УДК 519.246.87

Н.В. Абалов, В.В. Губарев, О.К. Альсова

ИСПОЛЬЗОВАНИЕ МЕТОДОВ СИНГУЛЯРНОГО СПЕКТРАЛЬНОГО АНАЛИЗА И МОДЕЛЕТЕКИ ПРИ ИДЕНТИФИКАЦИИ ВРЕМЕННЫХ РЯДОВ

Абалов Н.В., Губарев В.В., Альсова О.К. **Использование методов сингулярного спектрального анализа и моделетеки при идентификации временных рядов.**

Аннотация. Сингулярный спектральный анализ (ССА) является относительно новым методом анализа нестационарных временных рядов. Слабой стороной ССА является отсутствие аналитического модельного представления ряда, например, в виде суммы простых функций, компактное аналитическое представление которых могло бы быть нагляднее и доступнее для интерпретации, чем совокупность большого количества компонент. В настоящей работе описан оригинальный метод вариативного моделирования, позволяющий устранить отмеченную слабую сторону ССА путем совместного использования его и метода моделетеки для получения компактного и легко интерпретируемого модельного представления изучаемого временного ряда с желаемым уровнем его адекватности ряду, цели и условиям идентификации.

Первый этап предлагаемого метода заключается в разложении исходного временного ряда на компоненты с помощью ССА. Разложение исходного ряда завершается выделением интересующих исследователя компонент. На втором этапе компоненты идентифицируются моделями из априори сформированной моделетеки согласно целям идентификации. Результатом является результирующая модель исходного временного ряда в аддитивной или аддитивно-мультипликативной форме.

Применимость метода рассматривается на примерах идентификации искусственного ряда и реальных ежедневного данных изменения мутности воды в реке в г. Челябинске за 2005 г.

Ключевые слова: спектральный анализ, сингулярный спектральный анализ, моделетека, временные ряды, нестационарный временной ряд, вариативное моделирование, модель.

Abalov N.V., Gubarev V.V., Alsova O.K. **Use of methods of singular spectral analysis and modeleteka for the identification of time series.**

Abstract. Singular spectrum analysis (SSA) is relatively new method for analysis of non-stationary time series. The weakness of SSA is lack of analytical model representation of time series, e.g., as a sum of simple functions, which could be clearer and easier for interpretation than a large number of components in form of time series.

In this paper we propose to use variative modeling, based on joint use of SSA and method of modeleteka, for obtaining of analytical model of time series, providing necessary level of adequacy, compactness and interpretability. First, time series are decomposed into components using SSA, significant components are selected using formal indicators (e.g. variance contributed by component, etc.). Second, each significant component is identified according to the purpose of identification with simple and interpretable model from preformed modeleteka. The result is final model of time series in additive or additive-multiplicative form.

Applicability of the method is shown on synthetic data and time series of daily changes of water turbidity in the river in the city of Chelyabinsk in 2005.

Keywords: spectral analysis, singular spectral analysis, modeleteka, time series, non-stationary time series, variative modeling, model, identification.

1. Введение. Под идентификацией временных рядов (ВР) понимается построение их моделей по эмпирическим значениям ряда. Чаще всего при этом стремятся, чтобы модель задавалась аналитически и

удовлетворяла определенным требованиям. Среди них, например, такие: 1) модель должна быть как можно более компактной, т.е. содержать как можно меньше структурных составляющих, параметров; 2) она должна позволять прикладным специалистам осуществлять интерпретацию модели, её составляющих и параметров; 3) модель должна быть четко ориентирована на конкретную задачу, ради решения которой она создавалась, либо допускать многофункциональное (многозадачное) её применение. Это, например, задачи: сжатия данных (ВР), имитации, прогнозирования, анализа и синтеза объекта, порождающего ВР, и т.д.

Однако разработанный и применяемый на данный момент инструментарий для определения структуры и модельного представления ВР зачастую не позволяет достичь желаемого качества идентификации одновременно с простотой и интерпретируемостью полученных результатов. Часто это вызвано тем, что ВР имеют сложную априори неизвестную структуру и характеризуются наличием нестационарности и компонент различной формы. Во многих приложениях ВР состоит из тренда, затухающих колебаний или биений, а также включает набор периодических колебаний, соответствующих сезонным составляющим (см., например, [1]).

Наиболее распространённым при идентификации таких ВР является двухэтапный подход [2, 3]. На первом этапе изучается структура ВР с использованием периодограммного спектрального анализа (ПСА). На втором – строится модельное описание ВР с использованием полигармонических моделей. Основными недостатками такого подхода являются: необходимость предварительной обработки ВР для устранения нестационарности, поскольку периодограммный СА требует предположения о стационарности ВР; фиксированность базиса разложения, состоящего лишь из гармонических колебаний.

