примера два ряда. Первый – изменение интенсивности потока нейтронов с нейтронного монитора в Москве, второй – с Иркутского нейтронного монитора. Данные с минутным интервалом приведены на рис.2.

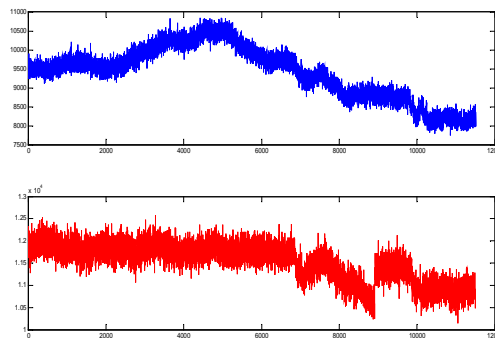


Рис.2

Применим описанные выше методы для выявления аномалий во время Форбуш - эффектов.

Для этого возьмем временное окно длиной N=240 минут и будем продвигать его с шагом в 10 минут.

Применение первой схемы дает результаты, представленные на рис. 3.

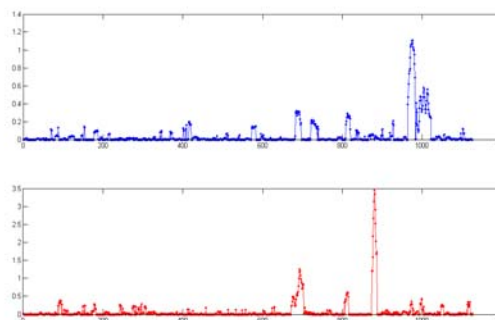


Рис.3

Результаты, полученные с помощью третьей схемы, представлены на рис. 4.

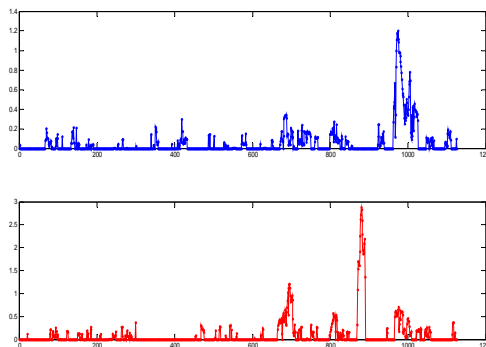


Рис.4

Видно, что обе схемы дают похожие результаты. Однако в каждом отдельно взятом из представленных двух рядов присутствуют процессы, не связанные с Форбуш – эффектом. Поэтому для выделения Форбуш – эффекта необходимо принимать во внимание только те процессы, которые одновременно наблюдаются на обеих установках и отражаются в рассматриваемых рядах.

С учетом комбинации представленных выше схем получаем результат по выделению Форбуш - эффекта, показанный на рис.5.

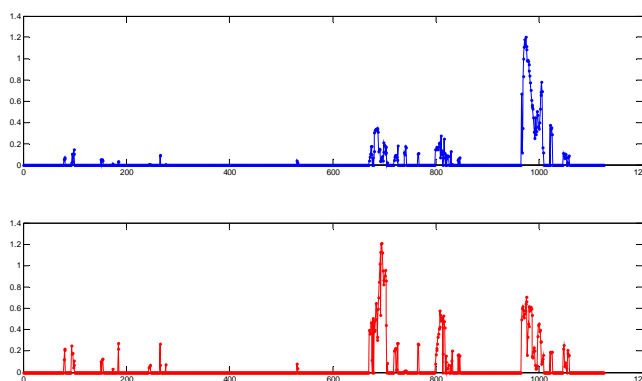


Рис.5

Таким образом, удалось выявить космический Форбуш – эффект, отделив его от эффектов изменения интенсивностей потока нейтронов, имеющих локальный атмосферный характер.

5. Заключение

Основной целью работы является выявление скрытых периодичностей и связанных с ними аномалий в нестационарных временных рядах. Представленные в работе схемы позволяют выявлять аномальные процессы, такие, например, как Форбуш-эффекты (резкое изменение потока космического излучения).

Для выделения трендовых, периодических и шумовых компонент нами используется: SSA; вейвлет - анализ; спектральный анализ, а также комбинированный метод с применением SSA и вейвлет - анализа. Для

реализации комбинированного метода с помощью SSA определяли и удаляли тренд, фильтровали шум, а к отобраным составляющим применяли вейвлет - анализ.

