Известия Гомельского государственного университета
имени Ф. Скорины, № 6(69), 2011

*УДК 004.8*

# О приведении последовательностей данных к нормированному виду для преобразования методом сингулярного спектрального анализа

Е. А. Якимов, О. М. Демиденко, Д. М. Албкеират, А. А. Ковалевич

Рассматриваются методика и результаты исследования постоянной составляющей в последовательностях данных, их сжатие/растяжение и оценка влияния на качество сингулярного спектрального анализа. Предложено аналитическое выражение для получения нормированного временного ряда, предназначенного для универсального применения метода сингулярного спектрального анализа.
**Ключевые слова:** сингулярный спектральный анализ, последовательность данных; постоянная составляющая, сжатие/растяжение, нормированный ряд.

The article considers the technique and results of the research of a constant component in data sequences, their compression/expansion and the estimation of influence on singular spectrum analysis quality. There is offered the analytical expression for reception of normalized time sequence intended for universal application of a singular spectrum analysis method.
**Keywords:** singular spectrum analysis, data sequence, constant component, compression/expansion, normalized line.

## Введение

В общем случае знания представляют собой обобщенное описание данных, которое отражает существенные закономерности, присущие исследуемым объектам. Эти закономерности могут принимать вид функциональных, логических или структурных связей. Знания, в отличие от данных, выполняют активную функцию: на основе знаний принимаются решения, вырабатываются стратегии, осуществляется планирование, проводится оптимизация и решаются другие задачи.

Интеллектуальный анализ данных (ИАД) – процесс управляемого извлечения зависимостей из больших баз данных. В этом процессе центральное место занимает автоматическое порождение моделей, правил и/или функциональных зависимостей, характеризующих анализируемые данные. В целом, процесс извлечения знаний в ИАД условно делят на следующие этапы [1], которые в совокупности могут быть использованы на этапе эксплуатации имитационной модели.

Шаг 1. Отбор данных: анализ задач пользователя, выбор целевого множества данных, определение переменных.

Шаг 2. Предобработка данных: устранение зашумленности, обработка пропущенных значений, итоговые показатели по группам данных.

Шаг 3. Редукция и проекция данных: ищутся полезные особенности данных для решения поставленных задач, сокращается пространство переменных.

Шаг 4. Поиск закономерностей: выбор метода поиска закономерностей с учетом объема и типа данных, их зашумленности и осуществление поиска закономерностей.

Шаг 5. Оценка и интерпретация найденных закономерностей: оценка и упорядочение закономерностей по их релевантности, проверка согласованности предыдущих и вновь найденных знаний. Возможно возвращение к любому шагу от 1 до 4 для дальнейших итераций.

Шаг 6. Использование найденных знаний: прямое использование, передача заинтересованным лицам, включение в интеллектуальные системы, основанные на знаниях.

Применительно к имитационному моделированию деятельности промышленных предприятий, итерационный процесс $M$ извлечения полезной информации представляется композицией следующих операторов:

$$M = T_6 \circ T_5 \circ T_4 \circ T_3 \circ T_2 \circ T_1,$$

где оператор $T_1$ представляет отбор данных, накопленных в комплексных информационных системах предприятий, определение (с привлечением экспертов в исследуемой предметной области) переменных для решения поставленной задачи; $T_2$ – предварительная обработка данных, устранение выбросов и пропущенных данных, применение методов кластерного анализа для группировки данных [2]; $T_3$ – выбор данных для исследования в соответствии с решаемой задачей, формирование XML-файла для автоматизированной обработки; $T_4$ – выбор метода поиска закономерностей, основанных в основном на применении статистических исследований, использовании метода сингулярного спектрального анализа и обобщенного закона распределения для оценки структуры распределения случайных величин; $T_5$ – построение модели исходных данных для применения в имитационной модели на этапе ее эксплуатации; $T_6$ – формирование XML-файла с моделью последовательности данных для последующего применения.

Для разработки автоматизированной технологии извлечения знаний из накопленных баз данных исследован сингулярный спектральный метод анализа (SSA-метод), включающий этапы: вложение, сингулярное разложение, группировку, диагональное усреднение [3].

## 1 Методика исследования SSA-метода на основе информационных технологий

Для исследования *SSA*-метода применяется комплекс информационных технологий, представленный табличным процессором MS Excel, математическим пакетом Mathcad и пакетом статистической обработки данных Statistica [4].

*Этап вложения*. Для экспериментальных исследований последовательность данных представлена временным рядом $G = (g_0, g_1, ..., g_{n-1})$, задается по известным функциям, формируется на рабочем листе MS Excel и затем в пакете Mathcad формируется матрица $A$, которая по правилам построения является ганкелевой [5]. Процедура вложения является преобразованием исходного одномерного ряда $G = (g_0, g_1, ..., g_{n-1})$ в последовательность $L$-мерных векторов, число которых равно $K = n - L + 1$:

$$\mathbf{A}_i = (g_{i-1}, \ldots, g_{i+L-2})^T, \; 1 \le i \le K. \qquad (1)$$

Эти вектора образуют траекторную матрицу $A = [\mathbf{A}_1; \ldots; \mathbf{A}_K]$ ряда $G$, в которой $a_{ij} = g_{i+j-2}$, т. е. матрица $A$ имеет одинаковые элементы на диагонали $i + j = \text{const}$.