Относительно новым методом анализа временных рядов является сингулярный спектральный анализ (ССА), описание которого и его объединение с методом «Гусеница» можно найти, например, в [4]. Среди основных сильных сторон этого метода можно отметить то, что он: не требует предположения о стационарности процесса; применим к коротким зашумленным рядам; позволяет выделять как сложные нестационарные компоненты, в частности аддитивный нелинейный тренд, медленно затухающие на интервале наблюдения колебания, так и периодические компоненты, а также шум.

Слабой стороной ССА является отсутствие аналитического модельного представления ряда.

Цель работы: описание разработанного метода и структуры программного обеспечения позволяющих осуществлять многофункциональную идентификацию ВР с достаточной для прикладных задач адекватностью на интервале наблюдения ВР в следующих условиях: а) априорная неопределенность о структуре компонент модели ВР; б) допустимость гипотезы о пригодности для описания ВР модели нестационарного случайного процесса в виде аддитивного или мультипликативного тренда и суммы стационарных (в широком смысле) периодических компонент и апериодических компонент, в частности шума; в) возможность априори неизвестного компактного описания выделенных компонент модели.

2. Описание предлагаемого метода. Построение модели ВР непосредственно на основе исходного временного ряда в условиях априорной неопределенности о её структуре, а именно её состава, формы отдельных компонент и их количества сложен и зачастую основывается на гипотетических предположениях о структуре временного ряда. Один вариант решения задачи – выбор сложной модели и подгонка её к ВР. Второй – представить модель аддитивной, мультипликативной или аддитивно-мультипликативной совокупностью простейших моделей, выбирать и подгонять множество простых моделей к каждой компоненте исходного ряда, а затем, при необходимости, переводить их в более сложное компактное аналитическое представление.

Суть предлагаемого метода сводится к двухэтапной идентификации, реализующей второй вариант. На первом этапе производится разложение ряда с использованием метода ССА на множество различных по свойствам компонент, с выделением, по процедурам, описанным в [4], из них наиболее значимых (по формальным показателям, например, по доле их вклада в общую дисперсию ВР, либо по их практической значимости или интерпретируемости). ССА является аналогом применяемого для исследования ВР многомерного метода главных компонент (МГК). Базовый вариант метода состоит в преобразовании одномерного ряда с использованием сдвиговой оконной процедуры («Гусеница») в представляемый в матричной форме многомерный ряд, разложении его на компоненты, число которых определяется размером окна, выделения в них наиболее значимых компонент и восстановления (аппроксимации) ряда, очищенного от нежелательных компонент.

На втором этапе каждая из отобранных компонент исследуется на пригодность включения её в модели как аддитивной или мультипликативной составляющей и идентифицируется как можно более про-

стой и хорошо интерпретируемой, согласно цели идентификации, моделью, выбираемой из предварительно сформированной моделетеки. Под моделетекой понимается упорядоченное множество моделей [3, 5]. В результате получается итоговая модель ряда в аддитивной, мультипликативной или аддитивно-мультипликативной форме.

Это позволяет использовать моделетеку из небольшого набора простых проверенных моделей для выбора наиболее адекватной (по формальным или функциональным показателям) модели к каждой из компонент по отдельности. Заметим, что использование на первом этапе ССА позволяет сразу выделить те компоненты, которые интересуют исследователя, а применение моделетеки позволяет автоматизировать апостериорный выбор модели этих компонент. Это следует из самого принципа формирования моделетеки как упорядоченного множества моделей, удовлетворяющих требованиям простоты, полноты, минимальной избыточности, уровня описания и исследованности в приложении к конкретной предметной области [3, 5].

3. Структура системы, реализующей подход. Предлагаемый метод может быть реализован системой, состоящей из нескольких крупных модулей, представленной на рисунке 1.

Первый этап предлагаемого метода заключается в разложении исходного ВР, формируемого в модуле 1, на компоненты с помощью ССА.

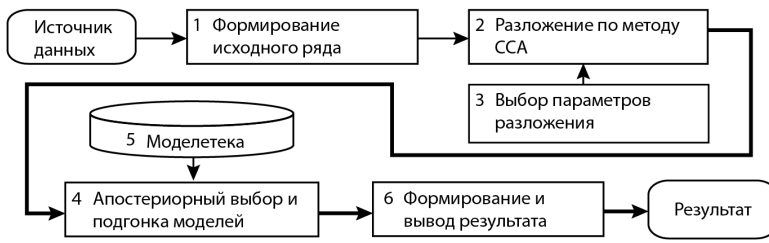

Рис. 1. Структура системы

Разложение исходного ряда завершается выделением интересующих исследователя компонент, которые могут быть проинтерпретированы как тренд, сезонные и циклические составляющие, а также шумы. За выполнение данного этапа отвечает модуль 2, который учитывает параметры разложения, полученные от специального модуля 3 (в простейшем случае параметр один - длина окна). Полученные в ходе разложения компоненты ВР, в целях уменьшения расчетов, могут быть сгруппированы (в частности парами, так как гармонические ко-

лебания раскладываются на две компоненты) автоматически или вручную.

