

УДК 519.2+551.510.534

ПРИМЕНЕНИЕ СИНГУЛЯРНОГО СПЕКТРАЛЬНОГО АНАЛИЗА ДЛЯ ВЫДЕЛЕНИЯ СЛАБО ВЫРАЖЕННЫХ ТРЕНДОВ

В.Б. Кашкин, Т.В. Рублева

Сибирский федеральный университет, г. Красноярск

E-mail: kafedra@nifti.krasnoyarsk.ru

Рассмотрены особенности нового метода анализа временных рядов – сингулярного спектрального анализа, обсуждаются достоинства метода и проблемы, связанные с его применением. Для повышения точности выделения тренда и устранения краевых эффектов предложено использовать прогноз ряда в оба конца. Сингулярный спектральный анализ применен к временному ряду спутниковых данных общего содержания озона в кольце широт от 40 до 60° в Северном и Южном полушарии. Найдены тренды. Показано, что за 1998–2005 гг. в этом кольце в Северном полушарии происходит уменьшение общего содержания озона на 0,38 %±0,01 % в год, а в Южном – на 0,10 %±0,01 % в год. Это не удивительно – большая часть озоноразрушающих веществ производится в средних широтах Северного полушария.

При изучении временных рядов обычно используется спектральный анализ Фурье или вейвлет-анализ. В настоящей работе применен аппарат сингулярного спектрального анализа (ССА «Гусеница»), разработанный и обоснованный в конце XX в. сотрудниками Санкт-Петербургского государственного университета [1, 2]. Метод основан на анализе главных компонент и позволяет исследовать стационарные и нестационарные временные ряды. Доказан ряд полезных теорем, предложены вычислительные методы для реализации ССА, создано программное обеспечение. Ранее связь между классическими методами анализа стационарных временных рядов и методом главных компонент была рассмотрена в [3].

ССА пока не нашел широкого применения, в частности, в технических приложениях. Ряд примеров использования ССА «Гусеница» приведен в [4]. В [5, 6] этот метод использован для анализа и прогноза данных о солнечной активности и данных о неравномерности вращения Земли.

Метод «Гусеница» основан на переходе от одномерного временного ряда длиной n с равномерным шагом $(x_1, x_2, x_3, \dots, x_n)$ к многомерным рядам, построенным из исходного одномерного. «Гусеница» позволяет выделить интересующие компоненты временных рядов, в частности, линейный и нелинейный тренд, компоненты с заранее известным и заранее не известным периодом, сгладить исходные данные, сделать прогноз ряда, заполнить пропуски.

Из исходного ряда составляется матрица X

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_k & \dots & x_{m+1} \\ x_2 & x_3 & x_4 & \dots & x_{k+1} & \dots & x_{m+2} \\ x_3 & x_4 & x_5 & \dots & x_{k+2} & \dots & x_{m+3} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_k & x_{k+1} & x_{k+2} & \dots & x_{2k-1} & \dots & x_n \end{pmatrix}.$$

Здесь $m < n$ – длина гусеницы, последней строкой с номером $k = n + m - 1$ являются элементы $(x_k, x_{k+1}, \dots, x_n)$, причем $x_{ij} = x_{i+j-1}$. Эту матрицу можно рассматривать как m -мерную выборку объема k или m -мерный временной ряд. В некоторых прило-

жениях могут быть вычислены средние значения μ_i по столбцам, тогда матрицу X можно центрировать. Далее вычисляется матрица $R = (1/k)X \cdot X^T$. Если матрица X центрирована, то R является выборочной корреляционной матрицей:

$$R = \begin{pmatrix} \sigma_1^2 & R_{12} & \dots & R_{1m} \\ R_{21} & \sigma_2^2 & \dots & R_{2m} \\ \dots & \dots & \dots & \dots \\ R_{m1} & R_{m2} & \dots & \sigma_m^2 \end{pmatrix}.$$

Элементами матрицы R служат выражения

$$R_{ij} = \frac{1}{k} \sum_{i=1}^k (x_{i+j-1} - \mu_i) \cdot (x_{j+i-1} - \mu_j).$$

По обычной методике анализа главных компонент [3] вычисляются собственные числа и собственные векторы матрицы R , т. е. проводится её сингулярное разложение $\sum_{i=1}^m \lambda_i = P \cdot \Lambda$ – диагональная матрица упорядоченных по убыванию собственных чисел матрицы R :

