МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

# МАТЕРИАЛЫ

## IV Международной молодежной научной конференции

# «МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ИНФОРМАЦИОННЫХ, ТЕХНИЧЕСКИХ И ЭКОНОМИЧЕСКИХ СИСТЕМ»

**Томск, 20–21 мая 2016 г.**

*Под общей редакцией*
*кандидата технических наук И.С. Шмырина*

Томск
Издательский Дом Томского государственного университета
2016

# ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ НА ОСНОВЕ СИНГУЛЯРНОГО СПЕКТРАЛЬНОГО АНАЛИЗА

**А.А. Замщикова, Ж.Н. Зенкова**
*Томский государственный университет*
zamashikova.n@mail.ru, thankoff@fpmk.tsu.ru

## Введение

В современных логистике и маркетинге широко используются различные методы прогнозирования временных рядов. Связанно это с тем, что на основе полученных прогнозов принимаются различные управленческие решения при планировании и оптимизации производства, поставок, запасов, распределения продукции и т.д.

В данной работе рассматривается метод Сингулярного Спектрального Анализа SSA (Singular Spectrum Analysis), также известный как метод «Гусеница». Метод SSA появился в 80-х годах 20 века, он основан на преобразовании одномерного временного ряда в многомерный ряд с последующим применением к полученному многомерному временному ряду метода главных компонент.

Метод SSA позволяет выделять тренд, обнаруживать периодичность, корректировать сезонность, сглаживать, подавлять шум и многое другое, он не требует стационарности ряда, знания модели тренда, а также сведений о наличии в ряде периодических составляющих и их периодах.

## Алгоритм прогнозирования методом «Гусеница»

Пусть имеется исходный вещественный временной ряд $x = (x_i)_{i=\overline{1,N}}$ длины *N*. Предположим, что ряд *x* — ненулевой (т.е. всегда существует хотя бы один номер *i*, для которого $x_i \neq 0$ ) [1].

**Первый этап алгоритма: Разложение**

**Шаг 1: Вложение.** Исходный временной ряд преобразуется в последовательность многомерных векторов. Для этого задается целое число *L* — длина окна или длина «гусеницы», при этом $1 < L < N$ и $L \le \frac{N}{2}$. Далее строятся $K = N - L + 1$ векторов вложения $X_i = (x_i, ..., x_{i+L-1})$, $i = \overline{1, K}$, затем — *L*-траекторная матрица *X*, состоящая из векторов $X_i$ в качестве столбцов:

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_L & \dots & x_K \\ x_2 & x_3 & x_4 & \dots & x_{L+1} & \dots & x_{K+1} \\ x_3 & x_4 & x_5 & \dots & x_{L+2} & \dots & x_{K+2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_L & x_{L+1} & x_{L+2} & \dots & x_{2L-1} & \dots & x_N \end{pmatrix}, \quad (1)$$

где $K = N - L + 1$. Полученная матрица является ганкелевой.

**Шаг 2: Сингулярное разложение.** Вычисляется матрица

$$S = \frac{1}{K} X \cdot X^{\mathrm{T}}, \quad (2)$$

затем рассматриваются $\lambda_1, \lambda_2, ..., \lambda_L$ — собственные числа матрицы *S*, $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_L$, и $U_1, U_2, ..., U_L$ — ортонормированная система собственных векторов матрицы *S*, соответствующих этим собственным числам, при этом

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_L \end{pmatrix},\ U = \left(U^1, U^2, ..., U^L\right) = \begin{pmatrix} u_1^1 & u_1^2 & ... & u_1^L \\ u_2^1 & u_2^2 & ... & u_2^L \\ ... & ... & ... & ... \\ u_L^1 & u_L^2 & ... & u_L^L \end{pmatrix}.$$

Тогда матрица *S* может быть представлена в виде:

$$S = U\Lambda U^{\mathrm{T}}. \tag{3}$$

**Шаг 3: Переход к главным компонентам.** Здесь матрица *U* представляется как матрица перехода к главным компонентам исходного ряда:

$$G = U^{\mathrm{T}} X = \left(G_1, G_2, .., G_L\right), \tag{4}$$

где величины $G_i = \left(U^i\right)^{\mathrm{T}} X$, $i = 1, 2, ..., L$, называются главными компонентами исходного вектора.

