УДК 519.2

Автоматизация выделения трендовых и периодических составляющих временного ряда в рамках метода «Гусеница»-SSA

Ф. И. Александров, Н. Э. Голяндина

В статье рассматривается задача выделения аддитивных составляющих временного ряда, трендовых и периодических, с помощью метода «Гусеница»-SSA. Метод не предусматривает знания параметрической модели ряда и позволяет работать с зашумленными нестационарными временными рядами. Он, в частности, позволяет выделять амплитудно-модулированные гармонические составляющие, что выгодно отличает его от методов, построенных на разложении Фурье.

В настоящее время уже достаточно хорошо развит аппарат интерактивного применения подхода «Гусеница»-SSA на основе визуального изучения промежуточных результатов применения метода. Основной целью данной работы является автоматизация выделения тренда, а также обнаружения и выделения периодичностей. Автоматизация решения рассматриваемой задачи важна, например, при пакетной обработке данных, а информация, предоставляемая методами автоматической идентификации, может быть полезна также и при интерактивной обработке. Предлагаемые методы управляются заданием параметров и пороговых значений используемых критериев.

Метод «Гусеница»-SSA используется в различных областях для исследования временных рядов, например в метеорологии, гидрологии, социологии, экономике, исследованиях транспортных потоков — везде, где естественно предположить существование тренда и/или периодическое поведение. Разработанные методы идентификации могут быть применены для автоматизации исследования таких рядов.

Работа может быть особенно интересна тем, кто хотел бы использовать преимущества метода «Гусеница»-SSA в автоматическом режиме, а также тем, кому интересны новые подходы в решении задачи выделения тренда и периодических составляющих временного ряда.

1. Введение

Рассмотрим общую задачу исследования структуры вещественнозначного временного ряда $(f_0, ..., f_{N-1})$ длины N. Здесь под структурой понимаются некоторые характеристики или свойства ряда, которые сохраняются с течением времени. Например, может стоять задача нахождения тренда или выделения периодических составляющих. Как правило, для нахождения структуры какого-либо явления необходима повторяемость этого явления (например, повторные выборки в статистике для нахождения характеристик генерального распределения). Если же мы имеем дело с временным рядом, который обычно существует в единственном экземпляре, то повторяемости в общем случае нет. В качестве способа выхода из этой ситуации в стандартных методах анализа временных рядов часто предполагается справедливость параметрической модели ряда (например, ряд состоит из линейного тренда и белого шума) или же стационарность ряда (т. е. априори предполагается повторяемость на уровне первых двух моментов). Однако эти ограничения нередко оказываются слишком жесткими.

Возникает вопрос: как же все-таки получить повторяемость, не накладывая на ряд предварительных ограни-

чений? Решение этой задачи кажется совершенно естественным. Давайте рассмотрим множество отрезков временного ряда заданной достаточно большой длины L (назовем ее длиной окна). Будем рассматривать эти отрезки последовательно, с первой по L-ю точку, со второй по $L+1$-ю и т. д. Рассмотренные отрезки (назовем их векторами L-вложения) будут наследовать свойства ряда. Если ряд содержит тренд (здесь под трендом мы понимаем медленно меняющуюся составляющую ряда), то и вектора вложения будут его содержать; если исходный ряд содержит периодическую составляющую, то такими же будут и вектора вложения. Таким образом, мы приходим к идее исследования всей совокупности векторов вложения для выявления их общей структуры.

Составим из векторов вложения так называемую *траекторную матрицу* размерности $L \times K$, где $K = N - L + 1$ есть число векторов вложения. Она будет иметь вид:

$$\mathbf{X} = \begin{pmatrix} f_0 & f_1 & \dots & f_{K-1} \\ f_1 & f_2 & \dots & f_K \\ \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & \dots & f_{N-1} \end{pmatrix} = [X_1, \dots, X_K], \quad (1.1)$$

$$X_j = \begin{pmatrix} f_{j-1} \\ \vdots \\ f_{j+L-2} \end{pmatrix}.$$

Теперь у нас есть повторяемость, и мы можем попытаться увидеть структуру векторов вложения. Воспользуемся следующим приемом: разложим всю траекторную матрицу на элементарные части (в сумму элементарных матриц), в некотором смысле независимые (ортогональные) и упорядоченные по их вкладу в разложение. Если разложение окажется «удачным», то мы сможем сгруппировать эти элементарные матрицы так, чтобы, например, одна группа соответствовала трендовой составляющей ряда, другая группа — циклической и т. д. Затем просуммируем матрицы внутри каждой группы и вернемся от разложения матриц к разложению ряда на тренд, циклическую составляющую и остаток.

АВТОРЫ

Александров Федор Игоревич
Аспирант кафедры статистического моделирования; Санкт-Петербургский государственный университет, г. Санкт-Петербург

Голяндина Нина Эдуардовна
Кандидат физико-математических наук, доцент кафедры статистического моделирования; Санкт-Петербургский государственный университет, г. Санкт-Петербург

Способом разложения траекторной матрицы, который оказывается очень хорошо согласованным с описанной выше задачей, является сингулярное разложение. Заметим, что если использовать статистические аналогии и рассматривать набор векторов вложения как выборку, то сингулярное разложение, с точностью до центрирования, эквивалентно анализу главных компонент этой многомерной выборки.

В 70–80-х годах прошлого столетия совершенно независимо и в разных точках земного шара практически одновременно возникла идея метода анализа временных рядов, основанного на описанном выше приеме. При этом его создатели приходили к нему с совершенно разных сторон, иногда не видя всей силы метода из-за узкой области, где он применялся. В зарубежной литературе метод наиболее известен под названием SSA (Singular Spectrum Analysis), он возник из анализа хаотического поведения ряда и аттракторов [1]. Название имеет довольно условное отношение к сути метода, так как сингулярный спектр здесь — это набор собственных чисел сингулярного разложения траекторной матрицы, понимаемый как сингулярный спектр соответствующего матрице оператора. В России метод получил название «Гусеница» [2] из-за скользящей процедуры нарезания векторов вложения из исходного ряда (подобно движению гусеницы) и возник из статистических аналогий с методом главных компонент. Мы будем называть метод «Гусеница»-SSA, соединив вместе эти два названия.

