УДК 378/1

Б.П. Бочаров, М.Ю. Воеводина

Харьковская национальная академия городского хозяйства, Харьков

АНАЛИЗ ЭФФЕКТИВНОСТИ АЛГОРИТМА ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ЗНАЧЕНИЙ ВРЕМЕННОГО РЯДА РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ ЗНАНИЙ

В статье приведены результаты имитационного моделирования процесса восстановления пропущенных значений временного ряда результатов тестирования знаний студентов методом сингулярного спектрального анализа временных рядов. Используемый в работе метод анализа для исследования временных рядов с пропусками, дающий точные результаты в достаточно жестких предположениях, оказывается применимым и к реальным рядам с пропусками, приводя в этом случае к приближенным результатам. Получены статистические оценки эффективности алгоритмов заполнения пропусков методом сингулярного спектрального анализа для конкретной предметной области – определения эффективности обучающих воздействий.

Ключевые слова: *временные ряды, сингулярный спектральный анализ, качество образования, тестирование знаний, обучающие воздействия.*

Введение

Формулировка проблемы. Получение качественного профессионального образования представляет собой комплексную проблему, решение которой отвечает существующим и будущим потребностям и вызовам времени. Модернизация управления системой образования – важная социальная задача, решение которой обеспечивает необходимое улучшение качества подготовки специалистов в вузе.

Объективными средствами управления качеством подготовки являются:

- модель образовательного процесса;
- схема оценки качества получаемого человеком образования, согласованная с системой предметных знаний и профессиональных задач в выбранной области деятельности;
- оценка возможности изменения системы образования, обеспечивающего улучшение качества предоставляемого образования;
- информационная система управления качеством образования.

Таким образом, исследования, ориентированные на разработку теоретических принципов, математических моделей и алгоритмов количественной оценки эффективности обучающих воздействий, являются актуальными и своевременными.

Анализ последних исследований. Одной из отличительных особенностей метода SSA в применении к анализу временных рядов без пропусков является то, что с его помощью возможно проводить исследование структуры ряда (выявление трендовой, гармонических и шумовых составляющих) без предположений о модели ряда. Однако прогноз выделенных методом составляющих ряда возможен только в рамках некоторой, достаточно широкой, но все-таки модели этих составляющих. Предполагается, что прогнозируемая составляющая является рядом конечного ранга [1].

Идея заполнения пропусков в рамках метода SSA в большой степени аналогична идее прогноза и состоит в продолжении выделенных этим методом составляющих ряда и их структуры на места пропущенных наблюдений. Соответственно и теоретические результаты относительно условий и способов точного восстановления пропущенных значений в составляющих наблюдаемого ряда с пропусками относятся к рядам конечного ранга. Так же, как и в базовом методе SSA, используемый в работе метод анализа для исследования временных рядов с пропусками, дающий точные результаты в достаточно жестких предположениях, оказывается применимым и к реальным рядам с пропусками, приводя в этом случае к приближенным результатам [1].

Цели статьи и формулировка задачи исследования. В [2, 3] рассматривается применение базового метода SSA для временных рядов результатов тестирования знаний студентов. Получены оценки эффективности статистических обучающих воздействий и эффективности использования фонда гибридной библиотеки. Рассматривались временные ряды результатов тестов (35 попыток в течение семестра без пропусков). Знания студентов, пропустивших по какой-либо причине одну-две попытки, проверялись с помощью опроса традиционными методами. Результаты студентов, пропустивших больше попыток, не рассматривались.

Однако далеко не всегда есть возможность получения временного ряда, который является основным рядом данных для подобного исследования. Например, трудно представить себе студенческую группу со стопроцентным посещением занятий, экзаменов, тестов и т.п. Вследствие, как правило, мы полу-

чаем для исследования ряды с пропущенными элементами.

Целью статьи является определение эффективности алгоритмов заполнения пропусков во временном ряде результатов тестирования знаний методом SSA. С помощью имитационного моделирования получены статистические оценки эффективности исследуемых алгоритмов.