Изучая закономерности временного ряда чисел Вольфа, удалось с помощью SSA выявить 27-дневный период, который при применении отдельно вейвлет – преобразования и Фурье-анализа не выявляется.

Комбинированный метод совместного использования SSA и вейвлет - анализа дает наилучший результат, что дает возможность прогнозировать аномальные явления, например, применение комбинированного метода позволило прогнозировать за несколько суток Форбуш - эффекты, связанные с резким возрастанием вариации солнечного коронарного излучения.

Литература

1. Барбашина Н.С., Борог В.В. и др. Широкоапертурный мюонный детектор для диагностики атмосферы и магнитосферы Земли. //Изв.РАН, сер. физ., т.71, №7, 2007.
2. Кононович Э.В. Солнечно - Земная Физика. М.:СИЗИФ, 2001.
3. Голяндина Н.Э. Метод Гусеница - SSA: анализ временных рядов. Санкт-Петербургский государственный университет, 2004.
4. Данилов Д.Л., Жиглявский А.А. Главные компоненты временных рядов: метод Гусеница. Санкт-Петербургский государственный университет, 1997.
5. Крянев А.В., Лукин Г.В. Математические методы обработки неопределенных данных. М.: Физматлит, 2006.
6. Витязев В.В. Спектрально - корреляционный анализ равномерных временных рядов. Санкт-Петербургский государственный университет, 2001.
7. Витязев В.В. Вейвлет - анализ временных рядов. Санкт-Петербургский государственный университет, 2001.
8. Отнес Р., Эноксон Л. Прикладной анализ временных рядов. Основные методы. М.: Мир, 1982.
9. Борог В.В., Крянев А.В., Удумян Д.К. Комбинированный метод выявления скрытых аномалий в одномерных и многомерных хаотических временных процессах// Препринт № 001-2008 М.: МИФИ, 2008.
10. Зрелов П.В., Иванов В.В. Интегральные непараметрические статистики и их свойства. Алгебраический вид, функции распределения и критерии согласия// Сообщения ОИЯИ, P10-92-461, Дубна, 1992.
11. E.V. Korotkov, M.A. Korotkova, N.A. Kudryashov. Information decomposition method to analysis symbolical sequences// Physics Letters A 312, 2003.

Из рисунка 3 видно, что наряду с трендовой компонентой исследуемый макроэкономический показатель имеет резко выраженные сезонные колебания, которые с помощью разработанных нами схем выделяются и прогнозируются с высокой точностью. Ошибка месячного прогноза на 2 года вперед для экономического показателя, представленного на рис.3 составляет менее 1%. Как показали многократные применения разработанных нами методов, эти методы позволяют в большинстве случаев прогнозировать исследуемые временные процессы с приемлемой точностью.

Выводы

В работе новые робастные схемы выделения трендовых, аномальных и хаотических компонент и прогнозирования применяются для анализа и прогнозирования динамики процессов. Используемые методы обладают вычислительной устойчивостью и устойчивостью к большим по амплитуде кратковременным выбросам в исследуемом временном процессе. С их помощью выделяются аномальные компоненты, в том числе не просматривающиеся явно, как по их положению во времени, так и по амплитуде. Это предоставляет возможность для более глубокого изучения аномальных и кризисных явлений исследуемых процессов. Разработанные нами схемы прогнозирования позволяют учесть дополнительную априорную экспертную информацию и, тем самым, получить более надежные прогнозные значения для исследуемых временных процессов.

Литература

1. Haykin S. Adaptive Filter Theory. Prentice-Hall, Englewood Cliffs. 1991.
2. Крянев А.В., Лукин Г.В. Математические методы обработки неопределенных данных. М.: Физматлит, 2006.
3. Себер Дж. Линейный регрессионный анализ. М.: Мир, 1980.
4. Хьюбер П. Робастность в статистике. М.: Мир, 1984.
5. Arsenin V.Ya., Kryanev A.V. Generalized maximum likelihood method and its application for solving ill-posed problems // Ill-Posed Problems in Natural Sciences. A. Tikhonov (Ed.). Netherland, VSP-BV. 1992.
6. Antoniou I., Akritas P., Burak D.A., Ivanov V.V., Kryanev A.V., Lukin G.V. Robust methods for stock market data analysis // Physica A 336, 2004.
7. Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of Time Series Structure. SSA and Related Techniques. Champan & Hall/CRS, 2001.