РЕЗЮМЕ

Задача

Выделение трендовых и периодических составляющих временных рядов.

Результаты

Рассмотрены алгоритмы идентификации для выделения тренда и сезонной составляющей ряда методом «Гусеница»-SSA.

Кроме описанных выше идей к методу SSA приходили со стороны анализа рядов, управляемых линейными рекуррентными формулами. Эти ряды обладают следующей замечательной особенностью: в сингулярном разложении их траекторной матрицы оказывается только небольшое число ненулевых компонент, причем это число (размерность ряда) не зависит от длины окна L (если L и N достаточно большие). Структура траекторной матрицы, порожденной рядами конечной размерности (точнее, структура пространства, порожденного столбцами траекторной матрицы, т. е. векторами вложения), исследовалась в работах Бухштабера [3]. В работе Cadzow [4] приводится решение задачи отделения рядов конечной размерности от шума путем итеративного применения метода SSA посредством аппроксимации траекторной матрицы матрицей ганкелевого типа (т. е. матрицей, у которой на антидиагоналях стоят одинаковые числа; примером такой матрицы служит траекторная матрица вида (1.1)). Так называемая аппроксимация Cadzow применяется в теории обработки сигналов, и в соответствующих статьях часто даже не упоминается название SSA. Список разных взглядов на метод SSA можно продолжать и дальше.

Ссылки на основную литературу по методу «Гусеница»-SSA можно найти в работах [2, 5–7]. За время своего существования метод расширился, возникли его обобщения для анализа многомерных

временных рядов, анализа изображений, поиска точек разладки в структуре временного ряда. Появились примеры его применения в широком круге областей: гидрологии, медицине, геофизике, экономике и пр.

Изначально идентификация составляющих ряда на основе сингулярного разложения его траекторной матрицы проводилась интерактивно, в основном визуальным способом, с использованием графического представления результатов и опираясь на имеющиеся теоретические сведения. Примеры такой идентификации приведены в [2, 5, 8]. С одной стороны, интерактивность является положительной стороной метода, т. к. дает возможность сознательного его применения со стороны пользователя и приводит к более качественному и глубокому анализу. С другой стороны, в ряде случаев (например, при необходимости анализа большого количества однотипных данных) возникает потребность в автоматизации процедуры идентификации составляющих ряда. На основе свойств разложения, используя теоретический аппарат метода «Гусеница»-SSA, были разработаны методы идентификации трендовых и гармонических (возможно, модулированных) компонент [7]. Методы позволяют автоматизировать процесс идентификации, а также предоставляют дополнительную информацию, которая может быть использована при интерактивной идентификации в анализе рядов со сложной структурой. Для демонстрации работы методов была написана программа AutoSSA, позволяющая выделять искомую составляющую автоматически.

Таким образом, целью работы является описание автоматизации технологии применения метода «Гусеница»-SSA для анализа одномерных временных рядов. Мы будем придерживаться терминологии и обозначений книги [5] (см. также [8]) и опираться на полученные там теоретические результаты. Большое число материалов по методу (теорию, примеры применения, программу, реализующую метод) читатель может найти на сайте `http://www.gistatgroup.com/gus/`.

2. Метод «Гусеница»-SSA

Алгоритм

Обсудим вкратце алгоритм метода «Гусеница»-SSA (более подробно он описан в [8, раздел 1], [5, разделы 1.1, 1.2]). Рассмотрим вещественнозначный временной ряд $F_N = (f_0, ..., f_{N-1})$ длины N.

Алгоритм можно разбить на четыре шага: вложение, сингулярное разложение, группировка и диагональное усреднение. Первые два в совокупности называются *разложением*, последние — *восстановлением*. Основным параметром алгоритма является так называемая длина окна L, $1 < L < N$. Результатом алгоритма является разбиение временного ряда на аддитивные составляющие.

Разложение. Первый шаг, вложение, состоит в формировании из ряда траекторной матрицы $\mathbf{X}$ размером $L \times K$, где $K = N - L + 1$ согласно (1.1). Далее проводится сингулярное разложение матрицы $\mathbf{X}$:

$$\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_d, \quad \mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T, \qquad (2.1)$$

где $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d > 0$ — упорядоченные ненулевые собственные числа матрицы $\mathbf{X}\mathbf{X}^T$, $\{U_i\}_{i=1}^d : U_i \in \mathbb{R}^L$. Соответствующие этим собственным числам собственные вектора $\{V_i\}_{i=1}^d$: $V_i = \sqrt{\lambda_i}\mathbf{X}^T U_i \in \mathbb{R}^K$ будем называть факторными векторами.

Восстановление. На третьем шаге проводится группировка компонент разложения. Разбив $\{1, ..., d\}$ на m непересекающихся подмножеств I_j, получим

$$\mathbf{X} = \mathbf{X}_{I_1} + \mathbf{X}_{I_2} + ... + \mathbf{X}_{I_m}, \quad \mathbf{X}_{I_j} = \sum_{k \in I_j} \mathbf{X}_k. \qquad (2.2)$$

Последним шагом является восстановление рядов $F_N^{(j)}$ по сгруппированным матрицам $\mathbf{X}_{I_J}$. Элемент ряда $f_n^{(j)}$ получается с помощью усреднения вдоль антидиагонали элементов матрицы $\mathbf{X}_{I_J}$ с индексами a, b такими, что $a + b = n + 2$. Таким образом, получаем разбиение ряда

$$F_N = F_N^{(1)} + ... + F_N^{(m)}. \qquad (2.3)$$

Самым неформализуемым шагом является шаг группировки. Вся информация о каждой из компонент $\mathbf{X}_i$ содержится в собственном числе λ_i, а также в собственном U_i и факторном V_i векторах. Собственный и факторный вектора называют *сингулярными векторами*, а совокупность $(\sqrt{\lambda_i}, U_i, V_i)$ — *собственной тройкой*. Поиск компонент для требуемой группировки, главным образом на основе анализа собственных троек, будем называть процедурой идентификации.

