Многомерная гусеница, выбор длины и числа компонент*

Л. Н. Леонтьева
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

В работе описывается метод гусеницы (SSA) и его применение для прогнозирования временных рядов. Алгоритм основан на выделении из изучаемого временного ряда некоторого набора его главных компонент и последующего построения прогноза по выбранному набору. Исследуется зависимость точности прогноза от выбора длины гусеницы и числа ее компонент. В вычислительном эксперименте приводятся результаты работы алгоритма на периодических рядах с разным рисунком внутри периода, на рядах с нарушением периодичности, а так же на реальных рядах почасовой температуры в Москве.
Ключевые слова: *прогнозирование, singular spectrum analysis, сингулярное разложение.*

Введение

Данная работа посвящена методу анализа и прогноза временных рядов “Гусеница”, в зарубежной литературе этот метод также называется Singular Spectrum Analisys (SSA). Данному методу посвящены книги [1, 2, 3, 4]. Мы использовали книгу [1] как основной источник сведений о методе гусеницы и придерживались используемых там обозначений и понятий. Одной из основных задач данной работы является исследование зависимости качества прогноза, построенного с помощью метода гусеницы, от длины гусеницы и числа ее компонент [3, 5]. В конце раздела 3 сделаны некоторые теоретические выводы по этому вопросу, а в разделе 4 представлены результаты работы алгоритма на данных почасовой температуры в зависимости от выбора параметров. Метод гусеницы применяется для решения довольно широкого круга задач [8] таких как: разбиение ряда на интерпретируемые составляющие [6], подавление шумов и сглаживание, заполнение пропущенных значений в данных [7] и многих других задач.

Постановка задачи

Дан временной ряд $T = \{x_i\}_{i=1}^{n}$, где n — длина временного ряда, i — номер отсчета. Требуется, задав параметр l, $1 < l < n$ (длину гусеницы) разложить ряд в сумму компонент (используя метод главных компонент), выбрать часть из них и построить по ним продолжение ряда $\{x_i\}_{i=1}^{n+\tau}$.

Предполагаем, что в рассматриваемом временном ряду нет пропущенных значений и он имеет периодическую составляющую с периодом τ, на который и производится прогноз.

Для контроля качества алгоритма прогноза разбиваем множество индексов $N = 1, \ldots, n$ на два подмножества $J_1 = n - \tau + 1, \ldots, n$ и $J_2 = 1, \ldots, n - \tau$. Выделяем во временном ряде T последовательных значений $\{x_i | i \in J_1\}$ (контрольную выборку), которые с помощью алгоритма прогнозируем по предыдущим значениям $\{x_i | i \in J_2\}$. В качестве критерия качества прогноза использовались два функционала: SSE (сумма квадратов ошибок) и MAPE (средняя абсолютная процентная ошибка):

$$SSE = \sum_{i=1}^{\tau} |\tilde{x}_i - x_i|^2,$$

Научный руководитель В. В. Стрижов

$$MAPE = \frac{1}{\tau}\sum_{i=1}^{\tau} 100\frac{|\tilde{x}_i - x_i|}{|x_i|},$$

где $\tilde{x}_i$ — спрогнозированное значение в точке i, x_i — фактическое значение в точке i.

Описание алгоритма

Анализ временного ряда

Для последующего разложение ряда по главным компонентам преобразуем ряд в траекторную матрицу X, которую строим следующим образом:

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ x_2 & x_3 & \dots & x_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_l & x_{l+1} & \dots & x_n \end{pmatrix}. \tag{1}$$

где $k = n - l + 1$, k — время жизни гусеницы. Матрицу (1) будем называть нецентрированной траекторной матрицей, порожденной гусеницей длины l.

Замечание 1. Проводимый в дальнейшем анализ главных компонент может проводиться как по центрированной, так и по нецентрированной выборкам. Для упрощения выкладок рассмотрим простейший нецентрированный вариант.

Построим ковариационную матрицу следующим образом:

$$C = \frac{1}{k}X^T X.$$

Выполним её сингулярное разложение:

$$C = V\Lambda V^T,$$

где $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_l)$ — диагональная матрица собственных чисел, $V = [v^{(1)}, \dots, v^{(l)}]$ — ортогональная матрица собственных векторов. При этом будем предполагать, что собственные векторы упорядочены по убыванию соответствующих собственных чисел, т. е. $\lambda_1 > \lambda_2 > \dots > \lambda_l$. Вычислим матрицу U нецентрированных главных компонент:

$$U = V^T X = (U_1, \dots, U_l)^T.$$

Восстановим траекторную матрицу по некоторому поднабору главных компонент, т. е. для $\tilde{V} = [v^{(i_1)}, \dots, v^{(i_r)}]$, $r \leqslant l$ и $\tilde{U} = \tilde{V}^T X$ вычисляется матрица $\tilde{X} = \tilde{V}\tilde{U}$.

После восстановления матрицы $\tilde{X}$ исходная последовательность восстанавливается усреднением по побочным диагоналям матрицы $\tilde{X}$:

$$\tilde{x}_s = \begin{cases} \frac{1}{s}\sum\limits_{j=1}^{s} \tilde{x}_{j,s-j+1} & 1 \leqslant s \leqslant l, \\ \frac{1}{l}\sum\limits_{j=1}^{l} \tilde{x}_{j,s-j+1} & l \leqslant s \leqslant k, \\ \frac{1}{n-s+1}\sum\limits_{j=1}^{n-s+1} \tilde{x}_{j+s-k,k-j+1} & k \leqslant s \leqslant n. \end{cases}$$

С геометрической точки зрения операция получения главных компонент есть изображение исходной выборки в базисе, составленном из выбранных собственных векторов, а восстановление — проектирование исходной выборки на гиперплоскость, порожденную выбранным набором собственных векторов ковариационной матрицы.