Изложение основного материала исследований

Результатом базового алгоритма метода «Гусеница» – SSA является разложение наблюдаемого временного ряда на аддитивные составляющие. Рассмотрим модификацию метода для анализа рядов с пропусками. Общая структура алгоритма та же, но шаги этапов будут несколько другими.

Пусть исходный временной ряд $F_N = (f_0,..,f_{N-1})$ состоит из N элементов, часть которых неизвестна. Опишем схему алгоритма для случая восстановления первой составляющей ряда $F_N^{(1)}$ на основе суммы двух: $F_N = F_N^{(1)} + F_N^2$.

Первый этап: разложение

1. Вложение. Зафиксируем длину окна $L: 1 < L < N$. Процедура вложения переводит исходный временной ряд в последовательность L-мерных векторов вложения $\{X_i\}_{i=1}^K$, где $K = N - L + 1$. Часть векторов вложения может иметь пропуски. Из векторов вложения без пропусков $X_i, i \in C$ образуем матрицу $\tilde{X}$, которая при отсутствии пропусков совпадает с траекторной матрицей ряда F_N.

2. Нахождение базиса. Пусть $\tilde{S} = \tilde{X} \cdot \tilde{X}^T$. Обозначим $\lambda_1,...,\lambda_L$ – собственные числа матрицы, взятые в невозрастающем порядке $(\lambda_1 \geq \ldots \lambda_L \geq 0)$ и $U_1,\ldots U_L$ – ортонормированную систему собственных векторов матрицы $\tilde{S}$, соответствующих собственным числам, $d = \max\{i : \lambda_i > 0\}$. В [4] предложен формальный вариант заполнения пропусков. Он состоит в замене скалярного произведения векторов на аналогичную формальную процедуру, применимую к векторам с пропусками. Зададим два вектора $A = (a_1,\ldots,a_n)^T$ и $B = (b_1,\ldots,b_n)^T$ и их множества пропущенных компонент A и B соответственно, причем $|A \cup B| < n$. Если ввести операцию "$*$" т.о.:

$$(A,B)^* = A^T * B = \frac{n}{n - |A \cup B|} \sum_{k:k \notin A \cup B} a_k b_k ,$$

то при умножении векторов без пропусков результат выполнения операции совпадает со скалярным произведением, а для векторов с пропусками будет численно заменять скалярное произведение.

В качестве матрицы $\tilde{S}$ можно взять матрицу, вычисляемую по формуле $\tilde{S} = X * X^T$, где X – траекторная матрица ряда F_N, содержащая пропуски.

Предложенный выше способ обобщим следующим образом: рассмотрим величину τ, $0 \leq \tau \leq L$, которую назовем *порогом количества пропущенных компонент*. Далее образуем матрицу $\tilde{X}_{(\tau)}$, состоящую из векторов вложения, содержащих не более τ пропущенных компонент, и $\tilde{S} = \tilde{X}_{(\tau)} * \tilde{X}_{(\tau)}^T$. Заметим, что матрица $\tilde{X}_{(0)}$ совпадает с матрицей $\tilde{X}$, составленной из векторов без пропусков, а $\tilde{X}_{(L)} = X$.

Второй этап: восстановление

3. Проекция векторов вложения. В начале производится выбор подпространства и проекция векторов вложения без пропусков. Выбирается набор номеров $I_r = \{i_1,\ldots,i_r\} \subset \{1,\ldots,d\}$, с помощью которых образуется подпространство $M_r = Sp(U_{i_1},\ldots,U_{i_r})$, соответствующее выделяемой компоненте. Выбор собственных векторов, соответствующих $F_N^{(1)}$, происходит аналогично тому, как это делается на этапе группировки в базовом алгоритме SSA. Одним из признаков нужного собственного вектора является то, что его форма подобна форме составляющей ряда $F_N^{(1)}$. Затем происходит проектирование векторов вложения без пропусков на выбранное подпространство M_r: $\hat{X}_i = \sum_{k \in I_r} (X_i, U_k) U_k$, $i \in C$. Затем строится проекция векторов вложения с пропусками. Для каждого вектора вложения с пропусками на местах из P (множество P свое для каждого вектора) процедура состоит из двух частей: (а) вычисления $\hat{X}_i$ для $I \setminus P$ и (б) вычисления $\hat{X}_i$ для P. Так как соседние вектора вложения имеют общую информацию, то ее использование приводит к большому числу возможных способов решения поставленной задачи, в том числе и для векторов $I = P$.