На втором этапе все компоненты, полученные в результате разложения, передаются в модуль 4 для поиска и подгонки моделей, выбираемых из моделетеки. В стартовом варианте реализации системы в состав моделетеки были включены следующие простейшие модели: линейная, кусочно-линейная, полигармонические, вейвлет (мексиканская шляпа), сумма вейвлета и гармонического колебания, рациональная функция, сумма двух гауссовых кривых. В простейшем варианте выбор модели может осуществляться перебором всех возможных моделей из множества, когда отбирается та из них, которая обеспечивает наилучшую в принятом смысле адекватность в результате аппроксимации ею эмпирических данных. При большем количестве перебираемых моделей можно использовать известные [3, 5] приемы автоматизации упорядочивания и выбора моделей. При отборе моделей учитывается не только качество (значение формальной меры адекватности модели ВР), но и сложность модели. Если значения показателей качества простой и сложной модели совпадают или близки, то приоритет следует отдавать более простой, и, как правило, легче интерпретируемой модели. С целью упрощения полученного результата в модуле 6 решаются вопросы повышения компактности полученного модельного представления, например, путем выявления общего мультипликативного тренда нескольких компонент. Итоговый результат окончательно формируется и выдается пользователю в модуле 6.

Рассмотрим результаты, получаемые с применением упрощенной реализации предложенного метода согласно рисунке 1 в программной среде MATLAB.

4. Примеры применения предлагаемого метода. Для проверки работоспособности и качества предлагаемого метода применим используемый для метрологической аттестации средств измерительной техники метод образцового сигнала. Для этого на вход системы подадим искусственный временной ряд, характеристики которого известны и близки реальным временным рядам. Например таким, с которыми приходится работать при анализе инфекционной обстановки и состояния окружающей среды [1, 2], содержащими нестационарные, периодические компоненты и шум.

Искусственный ряд $x(t)$ содержит 365 наблюдений (шаг дискретизации времени – 1 день, $t = 0,...,365$) и состоит из 5 детерминированных компонент, имеющих периодический характер:

$$x(t) = 4 + 4\sin(0{,}5 + 2\pi t / 1365) + 0{,}91\sin(1 + 2\pi t / 7) - 0{,}72\sin(1 + 2\pi t / 11) \cdot$$
$$\cdot \sin(2\pi t / 172) - 1{,}1\sin(1 + 2\pi t / 31) - 0{,}8\sin(0{,}065 + 2\pi t / 365{,}2) + \varepsilon(t).$$

Случайная компонента $\varepsilon(t)$ имеет распределение, близкое к нормальному с нулевым средним и среднеквадратическим отклонением, составляющим 10% от среднеквадратического отклонения суммы детерминированных компонент. Колебания с периодом в 1365 дней рассматриваются как монотонный тренд, так как длина окна наблюдения значительно меньше периода данной компоненты.

Согласно процедурам ССА [4] из сингулярного разложения ряда были выделены первые 11 компонент по степени убывания значимости, дающих долю объясненной дисперсии, равную примерно 0,99.

Полученные 11 компонент были сгруппированы на основе визуального анализа графиков компонент (на основе процедур описанных в [4]). В результате было выделено 4 отдельных компонент, которые были переданы в модуль подгонки моделей. При ручном выборе моделей было решено использовать следующие модели: 1 компонента (тренд, медленные колебания) – гармоническое колебание с аддитивным линейным трендом, 2, 3 компоненту – гармоническое колебание (синус). Поскольку 4-я компонента представляет собой биения, для ее описания была выбрана более сложная модель – «сумма двух синусоид» (что с учетом тригонометрических свойств эквивалентно произведению синусоид). В примере использовались достаточно простые модели.

Как видно из графиков, приведенных на рисунке 2, модели приемлемо точно подогнаны к исходным данным. Коэффициент детерминации R^2 – доля дисперсии, объясненная суммарной моделью, – составила 0,99. Мы получили адекватную и компактную модель $y(t)$, где $t = 0,...,365$, аналитически описывающую исходный временной ряд и имеющую вид:

$$y(t) = -1{,}64\sin(0{,}013t + 1{,}25) - 0{,}0022t + 7{,}66 + 1{,}08\sin(0{,}20t + 3{,}95) +$$
$$+ 0{,}9\sin(0{,}9t + 0{,}12) + 0{,}35\sin(0{,}53t - 1{,}16) + 0{,}35\sin(0{,}61t + 2).$$

В качестве второго примера рассмотрим реальные данные, представляющие собой значения мутности воды в реке в г. Челябинске за 2005 год. График временного ряда приведен на рисунке 3. Исходные данные были взяты из банка данных CliWaDIn (Climate, Water, Diseases, Infections) [6]. Данный ряд был выбран как пример явно нестационарного ряда с разладкой, к которому плохо применимы классические методы.