Теория

Целью метода является разложение ряда на трендовую составляющую, периодические (циклические) составляющие и шум:

$$F_N = F_N^{(\text{trend})} + F_N^{(\text{cycle})} + F_N^{(\text{noise})}. \qquad (2.4)$$

Заметим, что метод может решать также задачу разбиения ряда на низкочастотную и высокочастотную составляющие, но в данной работе мы не будем рассматривать этот случай. В связи с поставленной задачей возникают следующие вопросы:

1) можно ли выбрать и затем сгруппировать элементарные матрицы в (2.1) так, чтобы получить, возможно, приближенно, разложение (2.4) после последующего диагонального усреднения?
2) если это можно сделать, то
   а) какую длину окна выбирать для лучшей точности разложения?
   б) как идентифицировать матрицы в (2.1), чтобы собрать их в трендовую или циклическую составляющие?

Опишем коротко теорию, помогающую ответить на эти вопросы.

Когда выделение тренда и периодик возможно. Для ответа на вопрос о возможности построения разложения (2.4) было введено понятие разделимости [8, раздел 2], [5, раздел 1.5]. Два ряда, $F_N^{(1)}$ и $F_N^{(2)}$, являются (слабо) разделимыми при длине окна L, если для их суммы F_N существует сингулярное разложение (2.1) (неединственность разложения может быть вызвана совпадающими собственными числами), для которого может быть выбрана группировка, приводящая к (2.3) с $m = 2$.

Оказывается, что понятие такой точной разделимости накладывает слишком жесткие условия на ряды $F_N^{(1)}$ и $F_N^{(2)}$. От полиномиального ряда, например, никакой другой ряд не отделяется. Приведем пример с самыми «мягкими» условиями разделимости. Легко видеть, что ряды $F_N^{(1)} : f_n^{(1)} \equiv \text{const} \neq 0$ и $F_N^{(2)} : f_n^{(2)} = A\cos(2\pi\omega n + \varphi)$

являются разделимыми, если $L\omega$ и $(N+1)\omega$ — целые числа.

При наличии шума (который можно понимать как сумму большого числа гармоник с разными периодами и близкими небольшими амплитудами) точная разделимость вообще оказывается неосуществима. Поэтому более реальным оказывается понятие приближенной разделимости [8, раздел 2.3], [5, разделы 1.5.2, 6.1.2]. Приближенная разделимость чаще всего появляется из асимптотической по N ($N \to \infty$) разделимости при фиксированной длине ряда N.

Отметим два наиболее важных следствия асимптотической разделимости при $L \to \infty$, $K = N - L + 1 \to \infty$:

1) любой ряд, являющийся суммой экспонент и полиномов, асимптотически отделим от периодического ряда;
2) любой гармонический ряд с частотой ω_1 асимптотически отделим от другого гармонического ряда с частотой $\omega_2 \neq \omega_1$; это же верно и для экспоненциально-модулированных гармоник.

Эти два результата дают возможность, в частности, приближенно отделять тренд (понимаемый как медленно меняющаяся аддитивная составляющая ряда) от периодических компонент и от шума, а также разделять периодические компоненты на их гармонические составляющие.

В [8, раздел 2.3], [5, раздел 6.1.2] показано, что качество разделимости определяется величиной $\min(L, K)$. В соответствии с этим длина окна должна быть достаточно большой, но не больше, чем половина длины ряда. Заметим также, что кратность длины окна L периоду улучшает качество отделимости соответствующей периодической компоненты.

Как проводить группировку. Разделимость дает нам возможность выделить аддитивную составляющую. Но для того чтобы проводить идентификацию компонент, нам необходимо знать, сколько компонент соответствует искомой составляющей и какими свойствами они должны обладать.

Для ответа на этот вопрос вводится понятие рядов конечного ранга [8, раздел 3.1], [5, раздел 5]. К таким рядам относится любой ряд, являющийся суммой произведений полиномов, экспонент и гармоник. Ряд конечного ранга обладает тем свойством, что число ненулевых компонент сингулярного разложения (2.1) его траекторной матрицы конечно, причем оно не зависит от L (при достаточно больших L и N).

Приведем примеры рядов конечного ранга.

1. Экспоненциально-модулированный (э.-м.) гармонический ряд с общим членом $f_n = A\exp(\alpha n)\cos(2\pi\omega n+\varphi)$, $A \neq 0$, $\omega \in (0, 0.5]$. Ранг этого ряда зависит от частоты ω:
   а) Если $\omega \in (0, 0.5)$, то такой ряд имеет ранг 2. Оба собственных (факторных) вектора тоже имеют э.-м. гармонический вид с теми же экспоненциальным показателем и частотой, как у исходного ряда. Если $L\omega \in \mathbb{Z}$ ($K\omega \in \mathbb{Z}$), $\alpha = 0$ (случай обычной гармонической составляющей), то фазы сингулярных векторов различаются точно на $\pi/2$ (если один записать в виде синуса, то другой запишется косинусом от того же аргумента).
   б) Если $\omega = 0.5$, то ряд имеет ранг 1. Собственный (факторный) вектор имеет вид исходного ряда с теми же α и ω, а именно вид модулированной гармоники с периодом 2.
2. Экспоненциальный ряд с общим членом, заданным в виде $f_n = A\exp(\alpha n)$, имеет размерность 1. Собственный (факторный) вектор тоже имеют экспоненциальный вид, с таким же экспоненциальным показателем, как у исходного ряда.
3. Полиномиальный ряд. Ряд $f_n = \sum_{k=0}^{m} a_k n^k$, где $a_m \neq 0$, при достаточно больших N и L имеет ранг $m+1$, и элементы его сингулярных векторов тоже описываются полиномами степени не выше m.

