ИНФОРМАТИКА

УДК 001:002

ИССЛЕДОВАНИЕ SSA-МЕТОДА НА ОСНОВЕ КОМПЛЕКСНОГО ПРИМЕНЕНИЯ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Е.А. ЯКИМОВ

Белорусско-Российский университет
пр. Мира, 43, Могилев, 212005, Беларусь

Поступила в редакцию 7 декабря 2009

Предложена методика исследования SSA-метода с комплексным применением информационных технологий. Экспериментальные исследования детерминированного временного ряда проведены с использованием табличного процессора MS Excel, математического пакета MathCAD и пакета статистической обработки данных Statistica. Определены количественные характеристики погрешности восстановления временного ряда.

Ключевые слова: SSA-метод, временная последовательность данных, MS Excel, MathCAD, Statistica.

Введение

В процессе интеллектуального анализа данных (ИАД) центральное место занимает автоматическое порождение характеризующих анализируемые данные моделей, правил и/или функциональных зависимостей. В целом процесс извлечения знаний в ИАД условно делят на следующие этапы [1], которые в совокупности предложено использовать на этапе эксплуатации имитационной модели (ИМ) сложного объекта [2].

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.

Для разработки технологии извлечения знаний из временных последовательностей данных исследован сингулярный спектральный метод (SSA-метод), включающий этапы вложения, сингулярного разложения, группировки, диагонального усреднения [3].

Содержательное описание SSA-метода

SSA-метод используется для анализа временных рядов и может быть использован на каждом из этапов эксплуатации ИМ. Он позволяет выделить ряды, которые описывают тренд

первоначального ряда, гармонические колебания и те составляющие рядов, которые относят к "шуму". При этом метод не требует стационарности ряда, знания модели тренда, а также сведений о наличии в ряде периодических составляющих и их периодах. Также с помощью данного метода можно определить модель тренда и использовать это знание для дальнейшей обработки ряда уже с известной моделью тренда, что важно, например, при автоматизации определения длительности переходного процесса в ИМ.

Математической основой SSA-метода является сингулярное разложение [4]. Для успешного применения SSA-метода следует последовательно пройти несколько шагов.

Вложение. На этом шаге выбирается ширина окна L, от выбора которой зависят результаты применения SSA-метода. Из-за того, что нет общих рекомендаций по выбору ширины окна, параметр L зависит от решаемой задачи и предварительной информации, известной о ряде. Например, для выделения тренда рекомендуется выбирать ширину окна не слишком большой. С другой стороны, для выделения гармонических колебаний рекомендуется большая ширина окна. После выбора ширины окна в соответствии с L строится траекторная матрица A ряда, которая будет являться по условию ее построения ганкелевой [5].

Сингулярное разложение. Для матрицы $S=A\cdot A^T$ находятся собственные числа λ и ортонормированные собственные вектора **U**. Упорядоченные по убыванию собственные числа, которые являются большими нуля, часто называются сингулярными числами, а соответствующие им собственные вектора — левыми сингулярными векторами **U**. После этого вычисляются вектора **V**, которые называются правыми сингулярными векторами, и находятся элементарные матрицы, на сумму которых раскладывается первоначальная траекторная матрица.

Группировка. На этапе группировки элементарные матрицы группируются по принципу принадлежности к тренду, гармоническим колебаниям или к шуму. Этот этап является наиболее сложным при применении SSA-метода. Для нахождения тренда на диаграммах собственных векторов (по оси абсцисс откладывается порядковый номер координаты собственного вектора, по оси ординат откладывается значение координаты собственного вектора) выделяют медленно меняющиеся вектора. Сумма элементарных матриц, соответствующих этим векторам, будет являться траекторной матрицей тренда ряда. После этого восстанавливают гармонические колебания ряда. Для отделения шума можно воспользоваться несколькими замечаниями: нерегулярное поведение сингулярных векторов может говорить о принадлежности их к набору, порожденному шумовой компонентой; также об этом может свидетельствовать медленное, практически без скачков, убывание собственных значений с некоторого номера.