Для заполнения векторов с пропусками необходимо ввести разбиение всего множества пропусков на группы таким образом, чтобы не менее чем L подряд идущих известных значений ряда разграничивают группы пропущенных значений. Вычислительную процедуру (б) можно применять независимо к каждой группе пропусков (каждому набору векторов). Существуют различные способы восстановления позиций пропущенных компонент в наборе векторов вложения группы пропущенных значений, их можно условно разделить на две группы:

– *одновременный метод восстановления* (применение формулы, выражающей вектор X_P

через $X_{I\backslash P}$);

– группа *способов последовательного заполнения* (справа, слева, с двух сторон до середины, с двух сторон с усреднением). Эти методы основаны на том, что в траекторных матрицах значения на диагоналях с индексами $(i, j), i + j = const$ одинаковые. Поэтому можно восстановить пропуски в одном из векторов вложения, а пропуски в соседних векторах заполнить значениями, полученными по уже восстановленному вектору.

Преимуществом последовательных методов восстановления по сравнению с одновременным являются слабые условия применимости, а недостатком – то, что ошибка восстановления может накапливаться.

Результатом шага 3 является матрица $\hat{X} = [\hat{X}_1 \dots \hat{X}_K]$, служащая аппроксимацией траекторной матрицы ряда $F_N^{(1)}$ при правильном выборе множества I_r.

4. Диагональное усреднение. На последнем шаге базового алгоритма матрица $\hat{X}$ переводится в новый ряд $\tilde{F}_N^{(1)}$ (восстановленный ряд) с помощью операции диагонального усреднения.

Рассмотрим теперь имитационную модель процесса восстановления пропущенных значений временного ряда результатов тестирования знаний методом SSA. В качестве исходных данных используются реальные результаты тестов [2] для 105 студентов, которые выполнили по 35 попыток сдачи тестов.

Введем следующие обозначения: N_s – количество студентов; N_a – количество попыток для каждого студента; $\{f_{ij}\}, i = \overline{0, N_s - 1}; j = \overline{0, N_a - 1}$ – значения временного ряда оценок тестирования знаний для i-го студента в j-й попытке; $\{g_{ij}\}, i = \overline{0, N_s - 1}; j = \overline{0, N_a - 1}$ – значения восстановленного временного ряда оценок тестирования знаний для i-го студента в j-й попытке (использовался базовый метод SSA).

Временные ряды с пропущенными значениями получаем с помощью удаления из исходного $\{f_{ij}\}$ ряда n значений, $n = \overline{1, \tau}$, где τ – порог количества пропущенных компонент. В нашем случае $\tau = 15$.

Получаем семейство временных рядов $\{f_{ij}^{(n,k)}\}, i = \overline{0, N_s - 1}; j = \overline{0, N_a - n - 1}; n = \overline{1, \tau}; k = \overline{1, K_n}$, где K_n – количество вариантов удаления n значений из исходного временного ряда.

Очевидно, что $K_n = C_{N_a}^n$, а общее количество временных рядов (для всех студентов) равно $N_s K_n$. В нашем случае при $n > 4$ количество вариантов превышает 10^6. Так как миллион вариантов вполне достаточно для практических целей и современный компьютер может выполнить расчет такого количества за разумное время, то не будем генерировать больше 10^4 временных рядов для каждого студента. В этом случае конкретные временные ряды генерируются случайным образом с помощью метода Монте-Карло.