*Этап сингулярного разложения*. Обозначим $S = A \cdot A^T \in R^{L \times L}$. Матрица $A \cdot A^T$ симметричная и неотрицательно определенная, а значит ее собственные числа $\{\mu_k\}_{k=1}^{L}$ вещественны и неотрицательны. Представленные в виде $\mu_1 \ge ... \ge \mu_L \ge 0$ собственные числа называют сингулярными значениями матрицы *A*. Пусть $\mathbf{U}_1, \ldots, \mathbf{U}_L$ – соответствующие им ортонормированные собственные вектора. Будем называть $p = max\{k \mid \mu_k > 0\}$ порядком сингулярного разложения. Обозначим

$$\mathbf{V}_k = \frac{1}{\sqrt{\mu_k}} A^T \mathbf{U}_k, \; k = 1, ..., p. \qquad (2)$$

Тогда сингулярным разложением матрицы *A* называется ее представление в виде суммы элементарных матриц

$$A = A_1 + A_2 + ... + A_p, \quad A_k = \sqrt{\mu_k}\, \mathbf{U}_k \mathbf{V}_k^T. \qquad (3)$$

Каждая из матриц $A_k$ имеет ранг, равный единице. Поэтому их можно назвать элементарными матрицами. Вектор $\mathbf{U}_k$ называют *k*-м левым сингулярным вектором или просто *k*-м собственным вектором, вектор $\mathbf{V}_k$ – правым сингулярным вектором. Набор $< \sqrt{\mu_k}, \mathbf{U}_k, \mathbf{V}_k >$ называют *k*-ой собственной тройкой. Обозначим корень собственного числа через $\lambda_k = \sqrt{\mu_k}$ и будем использовать это обозначение в дальнейших исследованиях.

Собственные числа $\{\mu_k\}_{k=1}^{L}$ в пакете Mathcad представлены вектором **d**. Вектор **d** сингулярных значений в Mathcad определяется с использованием функции svds() [6]:

$$\mathbf{d} := \text{svds}(A). \quad (4)$$

Диагональная матрица *ds* сингулярных значений матрицы *A* в пакете Mathcad определяется с использованием функции diag():

$$ds := \text{diag}(\mathbf{d}). \quad (5)$$

Объединенная матрица *AS* с левыми и правыми сингулярными векторами определяется с использованием функции svd ():

$$AS := \text{svd}(A). \quad (6)$$

Для разделения левых и правых сингулярных векторов из матрицы *AS* используется функция submatrix() [6].

*Этап группировки*. Вид левых и правых сингулярных векторов, трактуемых в *SSA* как временные ряды, является очень важным для следующего шага метода – группировки [7]. При этом для одномерного *SSA* левые и правые сингулярные вектора обладают определенной симметрией, так как в этих случаях сингулярные разложения траекторных матриц с длиной окна $L$ и $K = n - L + 1$ эквивалентны.

Процедура группировки формально одинакова для всех разновидностей *SSA*. На основе разложения (3) процедура группировки делит все множество индексов $\{1, \ldots, p\}$ на $m$ непересекающихся подмножеств $I_1, ..., I_m$.

Пусть $I = \{i_1, \ldots, i_p\}$. Тогда результирующая матрица $A_I$, соответствующая группе $I$, определяется как $A_I = A_{i_1} + \ldots + A_{i_p}$. Такие матрицы вычисляются для $I = I_1, ..., I_m$, тем самым разложение (3) может быть записано в сгруппированном виде:

$$A = A_{I_1} + \ldots + A_{I_m}. \quad (7)$$

Процедура выбора множеств $I = I_1, \ldots, I_m$ и называется группировкой собственных троек. Для определения $I = I_1, \ldots, I_m$ в MS Excel используется лепестковая диаграмма, которая является аналогом графика в полярной системе координат, отображая распределение значений относительно начала координат. По особенностям представления сингулярных векторов на лепестковой диаграмме принимается решение о принадлежности их одной группе.

*Этап диагонального усреднения.* На последнем шаге базового алгоритма каждая матрица сгруппированного разложения переводится в новый ряд длины $n$. Для произвольной матрицы *X* процедуру приведения ее к ганкелевому виду и последующему преобразованию в ряд (обозначим его как $\hat{G}$) выразим следующим образом. Пусть $X$ – матрица размера $L \times K$ с элементами $x_{ij}$, $1 \le i \le L$, $1 \le j \le K$. Положим $L^* = min(L, K)$, $K^* = max(L, K)$ и $n = L + K - 1$. Пусть $z_{ij} = x_{ij}$, если $L < K$ и $z_{ij} = x_{ji}$ в остальных случаях. Тогда диагональное усреднение переводит матрицу $X$ в ряд $(\hat{g}_0, \ldots, \hat{g}_{n-1})$ по формуле

$$\hat{g}_k = \begin{cases} \frac{1}{k+1}\sum_{j=1}^{k+1} z_{j,k-j+2} \mid 0 \le k \le L^* - 1; \\ \frac{1}{L^*}\sum_{j=1}^{L^*} z_{j,k-j+2} \mid L^* - 1 \le k \le K^*; \\ \frac{1}{n-k}\sum_{j=k-K^*+2}^{n-K^*+1} z_{j,k-j+2} \mid K^* \le k \le n. \end{cases} \quad (2.8)$$

Это выражение соответствует усреднению элементов матрицы вдоль побочных диагоналей $i + j = k + 2$: выбор $k = 0$ дает $\hat{g}_0 = x_{11}$, для $k = 1$ получаем $\hat{g}_1 = (x_{12} + x_{21})/2$ и т.д. Применив диагональное усреднение к матрицам, полученным на этапе группировки, приходим к разложению исходного ряда в сумму $m$ рядов.