Диагональное усреднение. Если полученные сгруппированные матрицы являются ганкелевыми, то они являются траекторными матрицами некоторого ряда, который может быть легко по ним восстановлен. Однако обычно сгруппированные матрицы редко получаются ганкелевыми, поэтому для восстановления ряда прибегают к диагональному усреднению. В соответствии с этим этапом каждый член восстановленного ряда будет являться средним арифметическим соответствующей ему побочной диагонали траекторной матрицы.

В результате проделанных шагов получается несколько рядов, один из которых является рядом, описывающим тренд первоначального ряда, другой описывает гармонические колебания, а третий — шумовые составляющие.

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.

Этап вложения. Для экспериментальных исследований исходный ряд $G=(g_0, g_1, \ldots, g_{N-1})$ может быть задан по известным функциям либо определен результатами функционирования исследуемой системы. По известным функциям ряд 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 $G=(g_0, g_1, \ldots, g_{N-1})$ в последовательность L-мерных векторов, число которых равно $K=N-L+1$:

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

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 $i+j$=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 $\{\lambda_k\}_{k=1}^{L}$ вещественны и неотрицательны. Представленные в виде $\lambda_1 \geq \ldots \geq \lambda_L \geq 0$ собственные числа называют сингулярными значениями матрицы A. Пусть $\mathbf{U}_1, \ldots, \mathbf{U}_L$ — соответствующие им ортонормированные собственные вектора. Будем называть $p=\max\{k \mid \lambda_k>0\}$ порядком сингулярного разложения. Обозначим

$$\mathbf{V}_k = \frac{1}{\sqrt{\lambda_k}} A^T \mathbf{U}_k, \quad k = 1, ..., p. \tag{2}$$

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

$$A = A_1 + A_2 + ... + A_p, \quad A_k = \sqrt{\lambda_k}\, \mathbf{U}_k \mathbf{V}_k^T. \tag{3}$$

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 $<\sqrt{\lambda_k}, \mathbf{U}_k, \mathbf{V}_k>$ будем называть k-й собственной тройкой.

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

d := svds(A). (4)

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

ds := diag(**d**). (5)

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

AS := svd(A). (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 [3]. При этом для одномерного SSA левые и правые сингулярные векторы 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 $I=I_1, \ldots, I_m$, тем самым разложение (3) может быть записано в сгруппированном виде:

$$A = A_{I_1} + \ldots + A_{I_m}. \tag{7}$$

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 относи-

тельно 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 $G^{\text{в}}$) 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 $K=N–L+1$. Пусть $z_{ij}=x_{ij}$, если $L<K$ и $z_{ij}=x_{ji}$ в остальных случаях. Тогда диагональное усреднение переводит матрицу X в ряд $(g^{\text{в}}_0, \ldots, g^{\text{в}}_{N-1})$ по формуле

$$g^{\text{в}}_k = \begin{cases} \frac{1}{k+1}\sum_{j=1}^{k+1} z_{j,k-j+2} & 0 \le k < L^*-1; \\ \frac{1}{L^*}\sum_{j=1}^{L^*} z_{j,k-j+2} & L^*-1 \le k < K^*; \\ \frac{1}{N-k}\sum_{j=k-K^*+2}^{N-K^*+1} z_{j,k-j+2} & K^* \le k < N. \end{cases} \qquad (8)$$

Это выражение соответствует усреднению элементов матрицы вдоль побочных диагоналей $i+j=k+2$: выбор $k=0$ дает $g^{\text{в}}_0=x_{11}$, для $k=1$ получаем $g^{\text{в}}_1=(x_{12}+x_{21})/2$ 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 $G^{\text{в}}$ определяется аналогично формуле (1) с использованием функции СРЗНАЧ() в MS Excel. Затем исследуется в пакете Statistica.

Экспериментальная часть

Для проведения исследований выбрана функция $F(x)=y_1(x)+y_2(x)$, где $y_1(x)$ — линейная функция, $y_1(x)=0{,}1x+2$; $y_2(x)$ — гармоническая функция, $y_2(x)=\sin(x)$. Переменная x принимает значения от 0 до 42 с шагом, равным единице. Таким образом, длина N ряда $G=(g_0, g_1, \ldots, g_{N-1})$ равна 43. При этом длина окна $L=22$, число L-мерных векторов $K=N–L+1=22$.