Таким образом

$$K_n = \begin{cases} C_{N_a}^n, n \le 4, \\ 10^4, n > 4. \end{cases}$$

Применяем модификацию метода SSA для восстановления временного ряда с пропущенными значениями для каждого ряда из семейства $\{f_{ij}^{(n,k)}\}$. Получаем семейство восстановленных временных рядов $\{g_{ij}^{(n,k)}\}, i = \overline{0, N_s - 1}; j = \overline{0, N_a - 1}; n = \overline{1, \tau}; k = \overline{1, K_n}$.

Исследуем соотношение $\{g_{ij}\}$ и $\{g_{ij}^{(n,k)}\}$.

Будем считать, что для каждого n среднее (по попыткам) относительное значение модуля разности значений восстановленных временных рядов с пропусками и без пропусков есть случайная величина с количеством реализаций $N_s K_n$:

$$y_i^{(n,k)} = \frac{1}{N_a} \sum_{j=0}^{N_a - 1} \frac{\left| g_{ij}^{(n,k)} - g_{ij} \right|}{g_{ij}},$$

$$i = \overline{0, N_s - 1}; \quad n = \overline{1, \tau}; \quad k = \overline{1, K_n}.$$

Назовем эту случайную величину «ошибкой» алгоритма восстановления значений временного ряда с n пропущенными значениями. В [5] показано, что такая величина может быть использована как мера близости временных рядов. Статистический анализ этой случайной величины позволяет принять гипотезу о нормальном законе распределения для всех n с уровнем значимости не менее 0,95.

Математическое ожидание «ошибки» алгоритма восстановления определяется по формуле:

$$\bar{y}^{(n)} = \frac{1}{N_s K_n} \sum_{i=1}^{N_s} \sum_{k=1}^{K_n} y_i^{(n,k)},$$

а стандартное отклонение равно

$$\delta^{(n)} = \sqrt{\frac{1}{N_s K_n - 1} \sum_{i=1}^{N_s} \sum_{k=1}^{K_n} (y_i^{(n,k)} - \bar{y}^{(n)})^2}.$$

В табл. 1 приведены статистические результаты имитационного моделирования алгоритма восстановления временного ряда с пропусками. Для конкретных значений количества пропущенных значений (от 1 до 15) определялись доверительные интервалы «ошибки» алгоритма с уровнем доверия 90%. Статистический анализ показал, что если число пропущенных значений не превышает семи, то

«ошибка» алгоритма восстановления не больше 20%.

Таблица 1

Результаты статистического анализа случайной величины «ошибки» алгоритма восстановления значений временного ряда

Количество пропущенных значений	Среднее значение ($\overline{y}^{(n)}$)	Стандартное отклонение ($\delta^{(n)}$)	Нижняя граница доверительного интервала	Верхняя граница доверительного интервала
1	0,012	0,006	0,002	0,022
2	0,018	0,007	0,006	0,030
3	0,032	0,010	0,016	0,048
4	0,039	0,011	0,021	0,057
5	0,068	0,013	0,047	0,089
6	0,087	0,015	0,062	0,112
7	0,171	0,018	0,141	0,201
8	0,272	0,021	0,237	0,307
9	0,319	0,022	0,283	0,355
10	0,346	0,025	0,305	0,387
11	0,382	0,028	0,336	0,428
12	0,409	0,034	0,353	0,465
13	0,453	0,041	0,386	0,520
14	0,481	0,052	0,395	0,567
15	0,516	0,059	0,419	0,613

Это значит, что расхождение между оценками не превышает один балл по пятибалльной шкале. Семилетний опыт тестирования знаний в Харьковской национальной академии городского хозяйства показал, что при больших ошибках тесты уже не оценивают адекватно знания студентов. Поэтому при количестве пропущенных значений больше семи, алгоритм SSA нельзя использовать для восстановления временного ряда результатов тестирования знаний студентов.