Процедуру диагонального усреднения просто и наглядно предложено выполнить в MS Excel. Для этого матрица, подлежащая диагонализации, размещается на рабочем листе. Затем блок матрицы, следующий за первой строкой, сдвигается вправо на одну позицию. В сдвинутом блоке также определяется блок, следующий за первой строкой, который сдвигается вправо на одну позицию. Процедура повторяется до тех пор, пока в очередном блоке не останется ни одной строки. Восстановленный ряд $\widehat{G}$ определяется аналогично формуле (8) с использованием функции СРЗНАЧ() в MS Excel. Затем исследуется в пакете Statistica [4].

## 2 Численные исследования влияния постоянной составляющей временного ряда на качество сингулярного спектрального анализа

Методика проведения экспериментальных исследований включает выбор исходного временного ряда $G = G_T + G_H + G_N$, который задан известными моделями соответственно трендовой, гармонической и шумовой составляющих.

Задачей исследования, поставленной по результатам предварительного изучения метода сингулярного спектрального анализа [8; 9; 10], является оценка влияния постоянной составляющей тренда на качество восстановления гармонической и шумовой составляющей временного ряда, определяемого моделью по формулам (9)–(11):

$$G = F(x) = F_T(x) + F_H(x) + F_N(x) \mid x = 0, ..., n-1; \qquad (9)$$

где $F_T(x); F_H(x); F_N(x)$ – соответственно трендовая, гармоническая и шумовая составляющая.

Трендовую составляющую $G_T$ определим по формуле

$$G_T = F_T(x) = \frac{2x}{n-1} - 1 \mid x = 0, ..., n-1;\ n = 43. \qquad (10)$$

Представим $G_T$ в виде двух составляющих

$$G_T = G_{TC} + G_{TD},$$

где $G_{TD}$ – динамическая составляющая, $G_{TC}$ – постоянная составляющая, причем $G_{TC} = 0$.

В общем случае гармоническая составляющая с $k$ периодами

$$G_H = F_H(x) = \sin(k \cdot \frac{2\pi x}{n-1}) \mid\ x = 0,\ ..., n-1, k > 0. \qquad (11)$$

В настоящем исследовании будем полагать $k = 2$.

Шумовая составляющая задана моделью (рисунок 1)

$$G_N = Rnd(n_{inf}; n_{sup}), \qquad (12)$$

где $Rnd(n_{inf}; n_{sup})$ – функция, возвращающая случайные равномерные числа в интервале $[n_{\inf}; n_{\sup}]$ ($n_{\inf}$ – нижняя граница значений случайных чисел, $n_{\sup}$ – верхняя граница значений случайных чисел, причем $n_{\inf} = -1$, $n_{\sup} = 1$, чтобы не вносить дополнительных постоянных составляющих при исследовании временного ряда SSA-методом).

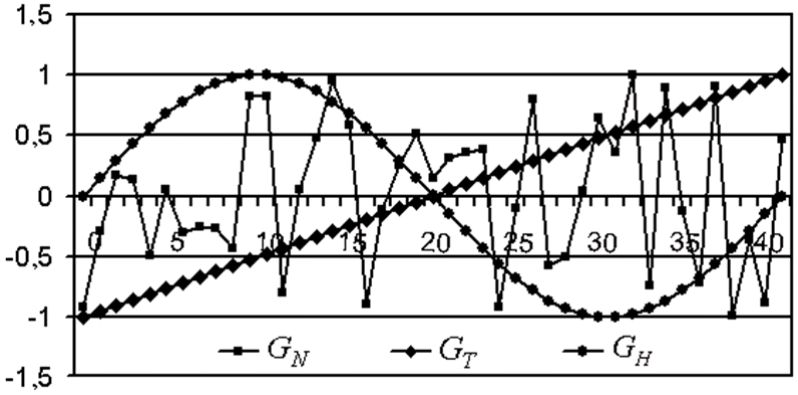


Рисунок 1 – Составляющие временного ряда

При исследовании постоянная составляющая временного ряда выбирается последовательно из набора

$$G_{TCj} \in <-100, -10, -1, 0, 1, 10, 100>, j = 1, \ldots, 7. \quad (13)$$

### 3 Обсуждение результатов исследования SSA-методом постоянной составляющей временного ряда

Для оценивания результатов исследования каждой из составляющих временного ряда используется сумма модулей значений элементов временного ряда, которая именуется Ф-оценкой соответственно для трендовой, гармонической и шумовой составляющих:

$$\Phi_M = \sum_{i=1}^{n} |g_{Mi}|, M = T, H, N, \quad (14)$$

где $g_{Mi}$ – значение *i*-го элемента *M*-ой составляющей временного ряда, *n* – длина временного ряда ($n = 43$).