Отдельно на рабочем листе MS Excel 2003 рассчитаны значения функций $y_1(x)$, $y_2(x)$ и $F(x)$. При этом ряду G_{y1} линейной функции, или тренда, соответствуют значения $G_{y1}=\{g_{y1_0}, g_{y1_1}, ..., g_{y1_{42}}\}=\{2{,}0; 2{,}1; \ldots; 6{,}2\}$. Ряд G_{y2} гармонической составляющей — $G_{y2}=\{g_{y2_0}, g_{y2_1}, ..., g_{y2_{42}}\}=\{0; 0{,}841; \ldots; -0{,}917\}$ и $G=\{g_0, g_1, ..., g_{42}\}=\{2{,}0; 2{,}94; \ldots; 5{,}28\}$.

Элементы ряда G копируются в траекторную матрицу A на рабочем листе MathCAD 14.0. Сформированная матрица A является ганкелевой размером 22×22. Следует отметить, что размеры матрицы при ручном вводе ограничены возможностями MathCAD и не должны превышать 600 позиций.

Сингулярные числа при сингулярном разложении траекторной матрицы A с применением функции svds() по формуле (4) принимают следующие значения: $\lambda_1=99{,}49$; $\lambda_2=10{,}99$; $\lambda_3=10{,}80$; $\lambda_4=2{,}10$; $\lambda_5=0{,}024$. Остальные значения $\lambda_6, \ldots, \lambda_{22}$ представляют убывающую последовательность: $\lambda_6, \ldots, \lambda_{22}=0{,}022; 0{,}018; \ldots, 3{,}06\cdot10^{-3}; 6{,}602\cdot10^{-4}$.

Сумма всех сингулярных чисел равна 116,577, причем сумма собственных чисел $\lambda_1+\lambda_2+\ldots+\lambda_5=116{,}413$.

Сингулярное разложение в пакете MathCAD 14.0 реализуется с помощью функции svd() по формуле (6), формируется обобщенная матрица *AS*, включающая левые и правые сингулярные вектора. Левые и правые сингулярные вектора разделяются по формулам *U*:=submatrix(*AS*, 0,21,0,21) и *V*:=submatrix(*AS*, 22,43,0,21), соответственно.

Диагональная матрица *ds* сингулярных чисел определена по формуле (5). Для контроля правильности выполненных вычислений определяется равенство

$$U \cdot ds \cdot V^T = A. \qquad (9)$$

Для последующих расчетов определяются скалярные значения собственных чисел по формулам с использованием функции submatrix(). Например, для λ_1 (обозначение *ds*0 в пакете MathCAD)

$$ds0 := \text{submatrix}(ds, 0, 0, 0, 0),\ ds0 = 92{,}49, \qquad (10)$$

для λ_2 (обозначение *ds*1 в пакете MathCAD)

$$ds1 := \text{submatrix}(ds, 1, 1, 1, 1),\ ds1 = 10{,}99 \text{ и т. д.} \qquad (11)$$

Левые сингулярные вектора определяются из матрицы *U* аналогично с использованием функции submatrix():

$$\mathbf{U}_0 := \text{submatrix}(U, 0, 21, 0, 0), \qquad (12)$$

$$\mathbf{U}_1 := \text{submatrix}(U, 0, 21, 1, 1) \text{ и т. д.} \qquad (13)$$

Правые сингулярные вектора определяются в пакете MathCAD по формуле

$$\mathbf{V}_i := \frac{1}{\sqrt{ds0}} \text{Ч} A^T \text{Ч} \mathbf{U}_i,\ i = 0, 1, ..., 4. \qquad (14)$$

Элементарные матрицы X_i определены по формуле

$$X_i := \sqrt{ds0} \text{Ч} (\mathbf{U}_i \text{Ч} \mathbf{V}_i^T),\ i = 0, 1, ..., 4. \qquad (15)$$

Для определения групп элементарных матриц X_i построены лепестковые диаграммы для векторов $\mathbf{V}_i$ в табличном процессоре MS Excel 2003 (рис. 1 и 2). Вектора $\mathbf{V}_1$, $\mathbf{V}_2$ и $\mathbf{V}_4$ имеют вид, аналогичный изображению на рис. 2. Векторам $\mathbf{V}_0$ и $\mathbf{V}_3$ на диаграммах соответствуют расходящиеся кривые (рис. 1).