Выводы исследования и перспективы дальнейших исследований

Использование имитационных моделей позволило получить статистические оценки эффективности восстановления значений временного ряда результатов тестирования знаний. Алгоритмы восстанавливают временной ряд с достаточной точностью (ошибка не больше 20%), если количество пропущенных значений не более 7. Применение метода сингулярного спектрального анализа для временных рядов с пропусками позволит определять эффективность обучающих воздействий в ситуациях, когда базовый метод не применим.

Перспективными представляются дальнейшие статистические исследования результатов тестирования студентов, построение статистических и имитационных моделей процесса обучения, разработка экспертных систем, баз знаний и систем поддержки принятия решений, позволяющих выбирать рациональные стратегии обучения для каждого студента.

Список литературы

1. Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of Time Series Structure: SSA and Related Techniques. – London: Chapman & Hall/CRC, 2001. – 305 p.

2. Воєводина М.Ю. К вопросу об определении эффективности обучающих воздействий // Системи обробки інформації. – Х.: ХУ ПС, 2007. – Вип. 2 (60). – С. 153-157.

3. Рябченко І.М., Бочаров Б.П., Воєводіна М.Ю. Особливості формування фонду електронної бібліотеки ВНЗ // Вісник книжкової палати. – 2006. – № 12. – С. 37-39.

4. Голяндина Н., Осипов Е. Метод "Гусеница"-SSA для анализа временных рядов с пропусками // Математические модели. Теория и приложения. – СПб: изд-во НИИХ, 2005. – С. 24-28.

5. Бурнаев Е.В., Оленев Н.Н. Меры близости на основе вейвлет коэффициентов для сравнения статистических и расчетных временных рядов // Труды XLVIII научной конференции МФТИ. Часть VII. –М: МФТИ, 2005. – С. 108-110.

Поступила в редколлегию 21.04.2008

Рецензент: д-р техн. наук, проф. Г.Н. Жолткевич, Харьковский национальный университет им. В.Н. Каразина, Харьков.

АНАЛІЗ ЕФЕКТИВНОСТІ АЛГОРИТМУ ВІДНОВЛЕННЯ ПРОПУЩЕНИХ ЗНАЧЕНЬ ЧАСОВОГО РЯДУ РЕЗУЛЬТАТІВ ТЕСТУВАННЯ ЗНАНЬ

Б.П. Бочаров, М.Ю. Воєводіна

В статті наведені результати імітаційного моделювання процесу відновлення пропущених значень часового ряду результатів тестування знань студентів методом сингулярного спектрального аналізу часових рядів. Метод аналізу, що використовується в роботі для дослідження часових рядів з пропусками, дає точні результати в досить жорстких припущеннях, виявляється також можливість його використання для реальних рядів з пропусками, що приводить у цьому випадку до приблизних результатів. Отримані статистичні оцінки ефективності алгоритмів заповнення пропусків методом сингулярного спектрального аналізу для конкретної предметної області – визначення ефективності навчаючих дій.

Ключові слова: *часові ряди, сингулярний спектральний аналіз, якість освіти, тестування знань, навчаючі дії.*

ANALYZING OF MISSING KNOWLEDGE TESTING RESULTS TIME SERIES VALUES RECONSTRUCTION ALGORITHM EFFICIENCY

B.P. Bocharov, M.Y. Voevodina

The article deals with the imitating modeling of the missing students' knowledge testing results time series values reconstruction process with the time series singular spectrum analysis method. The analysis method, used in this work for research of time series with missing data, makes it possible to get exact results in assumptions strict enough, can be applied to real time series with missing data, leading in this case to the approximate results either. Statistical estimations of missing data filling algorithms with singular spec-

trum analysis method efficiency for a concrete subject domain – training influences efficiency determination –have been received.
Keywords: *time series,* Singular spectrum analysis*, education quality, knowledge testing, training influences.*