В силу априорной неопределенности о структуре модели ВР проиллюстрируем, что даст применение для идентификации этого ВР классического периодограммного анализа и предполагаемого метода.

Нетрудно убедиться, что нестационарность ряда затрудняет получение адекватных результатов с использованием классического периодограммного анализа.

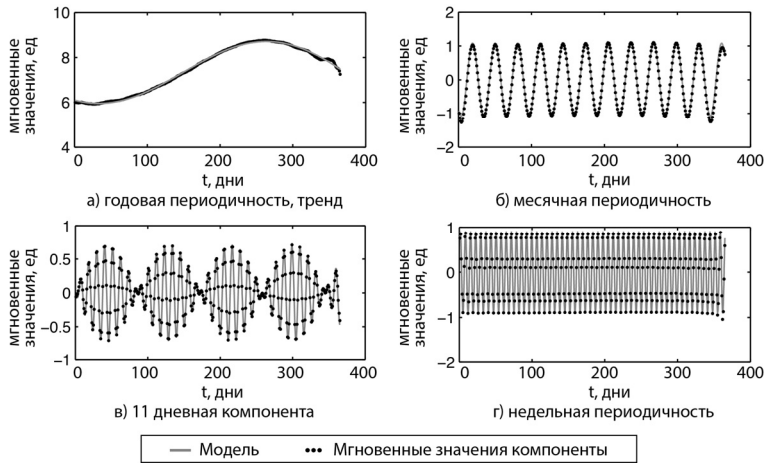

Рис. 2. Графики компонент и подогнанных моделей

Анализ периодограммы не позволяет сделать адекватные предположения о структуре ряда, поскольку медленные трендовые составляющие ряда и прочие нестационарности «размазывают» значения периодограммы в районе низких частот. Для повышения «структурной информативности» периодограммы предварительно вычтем трендовую составляющую. Воспользуемся простейшим экспоненциальным сглаживанием для его выделения. Из графика (рисунок 3) видно, что в данном случае вместе с трендом также были частично удалены колебания, что нежелательно с точки зрения возможной потери значимых периодических компонент идентификационной модели ВР.

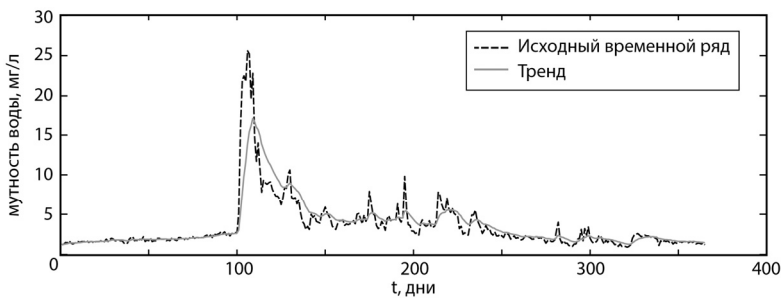

Рис. 3. Мутность воды в реке в городе Челябинске за 2005 год

Квазистационарная периодограмма, полученная для ряда за вычетом тренда, более наглядна и интерпретируема по сравнению с периодограммой, полученной для исходного ряда. Тем не менее, и в ней сохраняются значения трендовых компонент. Анализ такой периодограммы позволяет предложить модель в виде суммы трендовой и полигармонической составляющей, состоящей из 6 компонент (пар синусов и косинусов одной частоты).

$$y(t) = s(t, 0{,}15) - 0{,}51\sin(2\pi t \cdot 0{,}047) + 0{,}49\cos(2\pi t \cdot 0{,}047) +$$
$$+ 0{,}032\sin(2\pi t \cdot 0{,}044) - 0{,}67\cos(2\pi t \cdot 0{,}044) - 0{,}52\sin(2\pi t \cdot 0{,}036) -$$
$$- 0{,}39\cos(2\pi t \cdot 0{,}036) - 0{,}38\sin(2\pi t \cdot 0{,}016) - 0{,}38\cos(2\pi t \cdot 0{,}016) -$$
$$- 0{,}48\sin(2\pi t \cdot 0{,}027) + 0{,}22\cos(2\pi t \cdot 0{,}027) + 0{,}165\sin(2\pi t \cdot 0{,}011) +$$
$$+ 0{,}47\cos(2\pi t \cdot 0{,}011),$$

где $s(t,\alpha) = s(t-1) + \alpha\left[x(t) - s(t-1)\right]$.

Доля объясненной дисперсии суммарной модели составила 0,81. При этом часть ряда была описана моделью экспоненциального сглаживания, которая является менее информативной с точки зрения предварительного анализа, чем, например, гармонические или линейные модели, позволяющие оценить параметры процесса: периодичность, темп роста и т.п. Более того, трендовая составляющая, описанная экспоненциальным сглаживанием, дает долю объяснённой дисперсии исходного ряда равную 0,72, т.е. значительная часть структуры ряда объясняется слабо информативной моделью. Увеличение числа компонент с 6 до 8 дает увеличение коэффициента детерминации суммарной модели лишь на 0,016.