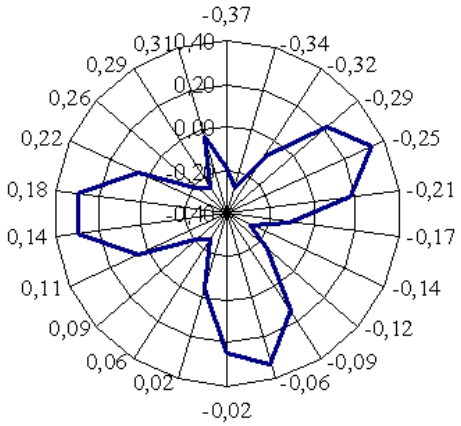

Рис. 1. Лепестковая диаграмма вектора $\mathbf{V}_2$

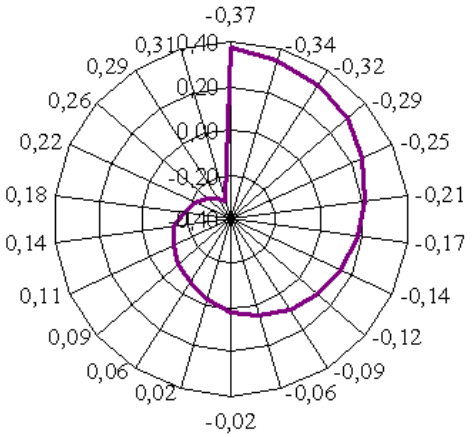

Рис. 2. Лепестковая диаграмма вектора $\mathbf{V}_3$

Тогда группа *X*1 элементарных матриц соответствует тренду ряда: $X1 := X_0 + X_3$. Аналогично группа $X2 := X_1 + X_2 + X_4$ соответствует гармонической составляющей ряда.

На этапе диагонализации для выполнения расчетов по формуле (8) используется специальная процедура в MS Excel, которая иллюстрируется на примере матрицы $X1$ с левым верхним блоком

$$X1^{4\times 4} = \begin{pmatrix} 2{,}141 & 2{,}242 & 2{,}276 & 2{,}303 \\ 2{,}242 & 2{,}345 & 2{,}377 & 2{,}402 \\ 2{,}276 & 2{,}377 & 2{,}407 & 2{,}431 \\ 2{,}303 & 2{,}402 & 2{,}432 & 2{,}453 \end{pmatrix}. \qquad (16)$$

На рис. 3 представлен фрагмент рабочего листа MS Excel с решением задачи диагонализации по предложенной процедуре.

Сдвиговая процедура блока матрицы $X1$ при диагонализации	**2,141**	**2,242**	**2,276**	**2,303**	$x1_{15}$	$x1_{16}$	$x1_{17}$
		2,242	**2,345**	**2,377**	**2,402**	$x1_{25}$	$x1_{26}$
			2,276	**2,377**	**2,407**	**2,431**	$x1_{35}$
				2,303	**2,402**	**2,431**	**2,453**
Восстановленные значения тренда, $g^{в}_{y1_i}$	2,141	2,242	2,299	2,34	…	…	…
Значения тренда, g_{y1_i}	2	2,1	2,2	2,3	…	…	…
Отклонения $g_{y1_i} - g^{в}_{y1_i}$	–0,141	–0,142	–0,099	–0,04	–0,0037	0,069	0,147

Рис. 3. Процедура диагонализации матрицы $X1$ тренда в MS Excel

Восстановленные значения тренда $g^{в}_{y1_i}$ определяются по столбцам с применением функции СРЗНАЧ(). Полученная таким образом матрица используется в пакете Statistica 6.0 для оценки характера распределения ошибки восстановления ряда.

Результаты и их обсуждение

В ходе проведенных исследований временного детерминированного ряда G, образованного функцией вида $F(x) = ax + b\sin x + c = 0{,}1x + \sin x + 2$, построена ганкелева матрица A размером 22×22. При сингулярном спектральном разложении матрицы A основными для восстановления ряда явились элементарные матрицы X_0 и X_3 для линейной составляющей (собственные числа λ_1=99,49 и λ_4=2,10, которым в пакете MathCAD соответствуют обозначения $ds0$ и $ds3$) и матрицы X_1, X_2, X_4 для гармонической составляющей (собственные числа λ_2=10,99; λ_3=10,80; λ_5=0,024, которым в пакете MathCAD соответствуют $ds1$, $ds2$ и $ds4$).