Оценки (13) принимают следующие значения: $\Phi_{TDj} = 22{,}0$, $\Phi_{Hj} = 26{,}7$, $\Phi_{Hj} = 21{,}9 \mid j = 1, ..., 7$, т.е. вклад каждой из составляющих во временном ряду примерно одинаков.

При анализе временных рядов и сравнении результатов *SSA*-преобразований используются относительные φ-оценки, определяемые соответственно для трендовой, гармонической и шумовой составляющих:

$$\varphi_M = \Phi_M \Big/ min(\Phi_T, \Phi_H, \Phi_N) \mid M = T, H, N. \quad (15)$$

Принятые φ-оценки характеризуют соотношение между значениями составляющих временного ряда, которые оказывают существенное влияние на качество восстановления составляющих SSA-методом [9]. Результаты расчета относительных φ-оценок следующие:

1) $\varphi_{Tj}$ для соответствующих рядов $F_j(x)$, $j = 1, \ldots, 7$ принимает значения из набора <196,5; 19,65; 1,96; 1,01; 1,96; 19,65; 196,5>;

2) $\varphi_{Hj}$ и $\varphi_{Nj}$ принимают одно значение во всех рассматриваемых рядах $F_j(x)$: $\varphi_{Hj} = 1{,}22$, $\varphi_{Nj} = 1$, $j = 1, \ldots, 7$.

При сингулярном спектральном анализе получено распределение корней собственных чисел $\lambda_i$, представленное в таблице 1 и являющееся одной из характеристик для оценки качества восстановления составляющих исходного временного ряда, задаваемого функцией $F_j(x)$, $j = 1, \ldots, 7$.

Таблица 1 – Корни собственных чисел при сингулярном спектральном анализе временных рядов $F_j(x)$

| $\lambda_i$ | $F_1(x)$ | $F_2(x)$ | $F_3(x)$ | $F_4(x)$ | $F_5(x)$ | $F_6(x)$ | $F_7(x)$ |
|---|---|---|---|---|---|---|---|
| $\lambda_1$ | **2198,0** | **218,09** | **21,916** | **10,953** | **25,723** | **222,26** | **2202,0** |
| $\lambda_2$ | **10,158** | **10,122** | **9,631** | **10,927** | **10,421** | **10,197** | **10,165** |
| $\lambda_3$ | **9,199** | **9,247** | **9,549** | 5,274 | **8,526** | **9,138** | **9,188** |
| $\lambda_4$ | 4,931 | 4,928 | 4,892 | 4,936 | 4,968 | 4,935 | 4,932 |
| ... | … | … | … | … | … | … | … |
| $\lambda_{21}$ | 0,643 | 0,645 | 0,665 | 0,538 | 0,627 | 0,641 | 0,643 |
| $\lambda_{22}$ | 0,052 | 0,052 | 0,059 | 0,034 | 0,048 | 0,051 | 0,052 |

Значения корней собственных чисел в таблице 1, которым соответствуют вектора, восстанавливающие трендовую составляющую, выделены полужирным шрифтом, гармоническую составляющую – полужирным шрифтом с подчеркиванием. Остальные $\lambda_i$ определяют шумовую составляющую.

Аналогично Ф-оценкам и относительным φ-оценкам составляющих временного ряда приняты оценки для сингулярной последовательности корней собственных чисел:

$$\Phi_{\lambda M} = \sum_{i=1}^{k} \lambda_i, | M = T, H, N; \tag{16}$$

$$\varphi_{\lambda M} = \Phi_{\lambda M} \Big/ min(\Phi_{\lambda T}, \Phi_{\lambda H}, \Phi_{\lambda N}) | M = T, H, N, \tag{17}$$

где $k$ – количество собственных чисел.

Расчеты абсолютных оценок $\Phi_\lambda$ сингулярной последовательности корней собственных чисел представлены в таблице 2.

Таблица 2 – Абсолютные $\Phi_\lambda$-оценки сингулярной последовательности корней собственных чисел

| Показатель | $F_1(x)$ | $F_2(x)$ | $F_3(x)$ | $F_4(x)$ | $F_5(x)$ | $F_6(x)$ | $F_7(x)$ |
|---|---|---|---|---|---|---|---|
| $\Phi_{\lambda T_j}$, $j$=1, …, 7 | **2200** | **218,1** | **21,92** | – | **25,72** | **222,3** | **2200** |
| $\Phi_{\lambda H_j}$, $j$=1, …, 7 | 19,357 | 19,369 | 19,18 | 21,88 | 18,947 | 19,335 | 19,353 |
| $\Phi_{\lambda N_j}$, $j$=1, …, 7 | 45,517 | 45,471 | 44,968 | 51,231 | 45,934 | 45,57 | 45,528 |

Следует заметить, что $\Phi_{\lambda H_j}$ и $\Phi_{\lambda N_j}$ остаются практическими неизменными для $F_j(x)$, $j = 1, \ldots, 7$; $\Phi_{\lambda T_j}$ изменяет свои значения в соответствии с заданным порядком постоянных составляющих (13). В то же время для $F_4(x)$ с нулевым значением постоянной составляющей не удается восстановить трендовую составляющую SSA-методом.