Наличие размазанности по частотам в периодограмме свидетельствует о том, что некоторые компоненты имеют непериодический характер и, возможно, могут быть описаны более компактно при использовании базиса, отличного от гармонического. Это наталкивает на использование моделетеки, в рамках которой следует совмещать полигармонические модели с другими видами моделей, например, линейными и вейвлетами. Как будет видно из дальнейшего, некоторые компоненты изучаемого ряда имеют вейвлет-подобную форму. Описание таких компонент только гармониками повышает сложность модели и понижает ее интерпретируемость.

Стоит отметить, что даже при использовании лишь гармонического базиса анализ структуры ряда в предлагаемом подходе упрощается по сравнению с классическим, так как рассматривается не весь ряд в целом, а его отдельные компоненты. Кроме того упрощается решение вопросов соответствия модели требуемому уровню адекватно-

сти, поскольку исследователь будет в итого иметь знания о вкладе каждой из выделенных компонент в мощность всего ВР.

Теперь используем предлагаемый метод. С помощью ССА из ряда были выделены 15 компонент. Часть компонент была сгруппирована в результате визуального анализа. После чего отобранные и сгруппированные компоненты были аппроксимированы моделями, выбранными из моделетеки.

Как уже упоминалось, метод позволяет широко изменять используемый базис. Для одной компоненты в принципе можно подобрать не одну модель. Это позволит исследователю рассмотреть одни и те же данные с разных сторон, варьируя модели в зависимости от текущих целей исследования. На рисунке 4 приведены примеры полученных моделей для первой и второй компонент, соответствующих трендовой составляющей ряда.

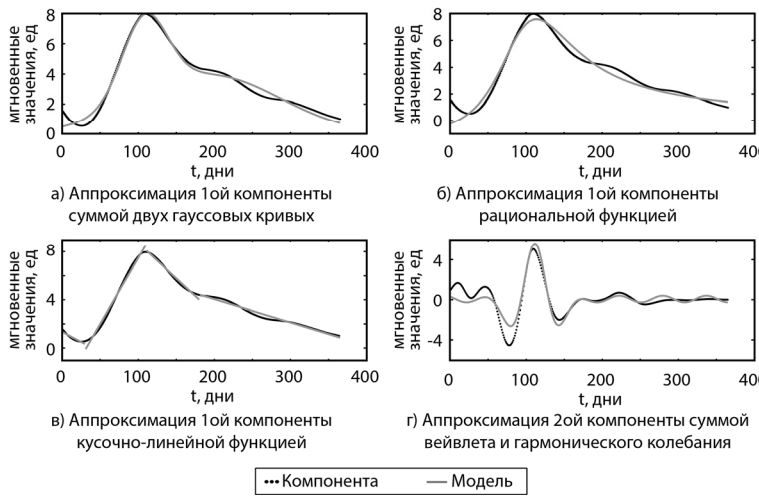

Рис. 4. Аппроксимации первой и второй компонент моделями различного вида

Для первой компоненты можно получить приемлемо адекватные модели, используя сумму двух гауссовых кривых (рисунок 4, а), рациональную функцию (рисунок 4, б) или кусочно-линейную аппроксимацию (рисунок 4, в). Варьируя модели, включаемые в финальную модель ряда, можно решать различные задачи. Так для задач краткосрочного прогнозирования могут использоваться более точные, но сложные модели, а для других задач могут быть использованы более простые в интерпретации модели. Вторая компонента может быть аде-

кватно и компактно описана в виде суммы вейвлета (мексиканская шляпа) и гармонического колебания (рисунок 4, г).

Для остальных 3 компонент на данном этапе были использованы вейвлет Морле и полигармонические модели, которые могли бы быть описаны более компактно при наличии в моделетеки соответствующих моделей. Отбор моделей, которые могли бы компактно и адекватно описать данные компоненты для включения в моделетеку, требует дополнительных исследований.