Абсолютная ошибка восстановления ряда определена величиной отклонения

$$\mathrm{D} = \max |g_i - g^{в}_i|. \qquad (17)$$

Относительная ошибка восстановления ряда определяется по формуле

$$\mathrm{d} = \frac{\max |g_i - g^{в}_i|}{\max |g_i|} 100\%. \qquad (18)$$

Абсолютная ошибка восстановления для тренда составляет 0,142, для гармонической составляющей — 0,143, для исходного ряда — 0,0055. Относительная ошибка восстановления тренда составляет 2,29%, для гармонической составляющей — 14,3 %, для исходного ряда — 0,08 %.

Обнаружены интересные факты: абсолютная погрешность восстановления тренда и гармонической составляющей для начальных и конечных значений ряда имеют наибольшие значения (рис. 4), распределения суммы квадратов отклонений значений побочных диагоналей матриц $X1$ и $X2$ имеют характерный вид (рис. 5).

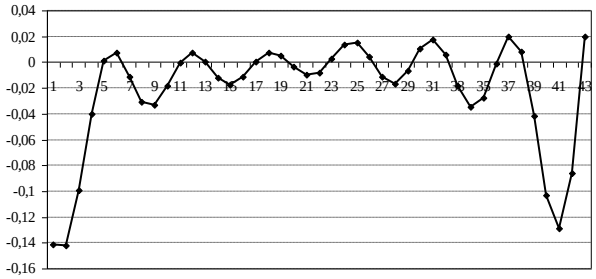

Рис. 4. Распределение абсолютной ошибки восстановления тренда

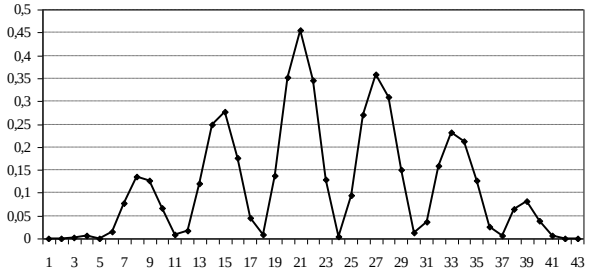

Рис. 5. Распределение суммы квадратов отклонений значений побочных диагоналей $X1$

Для тренда наибольшая ошибка принимает отрицательные значения, для гармонической составляющей — положительные значения. Выявленные особенности оказывают существенное влияние на прогнозирующие свойства SSA-метода.

Заключение

Предложена методика исследования SSA-метода, позволяющая проводить его изучение без разработки специализированного программного приложения. Получены оценки погрешности восстановления временной последовательности данных для детерминированной функции $F(x) = ax + b\sin x + c$.

Работа выполнена при финансовой поддержке гранта Ф09М-171 БРФФИ.

RESEARCH SSA-METHOD ON THE BASIS OF COMPLEX USING INFORMATION TECHNOLOGY

E.A. YAKIMAU

Abstract

The technique of research of a SSA-method with complex application of information technologies is offered. Experimental researches determined time of some are lead with use of tabulared processor MS Excel, mathematical package MathCAD and a package of statistical data processing Statistica. Quantitative characteristics of an error of restoration time are certain of some.

Литература

1. *Таран Т.А., Зубов Д.А.* Искусственный интеллект. Теория и приложения: учеб. пособие. Луганск, 2006.
2. *Якимов Е.А., Максимей И.В.* // Новые материалы, оборудование и технологии в промышленности: материалы междунар. науч.-техн. конф. молод. ученых. Могилев, 19–20 ноября 2009 г. Могилев, 2009. С. 121.
3. *Голяндина Н.Э.* Метод "Гусеница"-SSA: анализ временных рядов: Учебное пособие. СПб., 2004.
4. *Голуб Дж., Ван Лоун Ч.* Матричные вычисления: пер. с англ. М., 1999.
5. *Гантмахер Ф.Р.* Теория матриц. М., 1966.
6. *Ивановский Р.И.* Компьютерные технологии в науке и образовании. Практика применения систем MathCAD Pro: учеб. пособие / Р.И. Ивановский. М., 2003.