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \dots 0 \\ 0 & \lambda_2 & 0 \dots 0 \\ \dots & \dots & \dots \\ 0 & 0 & 0 \dots \lambda_m \end{pmatrix},$$

$\sum_{i=1}^m \lambda_i = P$, P – ортогональная матрица собственных векторов матрицы R :

$$P = \begin{pmatrix} P_{11} & P_{21} & \dots & P_{m1} \\ P_{12} & P_{22} & \dots & P_{m2} \\ \dots & \dots & \dots & \dots \\ P_{1m} & P_{2m} & \dots & P_{mm} \end{pmatrix}.$$

Матрица P есть решение уравнения $[R - \lambda I]P = 0$, I – единичная матрица. Матрица P удовлетворяет условию $P^{-1} = P^T$, что означает сохранение «средней мощности» ряда.

В методе «Гусеница» столбцы матрицы P играют роль переходных функций фильтров, настроенных на составляющие исходного процесса. Таким образом, фильтры порождаются самим исследуемым

процессом, «Гусеница» сама настраивается на те спектральные компоненты, которые присутствуют в этом процессе. Выделение спектральных компонентов ряда осуществляется линейным преобразованием исходного процесса с помощью дискретного оператора свертки:

$$y_j[l] = \sum_{q=1}^m x_{lq} P_{jq} = \sum_{q=1}^m x_{l+q-1} P_{jq}.$$

Выбор нескольких главных компонентов подобен параллельному соединению нескольких фильтров. Ширина полосы пропускания фильтра определяется видом собственного вектора и длиной интервала усреднения (длиной «гусеницы» m). Ширина полосы пропускания обратно пропорциональна m . Максимальное значение m равно половине длины ряда n , в этом случае при четном n матрица X квадратная. При небольших m , вплоть до $m=2$, происходит сглаживание ряда. Вид собственных векторов и главных компонент, полученных в результате преобразования матрицы R , дает информацию о структуре изучаемого процесса и свойствах его слагаемых. В частности, среди главных компонент можно наблюдать и интерактивно выбирать те, что относятся к тренду (медленно меняющиеся), периодические, шумовые.

Представляет интерес сравнение метода «Гусеница» с анализом Фурье. Если ряд состоит из набора строго гармонических компонентов, то «Гусеница» осуществляет разложение в ряд Фурье с выделением именно этих компонентов. «Гусеница» позволяет выделять спектральные составляющие с учетом изменения амплитуды и фазы квазипериодического сигнала, что невозможно при использовании обычного ряда Фурье. «Гусеница» выделяет также пары квазипериодических составляющих главных компонент ряда, которые имеют длину $m \leq n/2$. Эти компоненты сдвинуты по фазе на $\pi/2$, т. е. ортогональны. Арктангенс отношения этих компонентов дает зависимость полной фазы Φ квазипериодической составляющей от времени, правда, длина этого ряда короче длины исходного ряда n и равна $m \leq n/2$. Хотя фаза Φ определена с точностью до 2π , устранить скачки фазы величиной в 2π во временном ряду полной фазы не представляет труда. Иногда полная фаза может быть аппроксимирована линейной функцией: $\Phi(t) = A \cdot t + B$, $i=1 \dots m$. В этом случае коэффициент $A = \omega$ – это частота, т. е. производная от фазы. Возможна более сложная аппроксимация фазы, например, $\Phi(t) = A \cdot t + C \cdot t^2 + B$. Производная от фазы дает тренд частоты: $\omega(t) = A + C \cdot t$.

При использовании ССА бывает необходимо улучшить спектральное разрешение, которое зависит от длины ряда n . Авторы данной статьи увеличивали n , продлевая ряд в оба конца с использованием прогноза, сделанного с помощью той же «Гусеницы». Можно осуществлять «обратную децимацию», т. е. удлинять ряд путем вставки промежуточных интерполированных значений между исходными значениями с использованием свойства «Гусеницы» восстанавливать пропущенные данные. Это сложные

процедуры, в то время как дискретное преобразование Фурье позволяет улучшать спектральное разрешение простым добавлением нулей [7]. Отметим также, что по сравнению с быстрым дискретным преобразованием Фурье «Гусеница» – принципиально медленный метод, не позволяющий, в современной интерпретации, проводить обработку временно-го ряда в реальном времени, так как обычно требуется работать с матрицами большой размерности.