Для определения качества восстановления составляющих временного ряда приняты следующие показатели:

$$\Delta F_{Tj}(x) = F_{Tj}(x) - \hat{G}_{Tj},\ j = 1, ..., 7;$$

$$\Delta F_{Hj}(x) = F_{Hj}(x) - \hat{G}_{Hj},\ j = 1, ..., 7.$$

где $F_{Tj}(x)$, $F_{Hj}(x)$ – трендовая и гармоническая составляющая исходного временного ряда; $\hat{G}_{Tj}$, $\hat{G}_{Hj}$ – восстановленные трендовая и гармоническая составляющая временного ряда.

Восстановленная шумовая составляющая $\hat{G}_{Nj}$, а также $\Delta F_{Tj}(x)$ и $\Delta F_{Hj}(x)$, $j = 1, \ldots, 7$ определяются характеристиками положения: среднее (mean), медиана (med); характеристиками рассеяния: стандартное отклонение s; максимум (max); минимум (min), диапазон (range), коэффициенты асимметрии распределения ($\gamma_3$) и эксцесса ($\gamma_4$). Показатели качества восстановления исходного временного ряда представлены в таблице 3.

Таблица 3 – Основные показатели качества восстановления составляющих временного ряда SSA-методом

| Оценки | Mean | Med | Max | Min | Range | s | $\gamma_3$ | $\gamma_4$ |
|---|---|---|---|---|---|---|---|---|
| $\Delta F_{T1}(x)$ | –0,071 | –0,096 | 0,071 | –0,142 | 0,213 | 0,064 | 0,899 | –0,410 |
| $\Delta F_{H1}(x)$ | 0,029 | 0,047 | 0,245 | –0,176 | 0,421 | 0,134 | –0,031 | –1,088 |
| $\hat{G}_{N1}$ | –0,029 | –0,086 | 0,973 | –1,163 | 2,136 | 0,557 | –0,127 | –0,796 |

| Оценки | Mean | Med | Max | Min | Range | s | $\gamma_3$ | $\gamma_4$ |
|---|---|---|---|---|---|---|---|---|
| $\Delta F_{T2}(x)$ | –0,071 | –0,095 | 0,088 | –0,138 | 0,226 | 0,062 | 0,914 | –0,291 |
| $\Delta F_{H2}(x)$ | 0,032 | 0,053 | 0,260 | –0,178 | 0,439 | 0,136 | –0,056 | –1,126 |
| $\widehat{G}_{N2}$ | –0,027 | –0,086 | 0,972 | –1,160 | 2,132 | 0,556 | –0,128 | –0,792 |
| $\Delta F_{T3}(x)$ | –0,084 | –0,089 | 0,124 | –0,201 | 0,325 | 0,074 | 0,733 | 0,576 |
| $\Delta F_{H3}(x)$ | 0,065 | 0,078 | 0,610 | –0,229 | 0,839 | 0,187 | 0,535 | 0,222 |
| $\widehat{G}_{N3}$ | –0,006 | 0,027 | 0,971 | –1,139 | 2,111 | 0,550 | –0,143 | –0,674 |
| $\Delta F_{T4}(x)$ | – | – | – | – | – | – | – | – |
| $\Delta F_{H4}(x)$ | –0,007 | –0,059 | 0,873 | –0,862 | 1,735 | 0,495 | 0,096 | –0,699 |
| $\widehat{G}_{N4}$ | 0,006 | –0,027 | 2,057 | –2,142 | 4,199 | 0,755 | –0,119 | 1,304 |
| $\Delta F_{T5}(x)$ | –0,061 | –0,116 | 0,339 | –0,206 | 0,545 | 0,136 | 1,863 | 2,676 |
| $\Delta F_{H5}(x)$ | –0,001 | –0,028 | 0,261 | –0,264 | 0,525 | 0,142 | 0,213 | –0,752 |
| $\widehat{G}_{N5}$ | –0,049 | –0,083 | 0,988 | –1,187 | 2,174 | 0,570 | –0,161 | –0,708 |
| $\Delta F_{T6}(x)$ | –0,071 | –0,096 | 0,085 | –0,145 | 0,230 | 0,067 | 0,983 | –0,196 |
| $\Delta F_{H6}(x)$ | 0,027 | 0,041 | 0,248 | –0,173 | 0,421 | 0,131 | 0,008 | –1,032 |
| $\widehat{G}_{N6}$ | –0,031 | –0,086 | 0,974 | –1,165 | 2,140 | 0,559 | –0,127 | –0,798 |
| $\Delta F_{T7}(x)$ | –0,071 | –0,094 | 0,069 | –0,142 | 0,212 | 0,065 | 0,909 | –0,390 |
| $\Delta F_{H7}(x)$ | 0,029 | 0,045 | 0,245 | –0,175 | 0,421 | 0,133 | –0,024 | –1,078 |
| $\widehat{G}_{N7}$ | –0,029 | –0,086 | 0,973 | –1,163 | 2,136 | 0,558 | –0,127 | –0,797 |

Для проверки согласия восстановленных случайных величин $\widehat{G}_{Nj}$ теоретическому распределению используется критерий Колмогорова – Смирнова. Критическое значение $\Delta_p$ для наибольшего отклонения эмпирического распределения от теоретического при $p = 0{,}01$ и $n = 43$ равно 0,24332. Поскольку наблюдаемое значение во всех случаях меньше критического, гипотеза $H_0$ об известном распределении восстановленных случайных величин принимается.