Графики этих компонент и подогнанных моделей приведены на рисунке 5. Из них видно, что некоторые компоненты (рисунок 5, а, б) могут быть компактно описаны с помощью различных вейвлетов. Кроме того можно предположить наличие мультипликативного тренда, имеющего максимум, приходящийся на весенние месяцы. Суммарная модель может быть записана как $y(t) = y_1(t) + y_2(t) + y_3(t) + y_4(t) + y_5(t)$, где:

$$y_1(t) = 5{,}22 \exp\left[-\left(\frac{t - 108{,}1}{41{,}24} \right)^2 \right] + 3{,}98 \exp\left[-\left(\frac{t - 190{,}5}{133{,}7} \right)^2 \right]$$

$$y_2(t) = 0{,}051 + \frac{2 \cdot 26{,}16}{\sqrt{3 \cdot 18{,}6 \cdot \sqrt{\pi}}} \left(1 - \frac{(t - 111{,}9)^2}{18{,}6^2}\right) \exp\left(- \frac{(t - 111{,}9)^2}{2 \cdot 18{,}6^2} \right) + $$
$$+ 0{,}365 \sin(2\pi t / 56{,}15 + 8{,}5);$$

$$y_3(t) = 0{,}92 \sin(0{,}29t + 1{,}97) + 0{,}55 \sin(0{,}27t - 1{,}72) + 0{,}55 \sin(0{,}31x - 0{,}26);$$

$$y_4(t) = \frac{-7{,}933}{\sqrt{2\pi}} \exp\left[- \frac{(t - 91{,}16)^2}{2 \cdot 31{,}43^2} + i 2\pi \frac{(t - 91{,}16)}{31{,}43} \right];$$

$$y_5(t) = 0{,}53 \sin(0{,}43t - 0{,}32) + 0{,}54 \sin(0{,}44t - 0{,}30) + 0{,}28 \sin(0{,}35t + 1{,}41).$$

Отметим, что $y_3(t)$ и $y_5(t)$ можно преобразовать из аддитивной в аддитивно-мультипликативную запись для удобства интерпретации.

На рисунке 5, г) приведены графики результирующей модели и оригинального временного ряда. Доля объясненной дисперсии суммарной модели составила 0,81. Использование суммы 4-х гармоник (вместо 3-х) в моделях $y_3(t)$ и $y_5(t)$, позволяет увеличить коэффициент детерминации R^2 до 0,82, усложнив суммарную модель.

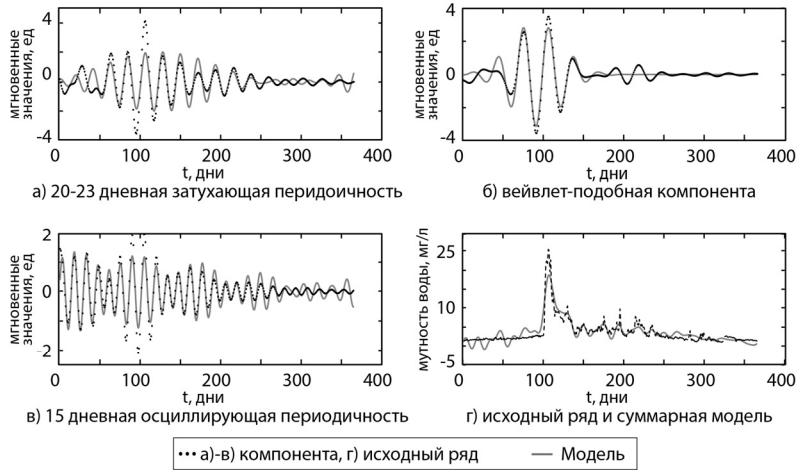

Рис. 5. а) – в) Графики компонент 3–5 и их моделей $y_i(t)$; г) график исходного ряда $x(t)$ и суммарной модели $y(t)$

Отметим, что при выборе моделей в данном примере мы ориентировались на компактность моделей и их пригодность для целей разведочного и дескрипторного анализа, а не целей построения точной аппроксимационной модели или прогнозирования. Таким образом, полученная модель, во-первых, не уступает по адекватности модели, полученной с помощью классического подхода с выделением тренда, во-вторых, дает более простое для анализа и интерпретации описание компонент. В частности тренд в данном случае представлен в виде суммы гауссовых кривых, которые более просты в анализе и интерпретации на этапах предварительного и дескрипторного анализа, чем модель экспоненциального сглаживания. При этом данный подход позволяет использовать более сложные суммарные модели, чем классический подход, без усложнения их восприятия и анализа, поскольку суммарная модель представляется в виде набора сравнительно более простых моделей, каждая из которых может рассматриваться независимо. Кроме того предлагаемый метод допускает варьируемость моделей в зависимости от целей исследователя.

5. Заключение. В работе рассмотрен новый подход к построению компактного и интерпретируемого модельного аналитического представления ВР. Подход был проверен на искусственных и на реальных данных, отражающих качество воды. Показана его работоспособность и преимущества перед ранее применяемыми подходами. Он сохраняет преимущества и позволяет устранить недостаток ССА, свя-

занный с отсутствием компактного аналитического модельного представления ряда.

Среди сильных сторон предлагаемого подхода можно отметить: отсутствие необходимости в предположении о стационарности исходного ВР; простота и наглядность получаемых моделей. Кроме того в отличие от классического подхода, при котором строится одна линейная комбинация гармонических функций, при данном подходе мы получаем схожую модель, но разбитую на отдельные группы. Тем самым повышается содержательность модели, так как исследователю доступна информация о связи между различными элементами модели.