Перечисленные выше результаты приводят к тому, что, во-первых, при выделении тренда нужно сгруппировать те компоненты, чьи сингулярные вектора медленно меняются. Их количество неизвестно, т. к. в реальных временных рядах тренд, скорее всего, только аппроксимируется некоторой суммой экспонент, полиномов и/или гармоник с большими периодами. Во-вторых, при выделении периодической компоненты с периодом T надо искать пары компонент сингулярного разложения, соответствующие периодам P, таким что T/P — целое. Каждая пара отличается тем, что ее изображение на двумерной диаграмме имеет вид «кружка» или «спирали». Исключение составляет случай $P = 2$, которому соответствует единственная собственная тройка с пилообразными сингулярными векторами.

Использование собственных чисел. Рассмотрим, какую информацию несут в себе собственные числа и как ими пользоваться при исследовании ряда и идентификации.

Во-первых, отношение $\sum_{k\in I}\lambda_k / \sum_{i=1}^{d}\lambda_i$ отражает вклад составляющей, восстановленной по группе компонент $I \subset \{1, ..., d\}$, в разложение исходного ряда. Так как мы упорядочили компоненты в порядке убывания собственных чисел, компонента с меньшим номером вносит больший вклад в вид ряда, а незначительные компоненты имеют большие номера. Этим можно пользоваться, в частности, для ограничения количества исследуемых компонент.

Известно, что при целых $L\omega$ и $K\omega$ собственные числа обоих компонент, соответствующих гармонической составляющей ряда, будут равны между собой. Для экспоненциально-модулированного гармонического ряда с невозрастающей амплитудой это свойство будет асимптотическим при $\min(L, K) \to \infty$ [8, раздел 3.2], [5, предложение 5.3]. Поэтому компоненты, соответствующие гармоническим (и э.-м. гармоническим) составляющим, будут соседними, а на графике собственных чисел им будет соответствовать «ступенька».

3. Пример интерактивного исследования ряда

На примере исследования реального временного ряда покажем, как пользоваться методом «Гусеница»-SSA для выделения тренда и периодической составляющей. Рассмотрим ряд TRAFFAT (traffic fatalities), содержащий ежемесячные данные о дорожных авариях с 1960 по 1974 годы в Онтарио. Ряд изображен на рис. 1, длина ряда равна 180. При взгляде на него видно, что присутствует составляющая, задающая регулярные пики, скорее всего ежегодные (сезонные). Поэтому возьмем длину окна $L = 60$, достаточно большую и кратную периоду 12. Для рассматриваемого ряда можно было бы взять $L = 84$, однако в данном случае такой выбор приводит к смешиванию компонент ряда в силу близких собственных чисел, им соответствующих.

После того как разложение (2.1) получено, нам необходимо идентифицировать его компоненты, относящиеся к тренду и сезонности.

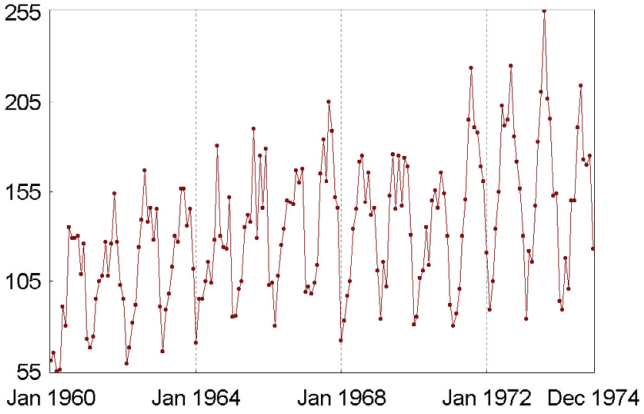

Рис. 1. Исходный ряд TRAFFAT.

Выделение тренда

Для идентификации тренда рассмотрим одномерные графики собственных векторов, первые четырнадцать из которых изображены на рис. 2.

Как мы знаем, для выделения тренда необходимо сгруппировать собственные тройки с медленно меняющимися сингулярными векторами. Видно, что 1-й вектор в целом изменяется медленно, и поэтому ET1 отнесем к трендовой группе. Вектора 4 и 5 тоже обращают на себя внимание тем, что, хотя, казалось бы, они меняются медленно, в них присутствует аддитивная осциллирующая составляющая. Такое бывает из-за неточной разделимости, когда составляющая ряда, которая задает медленно меняющуюся часть сингулярных векторов, смешивается с гармоникой. Тем не менее, поскольку медленно меняющаяся составляющая очевидно имеет существенно больший вклад, чем осциллирующая, включим компоненты ET4,5 в трендовую группу. Результат восстановления по компонентам приведен на рис. 3.

Выделение сезонной составляющей

Займемся теперь выделением сезонной составляющей. В ее разложении могут потенциально присутствовать гармоники с периодами $P = 12, 6, 4, 3, 2.4, 2$. Гармонике с периодом 2 соответствует одна собственная тройка с пилообразным собственным вектором, поэтому будем специально ее искать на одномерных графиках собственных векторов (рис. 2). 8-й собственный вектор обладает требуемой формой, поэтому идентифицируем компоненту ET8 как соответствующую гармонике с периодом 2. На рис. 2 также обращают на себя внимание последовательные пары собственных векторов, похожие на гармонические ряды с одинаковым периодом. Однако пары собственных векторов, соответствующих модулированным гармоническим составляющим ряда, удобнее искать на двумерных диаграммах, на которых по одной оси откладываются элементы первого вектора из пары и по другой — элементы второго вектора. Из-за свойств собственных чисел, соответствующих гармоникам, нам достаточно рассматривать 2D диаграммы только соседних собственных векторов. Приведем диаграммы таких пар на рис. 4.