Основным результатом исследования является вывод о независимости качества восстановления составляющих временного ряда от величины постоянной составляющей исходного временного ряда. Однако следует заметить, что при нулевой постоянной составляющей $G_{TC} = 0$ не удается восстановить динамическую составляющую тренда исходного временного ряда (таблица 3).

## 4 Численные исследования влияния сжатия/растяжения временного ряда на качество сингулярного спектрального анализа

Задачей исследования является оценка влияния процедуры сжатия/растяжения временного ряда на качество восстановления его составляющих методом сингулярного спектрального анализа. Под сжатием/растяжением будем понимать пропорциональное уменьшение/увеличение значений элементов ряда путем умножения на коэффициент 1/α:

$$\frac{1}{\alpha} \to \begin{cases} \text{сжатие} & | \ \alpha > 1; \\ \text{растяжение} & | \ 0 < \alpha < 1. \end{cases}$$

Методика проведения экспериментальных исследований включает выбор исходного временного ряда $G$, который задан по известным функциям (9)–(11) и шумовой составляющей

$$G_N = F_N(x) = Norm(\gamma; \beta),$$

где $Norm(\gamma; \beta)$ – функция, возвращающая случайные нормально распределенные числа; ($\gamma = 0$ – математическое ожидание случайных чисел, $\beta = 1$ – среднеквадратическое отклонение случайных чисел).

Затем исследуются временные ряды вида

$$G_\alpha = \frac{1}{\alpha} \cdot F(x),$$

где $1/\alpha$ – коэффициент сжатия/растяжения.

Для оценивания каждой из составляющих в исходном временном ряду используются оценки (14). Исходные данные для проведения исследований представлены в таблице 4.

Таблица 4 – Исходные данные для исследования временных рядов с нормальным шумом и $1/\alpha_1 = 5$, $1/\alpha_2 = 10$

| $F_j(x)$ | $F_{Tj}(x)$ | $F_{Hj}(x)$ | $F_{Nj}(x)$ | $\Phi_{Tj}$ | $\Phi_{Hj}$ | $\Phi_{Nj}$ |
|---|---|---|---|---|---|---|
| $F_1(x)$ | $5 \cdot (\frac{2x}{(n-1)} - 1)$ | $5 \cdot \sin(\frac{4\pi x}{n-1})$ | $5 \cdot Norm(0; 1)$ | 2150 | 133,4 | 109,4 |
| $F_2(x)$ | $10 \cdot (\frac{2x}{(n-1)} - 1)$ | $10 \cdot \sin(\frac{4\pi x}{n-1})$ | $10 \cdot Norm(0; 1)$ | 4300 | 266,9 | 218,9 |

## 5 Обсуждение результатов исследования SSA-методом сжатия/растяжения временного ряда

Для анализа временных рядов и сравнения результатов *SSA*-преобразований используются относительные φ-оценки (15), определяемые соответственно для трендовой, гармонической и шумовой составляющих. Для $F_1(x)$ и $F_2(x)$ (таблица 4) φ-оценки принимают равные значения: $\varphi_{Tj} = 19{,}65$, $\varphi_{Hj} = 1{,}22$, $\varphi_{Nj} = 1{,}0$, $j = 1, 2$.

При сингулярном спектральном анализе получено распределение корней $\lambda_i$ собственных чисел, представленное в таблице 5 и являющееся одной из характеристик для оценки качества восстановления составляющих исходного временного ряда.

Таблица 5 – Корни собственных чисел при сингулярном спектральном анализе временных рядов $F_1(x)$ и $F_2(x)$

| $\lambda_i$ | $F_1(x)$ | $F_2(x)$ | $\lambda_i$ | $F_1(x)$ | $F_2(x)$ |
|---|---|---|---|---|---|
| $\lambda_1$ | **1110** | **2220** | $\lambda_5$ | 24,416 | 48,869 |
| $\lambda_2$ | **50,963** | **101,944** | … | … | … |
| $\lambda_3$ | **45,683** | **91,4** | $\lambda_{21}$ | 3,222 | 6,449 |
| $\lambda_4$ | 24,652 | 49,34 | $\lambda_{22}$ | 0,235 | 0,457 |

Значения $\lambda_i$ в таблице 3.24, которым соответствуют вектора, восстанавливающие трендовую составляющую, выделены полужирным шрифтом. Корни собственных чисел, соответствующие гармонической составляющей, выделены полужирным шрифтом с подчеркиванием. Остальные $\lambda_i$ определяют шумовую составляющую.