Среди слабых сторон подхода можно отметить необходимость диалога с исследователем и значительное количество ручного ввода (группировка компонент, выбор моделей, упрощение результирующей модели) при данной реализации подхода. Данные этапы могут быть автоматизированы, в частности путем автоматической группировки компонент, полученных в ходе разложения с помощью ССА.

Среди направлений дальнейшего исследования можно выделить решение вопросов наполнения моделетеки различными моделями, автоматизацию подхода, в частности автоматизацию группировки компонент, и разработку адекватного алгоритма выбора моделей, одновременно учитывающего как требования к точности моделей, так и требования к их простоте.

Литература

1. *Naumova E.N., Jagai J.S., Matyas B., DeMaria A., MacNeill I.B., Griffiths J.K.* Seasonality in six enterically transmitted diseases and ambient temperature // Epidemiological Infection. 2007. vol. 135. pp. 281–292.
2. *Альсова О.К., Губарев В.В., Локтев В.Б.* Использование вариативного моделирования при идентификации временных рядов инфекционной заболеваемости // Известия Волгоградского государственного технического университета. Серия «Актуальные проблемы управления, вычислительной техники и информатики в технических системах». 2011. Т. 11. №12. С. 42–47.
3. *Губарев В.В.* Алгоритмы спектрального анализа случайных сигналов // Новосибирск: Издательство НГТУ. 2005. 660 с.
4. *Данилов Д.Л., Жиглявский А.А.* Главные компоненты временных рядов: метод Гусеница // СПб.: Издательство Санкт-Петербургского университета. 1997. 307 с.
5. *Губарев В.В.* Вероятностные модели: справочник: в 2 ч. // Новосибирский электротехнический институт. Новосибирск: НЭТИ. 1992. Ч. 2. С. 197–421.
6. *О. К. Альсова, В. В. Губарев, Н. А. Чистяков, С. Г. Юн и др.* Climate, Water, Diseases, Infections (CliWaDIn) // НГТУ. Свидетельство о государственной регистрации базы данных №2011620720 от 04.10.11.; заяв. 01.06.11; №2011620396.

References

1. Naumova E.N., Jagai J.S., Matyas B., DeMaria A., MacNeill I.B., Griffiths J.K. Seasonality in six enterically transmitted diseases and ambient temperature. Epidemiological Infection. 2007. vol. 135. pp. 281–292.

2. Alsova O. K., Gubarev V. V., Loktev V. B. [Use of the variant modeling for the identification of time series of infectious diseases]. *Izvestiya VolgGTU – Bulleting of the Volgograd state technical university*. 2011. vol. 11. no. 12. pp. 42–47. (In Russ.).
3. Gubarev V. V. *Algoritmy spektral'nogo analiza sluchajnyh signalov* [Algorithms for Spectral Analysis of Random Signals]. Novosibirsk: NSTU. 2005. 660 p. (In Russ.).
4. Danilov D., Zhigljavsky A.A. *Glavnye komponenty vremennyh rjadov: metod Gusenica* [Principal Components of Time Series: the Caterpillar Method]. SPB: St.Petersburg University. 1997. 307 p. (In Russ.).
5. Gubarev V. V. *Verojatnostnye modeli: spravochnik: v 2 ch.* [Probabilistic Models: A Handbook: in 2 parts]. Novosib. Elektrotekh. Inst. Novosibirsk. NJeTI. 1992. part 2. pp. 197–421 (In Russ.).
6. Alsova O. K., Gubarev, et al. Climate, Water, Diseases, Infections (CliWaDIn). Novosibirsk State Technical University. Patent RF, no. 2011620720, 04.10.2011. (In Russ.).

Абалов Николай Владимирович — аспирант кафедры вычислительной техники Новосибирского государственного технического университета. Область научных интересов: интеллектуальный анализ данных и вариативное моделирование. Число научных публикаций — 4. nickabalov@yahoo.com. 630073, г. Новосибирск, пр. К. Маркса, 20; р.т. +7-913-714-97-03.

Abalov Nikolay Vladimirovich — Ph.D student of Computer Sciences Department, Novosibirsk State Technical University. Research interests: intellectual data analysis and variative modelling. The number of publications — 4. nickabalov@yahoo.com. 20, Prospekt K. Marksa, Novosibirsk, 630073, Russia; p.t. +7-913-714-97-03.

Губарев Василий Васильевич — заслуженный деятель науки Российской Федерации, заслуженный работник высшей школы Российской Федерации, д-р техн. наук, профессор, кафедра вычислительной техники НГТУ. Область научных интересов: идентификация, измерение характеристик, имитация и прогнозирование случайных сигналов; вероятностное моделирование реальных объектов; статистические прикладные информационные системы; системный анализ в экспериментальных исследованиях; интеллектуальный анализ данных и вариативное моделирование; концептуальные основы информатики. Число научных публикаций — более 500. gubarev@vt.cs.nstu.ru; 630073, г. Новосибирск, пр. К. Маркса, 20; р.т. +7(383)346-11-33.