**Второй этап: Восстановление ряда по выбранным главным компонентам**

Зная главные компоненты, можно полностью восстановить траекторную матрицу следующим образом:

$$X = \left(U^1 \quad U^2 \quad .. \quad U^L\right) \cdot \begin{pmatrix} G_1 \\ G_2 \\ ... \\ G_L \end{pmatrix} = \sum_{i=1}^{L} U^i G_i .$$

При этом, зная траекторную матрицу, можно всегда перейти к искомому временному ряду $x = (x_i)_{i=\overline{1,N}}$.

Рассмотрим первые *r* из *L* компонент собственных векторов. Введем обозначения:

$$\tilde{U} = \left(U^1 \quad U^2 \quad .. \quad U^L\right),\ \tilde{G} = \tilde{U}^{\mathrm{T}} X .$$

Тогда восстановленная траекторная матрица определяется формулой:

$$\tilde{X} = \tilde{U}^{\mathrm{T}} \tilde{G} = \sum_{i=1}^{r} U^i G_i . \tag{5}$$

Она имеет размерность $L \times K$. Матрица $\tilde{X}$ позволяет получить восстановленный исходный временной ряд

$$\tilde{x}_s = \begin{cases} \dfrac{1}{s} \displaystyle\sum_{i=1}^{s} \tilde{X}_{s,s-i+1},\ 1 \le s \le L, \\ \dfrac{1}{L} \displaystyle\sum_{i=1}^{L} \tilde{X}_{i,s-i+1},\ L \le s \le K, \\ \dfrac{1}{N-s+1} \displaystyle\sum_{i=1}^{N-s+1} \tilde{X}_{i+s-K,K-i+1},\ K \le s \le N, \end{cases} . \tag{6}$$

который называется *SSA-сглаживанием исходного временного ряда* $(x_i)_{i=\overline{1,N}}$ по первым *r* компонентам из *L*.

**Третий этап: Прогнозирование**

На этом этапе нам необходимо построить числовой ряд $x = (x_i)_{i=\overline{N+1,M}}$, который называется прогнозом исходного временного ряда $x = (x_i)_{i=\overline{1,N}}$ на *M* шагов вперёд. Для этого рассмотрим некоторый набор главных компонент $G_i$, $i = 1, 2, ..., L$, обозначим:

$$U_* = \begin{pmatrix} u_1^1 & u_1^2 & \dots & u_1^r \\ u_2^1 & u_2^2 & \dots & u_2^r \\ \dots & \dots & \dots & \dots \\ u_{L-1}^1 & u_{L-1}^2 & \dots & u_{L-1}^r \end{pmatrix},\ Q = \begin{pmatrix} X_{N-L+2} \\ X_{N-L+3} \\ \dots \\ X_N \end{pmatrix},\ \tilde{u} = \left(u_L^1, u_L^2, \dots, u_L^r\right).$$

Тогда прогноз определяется как

$$x_{N+1} = \tilde{u}\left(U_*^{\mathrm{T}} U_*\right)^{-1} U_*^{\mathrm{T}} Q. \qquad (7)$$

Прогнозирование следующих значений может быть реализовано двумя способами. В основе первого способа лежит изменение соответствующим образом матрицы $Q$. Второй способ заключается в полном или частичном повторе алгоритма метода SSA. В этом случае, помимо матрицы $Q$, будут также меняться матрицы $\tilde{u}$ и $U_*$.

**Прогнозирование стоимости цен на нефть методом SSA**

Метод SSA был опробован для прогнозирования цен на нефть за период 03.04.2014 – 12.05.2016. Прогноз строился на период 03.05.2016 – 12.05.2016, итоговые прогнозные значения сравнивались с реальными.

При выборе входных параметров использовались два подхода.

1) Несмотря на то, что строгих правил при выборе длины гусеницы $L$ нет, существуют некоторые рекомендации [3], например, если взять $L$ слишком маленькой, то произойдёт смешивание содержательных частей временного ряда. При этом мы знаем, что $L$ не может быть больше, чем $N/2$ (так как SVD-разложения ряда размерности $N$ при выборе длины гусеницы $L$ и $N-L+1$ эквивалентны), и чем больше $L$, тем более подробным будет разбиение ряда. Из этих соображений и в зависимости от поведения графика значений реального и сглаженного ряда была выбрана длина гусеницы $L = 34$.

При выборе $r$ рассматривался тот факт, что собственные числа $\lambda_1, \lambda_2, \dots, \lambda_L$ играют роль доли информационного содержания главных компонент от общего информационного содержания исходного временного ряда. По рис. 1 и 2 можно увидеть, что содержательный смысл несут, в основном, первые две главные компоненты, а оставшиеся, вероятно, шум. Следовательно, зафиксируем $r = 2$.