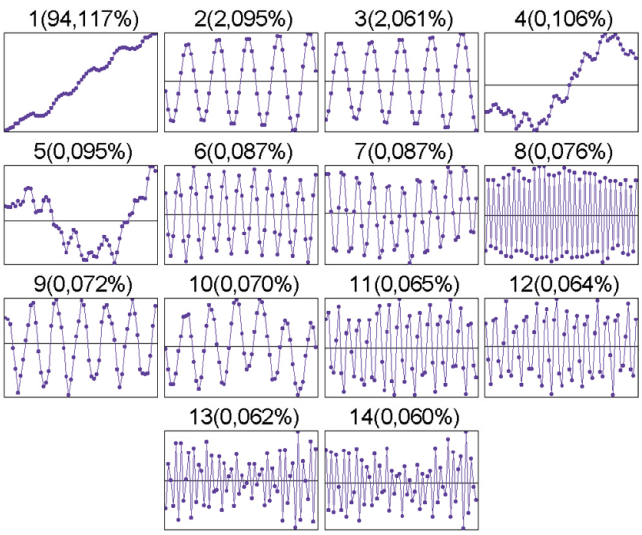

Рис. 2. Собственные вектора ET1–14.

Видно, что ET2–3, 6–7, 9–10, 11–12 образуют достаточно четкие спирали. Пара ET13–14 выделяется регулярностью поведения, диаграмма имеет четкий рисунок, достаточно симметричный относительно точки 0. Такая фигура соответствует э.-м. гармонике с нецелым периодом. Отсюда делаем вывод, что в ряде, возможно, присутствуют гармоники, которым соответствуют ET2–3, 6–7, 8, 9–10, 11–12, 13–14.

Теперь осталось найти все возможные гармоники, которые могут формировать сезонную составляющую. Для этого необходимо оценить периоды найденных гармоник. Будем использовать те же двумерные диаграммы, а оценивать период — через полярный угол одного «шага» на диаграмме. Если обозначить за $\bar{\Delta}$ средний угол, соответствующий одному шагу, то оценка выражается в виде $2\pi/\bar{\Delta}$. Оценки периодов для выделенных гармоник записаны в табл. 1.

Таблица 1. Оценка периодов для выделенных гармоник

ET	2–3	6–7	9–10	11–12	13–14
Оценка периода	11.97	5.97	9.86	4.00	2.43

Требуя, чтобы период гармоники P был таким, что $12/P \in \mathbb{Z}$, получаем, что сезонной составляющей соответствуют компоненты ET2, 3, 6–8, 11–14. Результат восстановления по ним приведен на рис. 3 (красный ряд).

Таким образом, с помощью интерактивного анализа визуального представления результатов сингулярного разложения траекторной матрицы ряда нам удалось выделить тренд и сезонную составляющую ряда.

4. Методы автоматической идентификации

В этом разделе мы опишем методы автоматической идентификации, которые также предоставляют дополнительную информацию при интерактивном выделении из ряда нужной составляющей. В основе разработанных методов лежит работа [7]. Описанные далее методы идентификации основаны на анализе сингулярных векторов.

Идентификация компонент, соответствующих тренду

В основание работы методов идентификации тренда положим следующую идею: сингулярные вектора компонент, соответствующих тренду, ведут себя подобно самому тренду (этим мы пользовались, когда выделяли тренд вручную; тогда мы, просто рассматривая сингулярные

вектора, решали, какие из них изменяются медленно). Исходя из этого, сформулируем методы идентификации тренда в применении к произвольному ряду (для наших задач — к последовательности элементов сингулярного вектора).

Метод Кендалла. Метод Кендалла основан на свойствах рангового коэффициента корреляции Кендалла [9, разделы 4.8, 5.5, 5.6]. Для ряда $G_M = (g_0, ..., g_{M-1})$ длины M рассмотрим коэффициент корреляции Кендалла $\tau(G_M)$ между последовательностью его элементов и участком натурального ряда $H_M = (1, ..., M)$: $\tau(G_M) = 2\mathcal{K}(G_M)/[M(M-1)] - 1$, где $\mathcal{K}(G_M)$ — количество пар (i, j), $i < j$, таких что $g_i < g_j$. Будем проверять гипотезу независимости элементов ряда G_M. Известно, что критерий независимости, основанный на статистике τ, является мощным относительно альтернативы с монотонным трендом. В то же время такой критерий маломощен против альтернативы со «стационарным» трендом вроде синуса. Перейдем к следующей мере, отражающей наличие монотонного тренда в G_M:

$$\alpha(G_M) = 2 - 2\Phi\left(\frac{|\tau(G_M) - a_M|}{\sqrt{\sigma_M}}\right),$$

где

$$a_M = \frac{2}{M(M-1)}, \qquad \sigma_M = \frac{2(2M+5)}{9M(M-1)},$$

а Φ — функция нормального распределения с нулевым средним и единичной дисперсией.

Для заранее заданного порогового уровня $\alpha_0 \in (0, 1)$ будем считать, что если ряд $G_M = (g_0, ..., g_{M-1})$ такой, что выполняется неравенство $\alpha(G_M) \le \alpha_0$, то он содержит монотонный тренд.

Заметим, что данный метод ориентирован на выделение монотонного тренда, то есть про выделенную им составляющую с большой долей уверенности можно сказать, что она содержит именно монотонный тренд.

Метод нулей. Следующие два метода основаны на представлении о тренде как о медленно меняющейся, необязательно монотонной, составляющей ряда.