По-видимому, точное выделение слабо выраженного тренда на фоне шума и периодического сигнала большой интенсивности представляется одной из сложных задач анализа временных рядов. Нами был проведен вычислительный эксперимент с рядом длиной 3000 шагов (см. левый верхний угол на рис. 1), где тренд – линейный, а периодический сигнал содержит компонент синусоидальной формы. На рис. 1 цифрой 1 показан исходный тренд, результатом выделения тренда полиномом первой степени по методу наименьших квадратов (МНК) (2) и результат выделения тренда с помощью «Гусеницы» (3). Значительно худшие результаты показало использование окна Ханна и вейвлета МНАТ. Краевые эффекты при использовании «Гусеницы» показаны стрелками. Хотя точность выделения тренда «Гусеницей» существенно лучше, чем по методу МНК, её необходимо было повысить, устранив краевые эффекты.

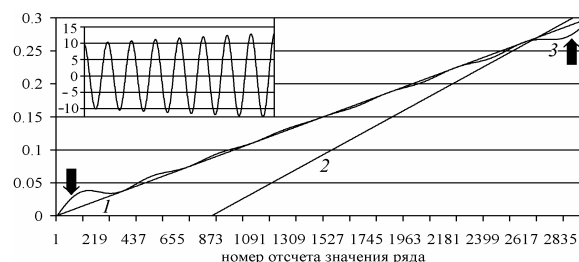


Рис. 1. Выделение тренда (1) по МНК (2) и с помощью «Гусеницы» (3)

Для этого был использован упомянутый выше прием – в данном случае продление ряда на 1000 шагов вперед с использованием прогноза «Гусеницей». Прогноз позволил существенно уменьшить краевой эффект – сдвинуть область неточной оценки тренда вправо, за пределы исследуемого ряда длиной 3000 шагов (рис. 2). На рис. 2 выброс из-за влияния краевого эффекта показан стрелкой.

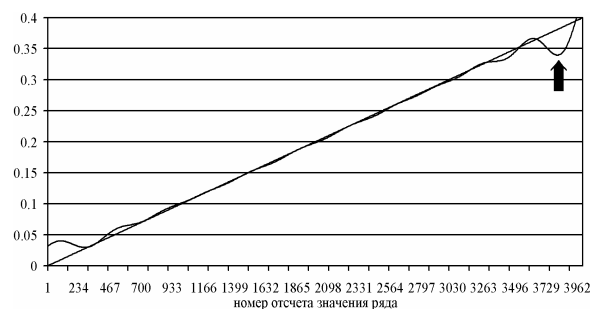


Рис. 2. Перемещение области неточной оценки тренда (показана стрелкой) вправо, за пределы исходного ряда длиной 3000 шагов

В качестве примера рассмотрим обработку временных рядов общего содержания озона (ОСО) в атмосфере Земли. Озоновый слой выполняет важную экологическую задачу – защищает все живое на Земле от губительного ультрафиолетового излучения Солнца. Отмечается постепенное глобальное истощение озонового слоя в течение десятилетий [8]. Выделяя тренды временных рядов ОСО, можно определить, с какой скоростью происходит истощение.

Определение ОСО, т. е. толщины слоя озона, приведенного к температуре 273 К и нормальному атмосферному давлению, осуществляется с помощью наземных и спутниковых методов. Искусственные спутники Земли оценивают ОСО, регистрируя рассеянное «назад» солнечное излучение в ультрафиолетовой области спектра. В нашей работе использованы ежесуточные глобальные данные об ОСО (ряды длиной $n=2920$ сут.), полученные космическим аппаратом EP/TOMS (США) за период с начала 1998 г. по конец 2005 г. и помещенные на сайте NASA [9].

Исследуется ОСО в области циркумполярных вихрей на широтах в кольце $40...60^\circ$ в Северном и Южном полушарии. Циркумполярный вихрь, т. е. обширный циклон, где ОСО всегда достаточно велико, в Южном полушарии окружает Антарктическую озоновую дыру. Выбор интервала широт обусловлен тем, что на широтах выше происходят значительные изменения ОСО, превышающие изменения в тропиках и субтропиках. На широтах более 60° эти изменения ещё значительнее, но непрерывные ряды ОСО по спутниковым данным для этих широт отсутствуют, так как спутниковая аппаратура оптического диапазона не может оценивать ОСО в зимние месяцы на неосвещенных Солнцем приполярных участках атмосферы.

На рис. 3 приведена зависимость от времени среднего суммарного ОСО в кольце широт от 40 до 60° для Северного и Южного полушария в единицах Добсона (е.Д.).