$\Phi_\lambda$-оценки (16), рассчитанные по таблице 5, имеют следующие значения: $\Phi_{\lambda T1} = 1110$, $\Phi_{\lambda T2} = 2220$, $\Phi_{\lambda H1} = 96{,}65$, $\Phi_{\lambda H2} = 193{,}30$, $\Phi_{\lambda N1} = 228{,}0$, $\Phi_{\lambda N2} = 456{,}1$.

Анализ векторов $\mathbf{U}_i$, $i = 1, \ldots, 22$ для временных рядов $F_1(x)$ и $F_2(x)$ не выявил различий между ними. Основные показатели восстановления временных рядов $F_1(x)$ и $F_2(x)$ представлены в таблице 6.

Таблица 6 – Основные показатели качества восстановления составляющих исходного ряда SSA-методом

| Оценки | Mean | Med | Max | Min | Range | s | $\gamma_3$ | $\gamma_4$ |
|---|---|---|---|---|---|---|---|---|
| $\Delta F_{T1}(x)$ | –0,348 | –0,473 | 0,428 | –0,716 | 1,144 | 0,334 | 0,981 | –0,204 |
| $\Delta F_{H1}(x)$ | 0,134 | 0,207 | 1,235 | –0,870 | 2,106 | 0,658 | –0,002 | –1,039 |
| $\hat{G}_{N1}$ | –0,156 | –0,440 | 4,884 | –5,827 | 10,711 | 2,794 | –0,125 | –0,800 |
| $\Delta F_{T2}(x)$ | –0,694 | –0,944 | 0,858 | –1,430 | 2,288 | 0,669 | 0,980 | –0,206 |
| $\Delta F_{H2}(x)$ | 0,267 | 0,418 | 2,464 | –1,740 | 4,204 | 1,314 | –0,004 | –1,040 |
| $\hat{G}_{N2}$ | –0,312 | –0,870 | 9,757 | –11,658 | 21,415 | 5,589 | –0,125 | –0,800 |

В соответствии с результатами таблицы 6, среднее ошибки восстановления трендовой составляющей $\Delta F_{T1}(x)$ находится в соотношении 1 : 2 к среднему ошибки восстановления трендовой составляющей $\Delta F_{T2}(x)$. В таком же соотношении находятся и другие показатели в таблице 6: Med, Max, Min, Range, s. При этом $\gamma_3$ и $\gamma_4$ не изменяются. Это же справедливо и для ошибок восстановления гармонической составляющей, а так же шумовой составляющей.

Таким образом, можно утверждать об исследовании нормированного временного ряда, который рассчитывается по следующей формуле:

$$G_{\text{н}} = \beta \frac{G - \text{mean}(G)}{\alpha} + \gamma,$$

где $\text{mean}(G)$ – среднее элементов $x_i$, $i = 0, \ldots, n-1$ временного ряда $G$:

$$\text{mean}(G) = \frac{1}{n} \sum_{i=0}^{n-1} x_i ;$$

$1/\alpha$ – коэффициент сжатия/растяжения, при этом параметр $\alpha$ определяется наибольшим по модулю значением элемента ряда $G - \text{mean}(G)$:

$$\alpha = \max_{x_i \in (G - \text{mean}(G))} \left| G - \text{mean}(G) \right| ;$$

$\beta$ – нормирующий параметр масштаба, определяемый верхней границей исследуемых значений элементов нормированного временного ряда, как правило, $\beta = 1$ или $\beta = 10$;

$\gamma$ – параметр сдвига элементов временного ряда в область положительных вещественных чисел, $\gamma \geq \beta$.

Пример. Пусть имеем временной ряд $G$.

$G =$ 950, 956, 971, 975, 980, 985, 990, 982, 992, 987, 993, 995, 1000, 1005, 1010, 1002, 994, 998, 1012, 994, 1020, 1034, 1044, 1055, 1040, 1050, 1042, 1052, 1043, 1054, 1056, 1066, 1076, 1058, 1068, 1072, 1071, 1061, 1056, 1060, 1076, 1077, 1090.

Методика приведения исходного ряда $G$ к нормированному виду состоит из следующих шагов:

Шаг 1. Находится величина $\text{mean}(G)$ временного ряда: $\text{mean}(G) = 1025{,}4$.

Шаг 2. Определяется разность между значениями элементов ряда $G$ и его средним $\text{mean}(G)$.

Шаг 3. Находится параметр $\alpha$ как максимальное по модулю значение из элементов ряда $G - \text{mean}(G)$, полученных на шаге 2: $\alpha = 75{,}395$.

Шаг 4. Полученные на шаге 2 элементы ряда $G - \text{mean}(G)$ умножаем на коэффициент $1/\alpha$.

Шаг 5. Значения, полученные на шаге 4, умножаем на нормирующий параметр $\beta$, в данном примере принимаем $\beta = 1$.

Шаг 6. Смещаем полученные значения на шаге 5 на величину параметра сдвига $\gamma = \beta$.

Получен нормированный временной ряд (рисунок 2), предназначенный для сингулярного спектрального анализа.