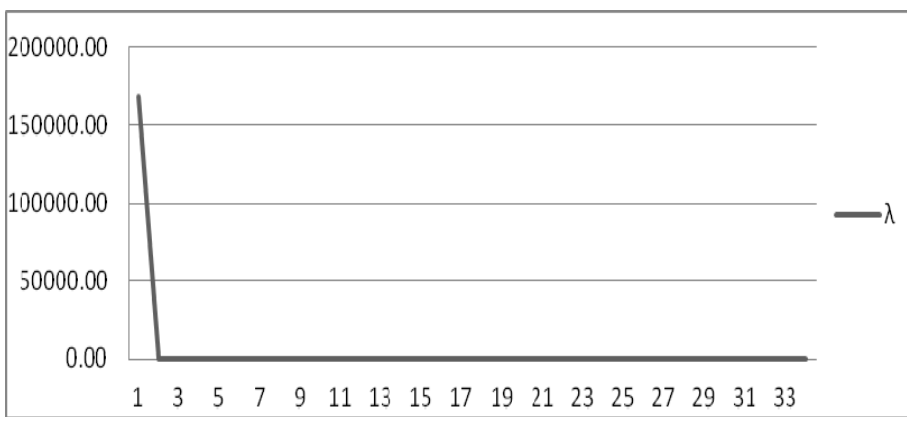



Рис. 1. График первых 34 собственных чисел ковариационной матрицы

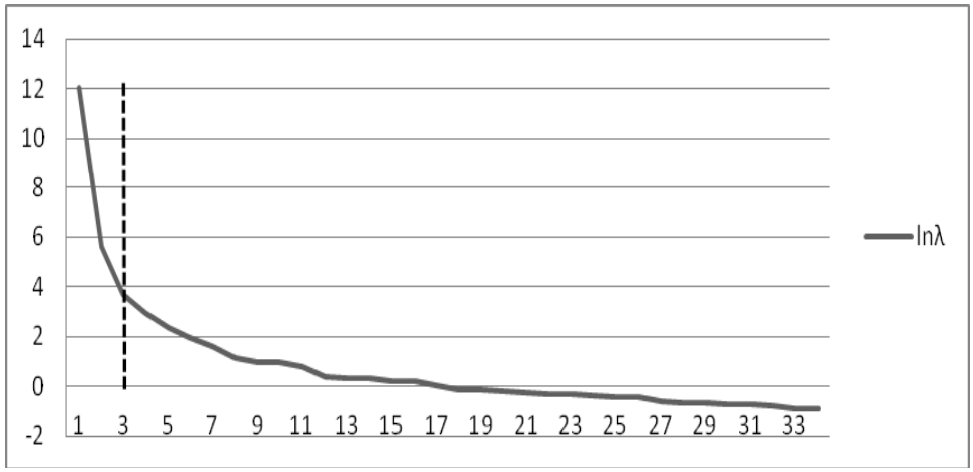



Рис. 2. График логарифма первых 34 собственных чисел ковариационной матрицы

2) Второй подход был выбран из соображений минимума отклонения между модельными и реальными значениями. Была разработана программа, которая подбирает такие пары $r$ и $L$ на каждом шаге прогнозирования, которые удовлетворяют вышеуказанному правилу. При этом ошибка прогнозирования, считается по формуле:

$$\varepsilon = \frac{1}{M}\sum_i \left| \frac{x_i - xx_i}{x_i} \right| \times 100\%, \qquad (7)$$

где $M$ — число шагов прогнозирования, $x_i$ — реальные цены, а $xx_i$ — прогнозные значения. Результаты моделирования при оптимальных параметрах представлены на рис. 3.




Рис. 3. Значение реальных и модельных цен на нефть за период 03.04.2014 – 12.05.2016, в долларах за баррель




Рис. 4. Значение реальных и модельных цен на нефть за период 01.04.2016 – 01.05.2016, в долларах за баррель




Рис. 5. Значение реальных и спрогнозированных цен на нефть за период 03.05.201 – 12.05.2016, в долларах за баррель

Заметим, что ошибка прогнозирования при первом подходе составила $\varepsilon_1 = 4.629\%$, что меньше 5%, следовательно, можно говорить о высоком качестве модели, однако при использовании оптимальных параметров $L$ и $r$ ошибка составила $\varepsilon_2 = 0.162\%$, что в 28.7 раз меньше $\varepsilon_1$. Следовательно, при использовании метода гусеницы имеет смыл предварительно оптимизировать параметры $L$ и $r$ в смысле минимизации отклонений реальных и модельных данных.