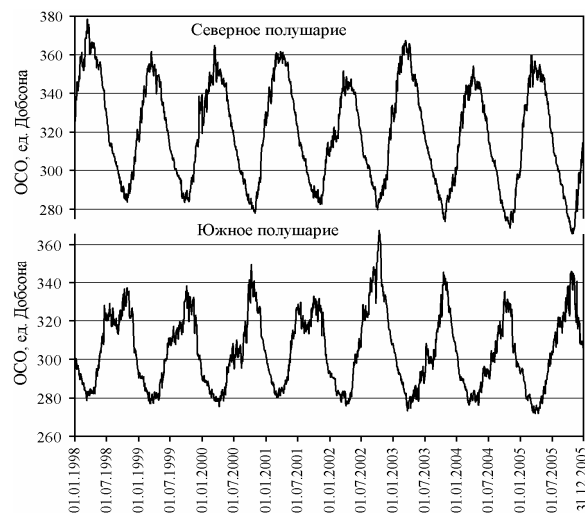


Рис. 3. Графики временных рядов среднего суммарного ОСО в кольце широт от 40 до 60° для Северного и для Южного полушария

В тот период, когда в Северном полушарии значение ОСО максимально, в Южном наблюдается минимум. Максимумы ОСО приурочены к весне, минимумы – к осени каждого полушария. Отметим, также, что в период максимума среднего суммарного ОСО в Южном полушарии Антарктическая озоновая дыра имеет наибольшую площадь. Каждый из представленных на рис. 3 временных рядов содержит тренд, периодические и случайные компоненты.

Линейные тренды рядов, представленных на рис. 3, первоначально были найдены по МНК в виде $x_i = a \cdot t_i + b, i = 1...n$. Параметр a несет важную информацию и характеризует скорость изменения среднего суммарного ОСО в кольце. Для Северного полушария оценка $a = -0,00726$ е.Д./сутки, стандартное отклонение оценки $\sigma_a = 0,00056$ е.Д./сут. Знак минус означает уменьшение ОСО. В расчете на год уменьшение ОСО составляет 2,6 е.Д. или 0,83 % по отношению к среднему значению.

Для Южного полушария МНК-оценка $a = -0,00044$ е.Д./сут., стандартное отклонение оценки $\sigma_a = 0,00042$ е.Д./сут., т. е. оценка a и её стандартное отклонение практически равны, следовательно, оценку a нельзя считать статистически значимой.

Далее анализ рядов ОСО был проведен методом ССА «Гусеница» с использованием соответствующего программного обеспечения [4]. Для повышения спектрального разрешения и устранения краевых эффектов ряды были продолжены в оба конца на 1000 дней с использованием прогноза. С помощью моделирования найдено, что при этом погрешность за счет краевых эффектов и шума при оценивании a составила $\Delta a \leq 0,0001$ е.Д./сут. Применение сингулярного спектрального анализа позволило существенно увеличить достоверность оценивания тренда ОСО, скорости деградации озонового слоя и не требуется заранее задавать вид тренда. Полученные тренды среднего суммарного ОСО для Северного и Южного полушария в кольце широт от 40 до 60° приведены на рис. 4.

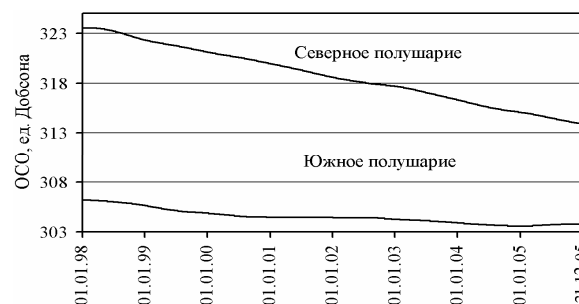


Рис. 4. Тренд среднего суммарного ОСО для Северного и Южного полушария в кольце широт от 40 до 60°

Из рис. 4 следует, что тренд ОСО в Северном полушарии хорошо описывается линейной зависимостью; тренд по Южному полушарию, в некотором приближении, также может быть аппроксимирован прямой. Для Северного полушария оценка параметра $a = -0,00334$ е.Д./сут., стандартное от-

клонение оценки $\sigma_a = 2,3 \cdot 10^{-6}$ е.Д./сут., для Южного оценка $a = -0,000811$ е.Д./сут., стандартное отклонение оценки $\sigma_a = 5,2 \cdot 10^{-6}$ е.Д./сут. В обоих случаях $\sigma_a \ll \Delta a$, поэтому погрешность оценивания a следует считать равной Δa .