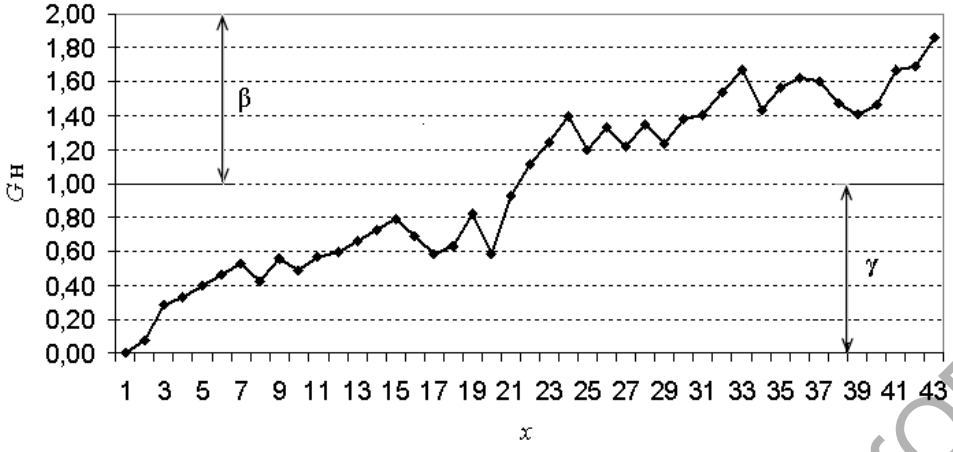


Рисунок 2 – Нормированный временной ряд

Таким образом, в результате проведенных исследований получено аналитическое выражение для нормированного временного ряда, включающее среднее элементов временного ряда, наибольшее по модулю значение элемента ряда, образованного разностью между элементами исходного ряда и средним его элементов, параметр масштаба, определяемый верхней границей нормированного временного ряда, параметр сдвига.

Аналитическое выражение для нормированного временного ряда позволяет независимым исследователям унифицировать результаты реальных исследований и создать базу знаний для выявления полезной информации с учетом особенностей применения метода сингулярного спектрального анализа.

## Заключение

Для практического применения метода извлечения полезной информации из временных последовательностей данных, накопленных в комплексных информационных системах действующих предприятий, основанного на сингулярном спектральном анализе, кластерном анализе и использовании обобщенных законов распределения случайных величин, предложено приводить временной ряд к нормированному виду путем его сжатия/растяжения и смещения, что создает перспективу построения банка знаний по исследованию временных последовательностей данных SSA-методом. Применение SSA-метода в имитационном моделировании на этапе эксплуатации имитационной модели позволит разрабатывать автоматизированные системы извлечения полезной информации.

## Литература

1 Таран, Т.А. Искусственный интеллект. Теория и приложения: учеб. пособие / Т.А. Таран, Д.А. Зубов. – Луганск: ВНУ им. В. Даля, 2006. – 240 с.: ил.

2 Методы, средства и технологии исследования временных последовательностей статистических данных в имитационном моделировании: отчет о НИР (заключ.) / Белорус.-Рос. ун-т; рук. Е.А. Якимов; исполн.: Р.В. Петров [и др.]. – Могилев, 2011. – 126 с. – Библиогр.: с. 124–126. – № ГР 20091957. – Инв. № Ф09М-171.

3 Golyandina, N. Analysis of Time Series Structure: SSA and Related Techniques / N. Golyandina, V. Nekrutkin, A. Zhigljavsky. – Boca Raton: Chapman & Hall/CRC, 2001. – 310 p.

4. Якимов, Е.А. Исследование SSA-метода на основе комплексного применения информационных технологий / Е.А. Якимов // Доклады БГУИР. – 2010. – № 2 (48). – С. 77–83.

5 Гантмахер, Ф.Р. Теория матриц / Ф.Р. Гантмахер. – 2-е изд., доп. – М.: Наука, 1966. – 576 с.

6 Ивановский, Р.И. Компьютерные технологии в науке и образовании. Практика применения систем MathCAD Pro: учеб. пособие / Р.И. Ивановский. – М.: Высш. шк., 2003. – 431 с.: ил.

7 Голяндина, Н.Э. Метод «Гусеница»-SSA: анализ временных рядов: учеб. пособие / Н.Э. Голяндина. – СПб.: С.-Петерб. гос. ун-т, 2004. – 76 с.

8 Якимов, Е.А. Исследование временных рядов с равномерным, нормальным и экспоненциальным шумом с помощью SSA-метода / Е.А. Якимов, В.Г. Замураев, А.И. Якимов // Вестн. Брест. гос. техн. ун-та. Физика, математика, информатика. – 2010. – № 5 (65). – С. 100–104.

9 Якимов, Е.А. О преобразовании методом сингулярного спектрального анализа последовательностей данных с равномерным шумом / Е.А. Якимов, В.Г. Замураев, А.И. Якимов // Информатика. – 2011. – № 1. – С. 52–61.

10 Якимов, Е.А. Особенности преобразования временных рядов методом сингулярного спектрального анализа / Е.А. Якимов, А.А. Ковалевич, Д.М. Албкеират // Вычислительный интеллект (результаты, проблемы, перспективы): материалы 1-й Междунар. науч.-техн. конф. (10–13 мая 2011 г., Черкассы). – Черкассы: Маклаут, 2011. – С. 504–505.

ГУВПО «Белорусско-Российский университет», Могилев

Поступило 08.11.11