Gubarev Vasiliy Vasilyevich — Ph.D., Dr. Sci., honored scientist of Russian Federation, honored worker of higher school of Russian Federation, professor of Computer Sciences Department, NSTU. Research interests: identification, measurement of characteristics, simulation and prediction of random signals; probabilistic modeling of real objects; applied statistical information systems; system analysis in experimental research; intellectual data analysis and variative modeling; conceptual foundations of informatics. The number of publications — more than 500. gubarev@vt.cs.nstu.ru; 20, Prospekt K. Marksa, Novosibirsk, 630073, Russia; p.t. +7(383)346-11-33.

Альсова Ольга Константиновна — к-т техн. наук, доцент кафедры вычислительной техники НГТУ. Область научных интересов: исследование и разработка методов и средств прогнозирования временных рядов, компьютерное моделирование систем, интеллектуальный анализ данных. Число научных публикаций — 30. alsowa@mail.ru; 630073, г. Новосибирск, пр. К. Маркса, 20; р.т. +7(383)346-04-92.

Alsova Olga Constantinovna — Ph.D., associate professor of Computer Sciences Department, NSTU. Research interests: research and development of methods and means of time series forecasting, computer modeling of systems, intellectual data analysis. The number of publications — 30. alsowa@mail.ru; 20, Prospekt K. Marksa, Novosibirsk, 630073, Russia; p.t. +7(383)346-04-92.

РЕФЕРАТ

Абалов Н.В., Губарев В.В., Альсова О.К. **Использование методов сингулярного спектрального анализа и моделетеки при идентификации временных рядов.**

Статья посвящена рассмотрению нового подхода к построению компактного и интерпретируемого модельного аналитического представления нестационарных временных рядов. В основе предлагаемого подхода лежит совместное применение методов сингулярного спектрального анализа (ССА) и моделетеки. Под моделетекой понимается упорядоченное множество моделей.

ССА является относительно новым методом анализа нестационарных временных рядов. Слабой стороной ССА является отсутствие аналитического модельного представления ряда, например, в виде суммы простых функций, компактное аналитическое представление которых могло бы быть нагляднее и доступнее для интерпретации, чем совокупность большого количества компонент. Описанный в статье метод вариативного моделирования позволяет устранить отмеченную слабую сторону ССА путем совместного использования его и метода моделетеки для получения компактного и легко интерпретируемого модельного представления изучаемого временного ряда с желаемым уровнем его адекватности ряду, цели и условиям идентификации.

Первый этап предлагаемого метода заключается в разложении исходного временного ряда на компоненты с помощью ССА. Разложение исходного ряда завершается выделением интересующих исследователя компонент. На втором этапе компоненты идентифицируются моделями из априори сформированной моделетеки согласно целям идентификации. Результатом является результирующая модель исходного временного ряда в аддитивной или аддитивно-мультипликативной форме.

В статье рассматривается применимость предлагаемого подхода на примерах идентификации искусственного ряда и реального временного ряда. В качестве реального временного ряда рассматривается нестационарный ряд ежедневных значений мутности воды в реке в г. Челябинске за 2005 г.

Среди сильных сторон предлагаемого подхода можно отметить: отсутствие необходимости в предположении о стационарности исходного временного ряда; простота и наглядность получаемых моделей.

SUMMARY

Abalov N.V., Gubarev V.V., Alsova O.K. **Use of methods of singular spectral analysis and modeleteka for the identification of time series.**

The article discusses new approach for obtaining compact and interpretable analytical model representation of nonstationary time series. Proposed approach is based on joint use of methods of singular spectrum analysis (SSA) and modeleteka. Modeleteka refers to an ordered set of models.

Singular spectrum analysis (SSA) is relatively new method for analysis of non-stationary time series. The weakness of SSA is lack of analytical model representation of time series, e.g., as a sum of simple functions, which could be clearer and easier for interpretation than a large number of components in form of time series.

Described method of variative modeling allows reducing highlighted drawback of SSA. This is achieved through joint use of SSA and method of modeleteka to obtain analytical model representation of time series, providing necessary level of adequacy, compactness and interpretability.

First, time series are decomposed into components using SSA, significant components are selected using formal indicators (e.g. variance contributed by component, etc.). Second, each significant component is identified according to the purpose of identification with simple and interpretable model from preformed modeleteka. The result is final model of time series in additive or additive-multiplicative form.

The article discusses applicability of the proposed approach based on examples of identification of synthetic and real time series. As a real time series a daily turbidity of river in city of Chelyabinsk in year 2005 is considered.

Strengths of the proposed approach include: nonnecessity of assumption of stationarity of the studied time series, simplicity and clarity of produced resulting models.