Для рассматриваемого ряда $G_M = (g_0, ..., g_{M-1})$ считается

$$\mathcal{N}(G_M) = \#\{i : g_i g_{i+1} \le 0, |g_i - g_{i+1}| > \varepsilon, 0 \le i \le M-2\},$$

где параметр ε — заданный допуск, сравнимый с точностью вычислений. Для очень маленького ε величину $\mathcal{N}(G_M)$ можно воспринимать как количество нулей кусочно-линейной функции, образованной рядом G_M следующим образом: в точках i, $0 \le i \le M-1$, она принимает значения g_i. Отсюда и название метода.

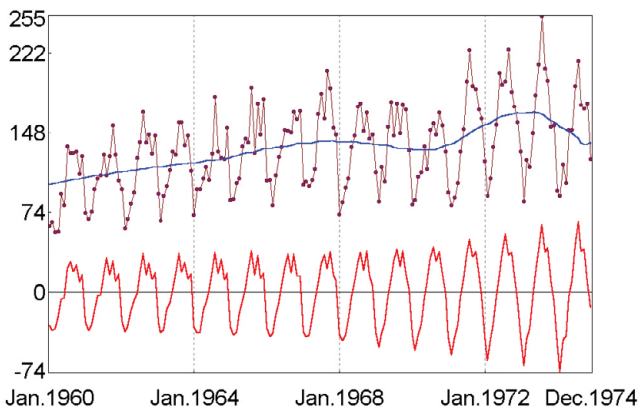

Рис. 3. Исходный ряд, тренд (ET1, 4, 5) и сезонная составляющая (ET2, 3, 6–8, 11–14).

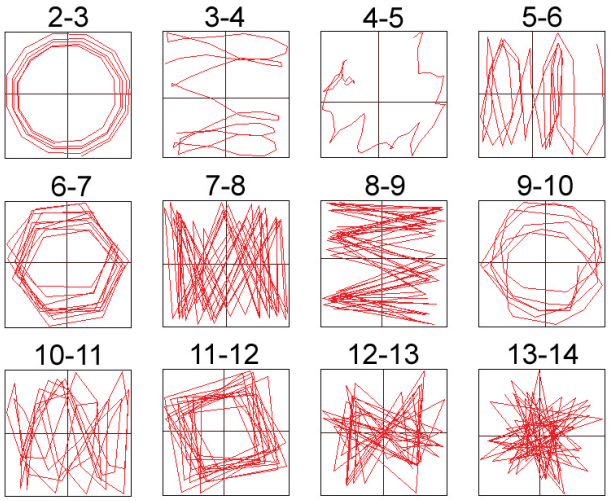

Рис. 4. Двумерные диаграммы для собственных векторов ET2–14.

Для ряда G_M, являющегося трендом, значение $\mathcal{N}(G_M)$ не должно быть большим, поэтому введем пороговое значение $\mathcal{N}_0$ — ограничение сверху на $\mathcal{N}(G_M)$. При этом будем считать, что G_M является трендом, если имеет местот неравенство $\mathcal{N}(G_M) \le \mathcal{N}_0$.

Метод низких частот. Метод нулей для определения того, насколько медленно меняется ряд, использует искусственную характеристику — число нулей. Опишем более естественный подход, основанный на периодограмме.

Рассмотрим разложение Фурье вещественного временного ряда $G_M = (g_0, ...g_{M-1})$:

$$g_n = c_0 + \sum_{1 \le k \le (M-1)/2} \left[c_k \cos\frac{2\pi nk}{M} + s_k \sin\frac{2\pi nk}{M}\right] + (-1)^n c_{M/2},$$

$$0 \le n \le M-1,$$

где $k \in \mathbb{Z}$ и $c_{M/2} = 0$, если M — нечетное. *Периодограммой* $\Pi_G^M(\omega)$ *ряда* G_M назовем функцию, определенную следующим образом при $\omega \in \{k/M\}_{k=0}^{[M/2]}$:

$$\Pi_G^M(k/M) = \frac{M}{2}\begin{cases} 2c_0^2, & k = 0, \\ c_k^2 + s_k^2, & 1 \le k \le \dfrac{M-1}{2}, \\ 2c_{M/2}^2, & M\text{— четное и } k = M/2. \end{cases}$$

Видно, что значение $\Pi_G^M(\omega)$, $\omega \in \{k/M\}_{k=0}^{[M/2]}$, отражает вклад в разложение ряда G_M гармоники с частотой ω.

Будем считать, что ряд является трендом, если гармонические составляющие с низкими частотами дают большой вклад в его разложение Фурье. Задав параметр ω_0, $0 < \omega_0 < 0.5$, будем считать областью низких частот интервал $[0, \omega_0]$. Вычислим для ряда G_M отношение

$$\mathcal{C}(G_M) = \frac{\sum\limits_{M\omega_0 < k \le M/2} \Pi_G^M(k/M)}{\sum\limits_{0 < k \le M/2} \Pi_G^M(k/M)}.$$

Величину $\mathcal{C}(G_M)$ можно интерпретировать как вклад гармоник со средними и высокими частотами в разложение Фурье последовательности g_0, ... g_{M-1}.

Будем считать, что ряд G_M содержит трендовую составляющую, если $\mathcal{C}(G_M) \le \mathcal{C}_0$ для заданного порогового уровня $\mathcal{C}_0$.

Идентификация компонент, соответствующих э.-м. гармонике

Метод Фурье, применяемый нами для автоматической идентификации компонент, соответствующих э.-м. гармонической составляющей, основан на анализе периодограмм сингулярных векторов. Алгоритм метода Фурье можно поделить на две части.