## Заключение

В данной работе был описан метод SSA «Гусеница», алгоритм применялся к реальным данным о значениях цен на нефть за период 03.04.2014 – 12.05.2016. Было реализовано сглаживание и прогнозирование временного ряда, произведено исследование зависимости качества прогнозирования от выбора входных параметров. В результате выбран набор параметров алгоритма, дающий очень маленькую ошибку прогнозирования. Процесс подбора параметров был полностью автоматизирован.

## ЛИТЕРАТУРА

1. *Голяндина Е.Э.* Метод «Гусеница»-SSA: прогноз временных рядов: Учеб. пособие. – СПб., 2004. – 52 с.

2. *Солнцев В.Н., Данилов Д.Л., Жиглявский А.А.* Главные компоненты временных рядов: метод “Гусеница”, С.-Петербургский государственный университет, 1997. 307 с. http://www.gistatgroup.com/gus/

3. *Леонтьева Л.Н.* Многомерная гусеница, выбор длины и числа компонент // Машинное обучение и анализ данных. – 2011. – № 1. – С. 2–10.

4. Нефть Brent — котировки за прошедший период [Электронный ресурс] – Электрон. дан. — URL: http://ru.investing.com/commodities/brent-oil-historical-data

# ВЛИЯНИЕ ВИДА РАСПРЕДЕЛЕНИЯ СТОИМОСТИ ОБОРОТНЫХ СРЕДСТВ НА ПОКАЗАТЕЛИ ОБОРАЧИВАЕМОСТИ ПО ИНТЕРВАЛЬНЫМ ДАННЫМ

**Ж.Н. Зенкова, О.Б. Макеева**

*Томский государственный университет*

oksi-mak-tsu@mail.ru, thankoff@fpmk.tsu.ru

В работе исследуется проблема влияния вида распределения стоимости вложений в оборотные средства предприятия на коэффициент оборачиваемости и оборот. Задача обобщается на случай интервально-цензурированных данных о размерах вложений. Рассматривается случай с использованием дополнительной информации о квантиле функции распределения. С помощью имитационного моделирования показано, что учёт информации о квантиле позволяет улучшить качество рассматриваемых статистических процедур в смысле снижения среднеквадратического отклонения и восполнить потери информации, возникающие при цензурировании, независимо от вида распределения.

## Введение

На сегодняшний день одной из актуальных проблем при оценке эффективности управления предприятием является анализ и прогноз стоимости его оборотных активов [1-4]. Их эффективное использование зависит от того, насколько точно определена потребность в оборотных средствах, при этом и в случае занижения, и в случае завышения величины вложений в оборотный капитал возникают риски, чреватые отрицательными последствиями для организации. Однако специфика исходных данных изначально не позволяет получить адекватную оценку среднего размера вложений, так как фактически аналитики работают со значениями моментного ряда, отображающими текущую стоимость активов, которая может существенно меняться даже в течение одного дня, при этом подобные колебания могут не фиксироваться, а в традиционных методиках расчетов никак не учитываются. В работах [3,4] были предложены новые подходы к оцениванию показателей оборачиваемости, когда исходная выборка рассматривалась как цензурированная, в [4-6] исследовался вопрос привлечения дополнительной информации при оценивании стоимости вложений. Заметим, что привлечение дополнительной информации приводит к существенному улучшению качества статистических процедур.

В данной работе с помощью имитационного моделирования исследовано влияние вида распределения стоимости вложений в оборотные средства на показатели оборачиваемости по интервальным данным с учётом априорной информации о квантиле функции распределения (ф.р.).

## 1. Непараметрический алгоритм Тёрнбулла и его модификация с учётом известного квантиля в оценивании показателей оборачиваемости

Пусть стоимость оборотных средств предприятия $\tau > 0$ есть случайная величина (с.в.) с ф.р. $F(t)$. Исходная выборка объёма $N$ состоит из интервалов $(L_i, R_i]$, $i = \overline{1, N}$, т.е. известно, что $i$-ое наблюдение попало в промежуток $(L_i, R_i]$, но его точное значение не известно. Для оценивания $F(t)$ рассмотрим разбиение $0 < \tau_0 < \tau_1 < \ldots < \tau_m$, состоя-