Таким образом, в Северном полушарии общее содержание озона за 1998–2005 гг. уменьшилось в среднем на 1,222 е.Д./год, т. е. на 0,38 % \pm 0,01 % по

отношению к среднему значению. В Южном полушарии существует тенденция к уменьшению ОСО на 0,296 е.Д./год, т. е. на 0,10 % \pm 0,01 % по отношению к среднему значению. Применение сингулярного спектрального анализа позволило существенно повысить точность выделения тренда.

Работа выполнена при поддержке РФФИ, грант № 07-01-00326

СПИСОК ЛИТЕРАТУРЫ

1. Главные компоненты временных рядов: метод «Гусеница» / Под ред. Д.Л. Данилова и А.А. Жиглявского. – СПб.: Преском, 1997. – 308 с.
2. Голяндина Н.Э. Метод «Гусеница» SSA: анализ временных рядов. – СПб.: СПб ун-т, 2004. – 74 с.
3. Бриллинджер Л. Временные ряды. Обработка данных и теория. – М.: Мир, 1980. – 536 с.
4. <http://www.gistatgroup.com>
5. Loskutov A., Istomin I.A., Kuzanyan K.M., Kotlyarov O.L. Testing and Forecasting the Time Series of the Solar Activity by Singular Spectrum Analysis // Nonlinear Phenomena in Complex Systems. – 2001. – V. 4. – № 1. – P. 47–51.
6. Кашкин В.Б., Баскова А.А. Исследование неравномерности вращения Земли с помощью сингулярного спектрального анализа // Вестник Красноярского государственного университета. Физ.-мат. науки. – 2006. – № 7. – С. 53–60.
7. Марпл-мл. С.Л. Цифровой спектральный анализ и его приложения. – М.: Мир, 1990. – 586 с.
8. Александров А.Л., Израэль Ю.А., Кароль И.Л., Хргиан А.Х. Озонный щит Земли и его изменения. – СПб.: Гидрометиздат, 1992. – 287 с.
9. <http://jwocky.toms.gsfc.nasa.gov>

Поступила 05.07.2007 г.

УДК 004.93'12

РАСПОЗНАВАНИЕ СТРУКТУРИРОВАННЫХ СИМВОЛОВ НА ОСНОВАНИИ МЕТОДОВ МОРФОЛОГИЧЕСКОГО АНАЛИЗА

А.В. Афонасенко

Томский государственный университет
E-mail: anuta8181@mail.ru

Рассматривается технология распознавания структурированных символов на основании методов морфологического анализа. Разработанный метод позволяет повысить надежность распознавания в условиях изменения масштаба, ориентации и проективных преобразований символов.

Распознавание изображений структурированных (печатных) символов обеспечивает решение ряда научных и прикладных задач при идентификации объектов различной природы. Современные методы распознавания символов используются для решения широкого круга задач, офисных (электронная подпись, расшифровка сообщений, распознавание текста и др.), так и специализированных задач, распознавание изображений маркировки на поверхностях различных объектов и др. При регистрации изображений наибольшие искажения, влияющие на результат распознавания, вносят аффинные и проективные искажения. Они существенно снижают надежность распознавания методами, используемыми в современных системах распознавания печатных символов (например, FineReader, Readiris, ScanSoft OmniPage, CuneiForm и др.). На сегодняшний момент выделяют три основных подхода для решения задачи распознавания символов: структурный, признаковый и шаблон-

ный [1–6]. Каждому из этих методов присущи свои достоинства и недостатки.

Шаблонные методы [1] сравнивают изображение символа со всеми имеющимися в базе системы шаблонами. Наиболее подходящим шаблоном считается тот, у которого будет наименьшее количество точек, отличных от исследуемого изображения. Шаблонные методы хорошо распознают дефектные символы (разорванные, склеенные), но основной недостаток шаблонных методов – невозможность распознать шрифт, хоть немного отличающийся от заложенного в систему (размером, наклоном или начертанием).

Признаковые методы [2–4] наиболее распространены. В их основу положено упрощающее предположение, что можно анализировать не все изображение символа, а только набор признаков, вычисленных по изображению. Подразумевается, что значения признаков несут достаточно информации о символе. Однако слабым местом призна-