Метод Фурье, часть 1. Воспользуемся тем, что периодограммы двух сингулярных векторов, соответствующих э.-м. гармонике, должны иметь максимумы при одном и том же значении частоты. Это и будем проверять. Пусть для рассматриваемой пары компонент с номерами i и $i+1$ величины θ_i и θ_{i+1} — аргументы максимумов периодограмм их сингулярных векторов. Введем пороговое значение метода s_0, $s_0 \in \mathbb{Z}$. Если $M|\theta_i - \theta_{i+1}| \le s_0$, то будем считать, что пара $(i, i+1)$ соответствует э.-м. гармонике. Заметим, что θ_i является оценкой частоты найденной э.-м. гармоники. Конечно, поиск компоненты, соответствующей э.-м. гармонике с периодом 2, должен проводиться отдельно, т. к. ее ранг равен 1. Для этого используется критерий $M|\theta_i - 0.5| \le s_0$.

Метод Фурье, часть 2. В первой части метода мы использовали только одно свойство периодограммы — аргумент ее максимума. Этого недостаточно, т. к. метод может ошибочно идентифицировать пары компонент, вовсе не соответствующих э.-м. гармонике. Учтем тот факт, что два гармонических сингулярных вектора (собственный и факторный), соответствующих гармонике, не только имеют такой же период, как и сама гармоника, но также имеют разницу в фазе, примерно равную $\pi/2$.

Зададим величину $\rho_{\{a,b\}}$, где a и b — номера двух сингулярных векторов $Y_a, Y_b \in \mathbb{R}^M$, формулой

$$\rho_{\{a,b\}} = \frac{1}{2} \max_{0 \le k \le M/2} [\Pi^M_{Y_a}(k/M) + \Pi^M_{Y_b}(k/M)].$$

Нетрудно увидеть, что если элементы векторов Y_a и Y_b образуют гармонические ряды с одной той же частотой ω и сдвигом фазы $\pi/2$, а $M\omega$ — целое число, то $\rho_{\{a,b\}} = 1$.

Воспользуемся этим для усовершенствования метода Фурье. Рассмотрим пары компонент, уже идентифицированные в первой части метода, и будем считать, что пара компонент с номерами a и b соответствует гармонике, только если выполняется $\rho_{\{a,b\}} \ge \rho_0$, где величина $\rho_0 \in (0,1)$ — заранее заданное пороговое значение. Ясно, что чем больше ρ_0, тем строже условие.

Похожим образом формулируется критерий и для гармоники с периодом 2.

Методы оценки частоты гармоники

Напомним, что периодика с периодом T может быть составлена из гармоник с периодами $T_i : T/T_i \in \mathbb{Z}$. Поэтому для того чтобы выделить периодическую составляющую ряда, необходимо найти всевозможные гармонические составляющие и объединить те из них, периоды которых удовлетворяют данному условию. Для этого необходимо уметь оценивать период э.-м. гармоники. Все приведенные далее методы применяются, конечно же, к компонентам, идентифицированным как соответствующие гармонике.

Один способ оценки, работающий с двумерной диаграммой пары собственных или факторных векторов, мы уже использовали при интерактивном выделении сезонной составляющей. Напомним, что оценка строится через средний угол, приходящийся на «шаг» диаграммы. Этот метод быстр и удобен для вычислений, но в пограничных случаях, при больших искажениях гармоники, его погрешность резко увеличивается.

Второй метод использует корни характеристического полинома линейной рекуррентной формулы [5, раздел 5.2] и в общем случае является, возможно, лучшим способом оценки частоты.

Перейдем для удобства описания еще одного метода от периода T_i к частоте $\omega_i = 1/T_i$. Этот метод оценивает частоту через аргумент максимума периодограммы. В случае, когда оцениваемая частота ω_i гармонической последовательности длины M попадает на решетку $\{k/M\}_{k=0}^{[M/2]}$, эта оценка будет точной. В противном случае мы можем получить оценку с ошибкой $1/M$. Поэтому, вообще говоря, лучше рассматривать не периодограмму сингулярного вектора, а периодограмму более длинного ряда, восстановленного по идентифицированной паре компонент (одной компоненте в случае периода 2).

5. Пример применения методов

Применим теперь описанные выше методы идентификации для выделения тренда и сезонной составляющей ряда TRAFFAT. Для исследования будем использовать программу AutoSSA, позволяющую как выделять заданную составляющую, так и проводить разбиение ряда на тренд, сумму периодических составляющих и остаток.

Возьмем $L = 60$, как и при интерактивном исследовании. Применим трендовые методы идентификации ко всем компонентам и исследуем пары соседних компонент с помощью метода Фурье. Пороговые значения задавать пока не будем, лишь установим значения параметров, которые используются при вычислении критериев. Для метода нулей возьмем $\varepsilon = 10^{-4}$. Для метода низких частот в качестве интервала низких частот возьмем $[0, 0.08]$. Число 0.08 было выбрано из следующих соображений. Мы предполагаем наличие сезонной составляющей, а минимальная частота гармоники, входящей в сезонную составляющую, равна $1/12 \approx 0.083$. Так как такая гармоника, а также все остальные гармоники с большими частотами, которые могут составлять сезонную составляющую, не должны быть идентифицированы, то возьмем интервал, лежащий слева от $1/12$, например $[0, 0.08]$. Результаты работы методов приведены в табл. 2.

Таблица 2. Результаты работы трендовых методов идентификации

ET	м. Кендалла α	м. нулей $\mathcal{N}$	м. низк. част. $\mathcal{C}$
1	0.0	0	0.0
2	0.24	9	1.0
3	0.93	10	1.0
4	0.0	1	0.05
5	0.39	2	0.12
6	0.9	20	1.0
7	0.79	20	0.95
8	0.49	59	1.0

Как мы видим, первая компонента была идентифицирована всеми тремя методами, причем вне зависимости от выбранных пороговых значений. То же можно сказать и про ET4. Значение критерия Кендалла для ET5 слишком велико, эта компонента не будет идентифицирована

методом Кендалла ни при каком разумном пороговом значении. Что же касается остальных методов, то если задать для второго метода возможное количество нулей, равное 5% от длины собственного вектора (а именно 3), то ET5 им будет идентифицирована. Значение критерия метода низких частот тоже невелико (0.12), так что 5-ая компонента будет идентифицирована двумя методами как относящаяся к тренду. Напомним, что особенность метода Кендалла в том, что он годен только для выделения составляющих, содержащих именно монотонный тренд. Возможно, поэтому им не была идентифицирована компонента ET5. Включим эту компоненту в трендовую группу, но на самом деле пара компонент ET4, 5 требует отдельного изучения. Таким образом, группа компонент для восстановления тренда — ET1, 4, 5, что совпадает с результатами интерактивного исследования.

Приведем в табл. 3 результаты работы метода Фурье для первых 14-ти компонент разложения.

Таблица 3. Результаты работы «периодических» методов идентификации

Пары ET	м. Фурье, ч. 1	м. Фурье, ч. 2
1–2	5	
2–3	0	0.99
3–4	4	
4–5	0	0.86
5–6	9	
6–7	0	0.96
7–8	20	
8	0	0.98
8–9	24	
9–10	0	0.9
10–11	9	
11–12	0	0.93
12–13	10	
13–14	0	0.86

Так как вторая часть метода Фурье применяется только к тем парам, которые были идентифицированы первой частью метода, то только для них в таблице приведены значения в колонке «м. Фурье, ч. 2». Для всех же остальных значение критерия первой части метода Фурье слишком велико. Компонента ET8 приведена также отдельно, а не только в парах с соседями, так как она была идентифицирована модификацией метода Фурье, ориентированной на идентификацию гармоники ранга 1 (то есть с периодом, равным 2).

Возьмем пороговое значение метода $\rho_0 = 0.8$. Этот выбор основан на исследованиях применения метода Фурье к реальным и модельным рядам. При таком пороговом значении будут идентифицированы компоненты ET2–3, 4–5, 6–7, 8, 9–10, 11–12, 13–14. Обратим внимание на качество идентификации. Видно, что худший результат показывает, в том числе, пара ET4–5. Вспомним, что ранее компонента ET4 была идентифицирована как трендовая всеми тремя методами, а ET5 — методом нулей и методом низких частот. Подобная коллизия должна быть разрешена исходя из условий поставленной задачи. Если требуется разбиение ряда на тренд, периодические составляющие и остаток, то необходимо решить, в какую группу относить пару ET4–5. Если же задача состоит в выделении сигнала, то можно внести эти компоненты как в трендовую, так и в периодическую группу.

Мы хотели выделить сезонную составляющую, для чего необходимо оценить периоды найденных гармоник и сгруппировать «сезонные» гармоники. Применим методы оценки частоты гармоники к соответствующим найденным парам собственных векторов (периодограммный метод применим к восстановленному по паре компонент ряду). Результаты приведены в табл. 4.

Таблица 4. Результаты оценки периодов найденных гармонических составляющих ряда

Пары ET	м. полярн. угла	м. характ. корней	периодограм. м.
2–3	11.97	11.95	12
4–5	42.25	61.80	60
6–7	5.97	5.95	6
8	—	2.00	2
9–10	9.86	9.65	10
11–12	4.00	3.98	4
13–14	2.43	2.40	2.4

Все методы показали в целом похожие результаты для всех гармоник, за исключением гармоники, восстановленной по ET4–5. Вид собственных векторов позволяет заключить, что метод полярного угла ошибся и эта составляющая все же ближе к гармонике с периодом 60.

Видно, что метод автоматической идентификации периодических компонент нашел как компоненты, относящиеся к сезонной составляющей ряда (ET2–8, 11–14), так и еще две пары собственных троек, ET4–5 и ET9–10. Пара ET4–5 имеет довольно большой период, соизмеримый с длиной окна, поэтому ее можно как отнести к тренду, так и рассматривать отдельно. Мы отнесем ее к тренду.

Вторую пару ET9–10 с периодом из интервала мы отнесем к остатку, т. к. ее период не соответствует сезонной цикличности.

Таким образом, на основе автоматической идентификации трендовых и периодических компонент, пользуясь дополнительной информацией о рассматриваемом ряде, мы получили разложение ряда такое же, как представлено на рис. 3.

Литература

1. Broomhead D., King G. Extracting qualitative dynamics from experimental data. // Physica D.— 1986.— V. 20.— P. 217–236.
2. Главные компоненты временных рядов: метод «Гусеница» / Под ред. Д. Л. Данилова, А. А. Жиглявского.— СПб.: Пресском, 1997.— 308 с.
3. Бухштабер В. М. Многомерные развертки временных рядов. Теоретические основы и алгоритмы // Обозрение прикл. промышл. матем. Сер. Вероятн. и статист.— 1997.— Т. 4.— Вып. 4.— С. 629–645.
4. Cadzow J. A. Signal Enhancement — A Composite Property Mapping Algorithm // IEEE Transactions on Acoustics, Speech and Signal Processing.— 1988.— V. 36.— P. 49–62.
5. Golyandina N. E., Nekrutkin V. V., Zhigljavsky A. A. Analysis of Time Series Structure: SSA and Related Techniques.— Boca Raton: Chapmap & Hall/CRC, 2001.— 305 p.
6. Elsner J. B., Tsonis A. A. Singular Spectrum Analysis: A New Tool in Time Series Analysis.— New York, London: Plenum Press, 1996.— 164 p.
7. Vautard R., Yiou P., Chil M. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals // Physica D.— 1992.— V. 58.— P. 95–126.
8. Голяндина Н. Э. Метод «Гусеница»-SSA: анализ временных рядов: Учебное пособие.— СПб: ВВМ, 2004.— 76 с.
9. Кендэл М. Ранговые корреляции.— М: Статистика, 1975.— 212 с.