Прогноз временного ряда Перейдем к прогнозированию временных рядов методом гусеницы. Для начала определимся с тем, что мы будем понимать под продолжением ряда.

Определение 1. *Числовой ряд $\{x_i\}_{i=1}^{n+1}$ называется продолжением ряда $\{x_i\}_{i=1}^{n}$, если порождаемая им при гусеничной обработке выборка лежит в той же гиперплоскости, что и у исходного ряда.*

Рассмотрим систему уравнений:

$$\begin{cases} \sum\limits_{j=1}^{s} h_j v_1^j &= x_{n-l+1}, \\ \qquad\qquad \dots & \\ \sum\limits_{j=1}^{s} h_j v_{l-1}^j &= x_n. \end{cases} \tag{2}$$

Введем следующие обозначения: $\mathbf{v} = (v_l^{(i_1)}, v_l^{(i_2)}, \dots, v_l^{(i_r)})$, где $0 < i_1 < \dots < i_r < l$, и

$$V^* = \begin{pmatrix} v_1^{(i_1)} & v_1^{(i_2)} & \dots & v_1^{(i_r)} \\ v_2^{(i_1)} & v_2^{(i_2)} & \dots & v_2^{(i_r)} \\ \vdots & \vdots & \ddots & \vdots \\ v_{l-1}^{(i_1)} & v_{l-1}^{(i_2)} & \dots & v_{l-1}^{(i_r)} \end{pmatrix}.$$

Заметим, что

$$\tilde{V} = \begin{pmatrix} V^* \\ \mathbf{v} \end{pmatrix}.$$

Также пусть $Q = (q_i)_{i=1}^{l-1} = (x_{n-l+2}, \dots, x_n)^T$ и $\bar{h} = (h_1, \dots, h_r)^T$. В этих обозначениях система (2) запишется как

$$V^* \bar{h} = Q. \tag{3}$$

Определение 2. *Обобщенным решением системы (3) назовем решение системы*

$$(V^*)^T V^* \bar{h} = (V^*)^T Q.$$

Определение 3. *Величину*

$$b = \mathbf{v}\bar{h}^*,$$

где $\bar{h}^*$ — решение системы (3), назовем обобщенным продолжением рассматриваемого ряда.

Учитывая (3), можно записать для прогнозируемого значения x_{n+1} следующую формулу:

$$x_{n+1} = \mathbf{v}((V^*)^T V^*)^{-1} (V^*)^T Q. \tag{4}$$

Выбор параметров В этом разделе мы обсудим роль параметров базового метода SSA и принципа их выбора. В базовом методе SSA есть два параметра. Первый — это

целое число l, длина гусеницы, а второй параметр является структурным — это способ группировки главных компонент.

Дадим несколько рекомендаций по выбору длины гусеницы:

— Сингулярные разложения одного и того же ряда длины n, соответствующие выбору длины гусеницы l и $n - l + 1$ эквивалентны. Следовательно, для анализа структуры временного ряда не имеет смысла брать длину гусеницы, большую чем половина длины ряда.
— Чем больше длина гусеницы, тем более детальным получается разложение исходного ряда. Таким образом, наиболее детальное разложение достигается при выборе длины гусеницы, приблизительно равной половине длины ряда ($l \sim n/2$). Причем, чем больше длина гусеницы, тем более детальным получается разложение исходного ряда.
— Маленькая длина гусеницы может привести к смешиванию интерпретируемых компонент ряда.
— При решении задачи выделения периодической компоненты с периодом τ следует выбирать длину гусеницы l кратной τ.
— В общем метод гусеницы устойчив относительно изменения длины гусеницы. Эффект проявляется не столько в количественном, сколько в качественном смысле.

Теперь обсудим важные моменты связанные с отбором главных компонент. Пусть длина гусеницы l фиксирована и мы уже имеем сингулярное разложение траекторной матрицы исходного ряда. Тогда следующим шагом является группировка членов сингулярного разложения:

— Если мы восстановили компоненту ряда только с помощью одной собственной тройки (собственное значение, собственный вектор и главная компонента) и оба сингулярных вектора имеют похожую форму, то восстановленная компонента будет иметь примерно такую же форму. Это правило означает, что, имея дело с единственной собственной тройкой, часто можно предсказать поведение соответствующей компоненты временного ряда. Например, если оба сингулярных вектора собственной тройки похожи на линейные ряды, то соответствующая составляющая ряда также будет близкой к линейной. Если сингулярные векторы имеют экспоненциальную форму, то и компонента ряда будет такой же. Монотонные сингулярные векторы соответствуют монотонной компоненте ряда. Синусоидальные векторы порождают гармоническую составляющую ряда.
— Чем больше собственное значение главной компоненты, тем больше вклад соответствующей восстановленной компоненты ряда.

Вычислительный эксперимент

Сперва протестируем работу алгоритма на простых периодических рядах. Строим прогноз на период для зашумленного синуса с периодом 50. Даже для длины гусеницы не кратной периоду ряда результат работы алгоритма хороший. На рис. 1 показан прогноз, сделанный по двум компонентам и длине гусеницы равной 80.

А теперь мы возьмем только одну первую главную компоненту, а длину гусеницы оставим прежней. Мы существенно не угадали период и длина гусеницы ему не кратна. Первая ГК не отражает основной период, а пытается отразить тренд, которого нет.

Теперь вместо зашумления внесем в тот же синус с периодом 50 два сильных выброса и спрогнозируем его также по двум компонентам с длинной гусеницы 80. Результат работы показан на рис. 3, из которого можно сделать вывод, что алгоритм в некоторой степени устойчив к выбросам. Однако при увеличении числа компонент регрессионные остатки

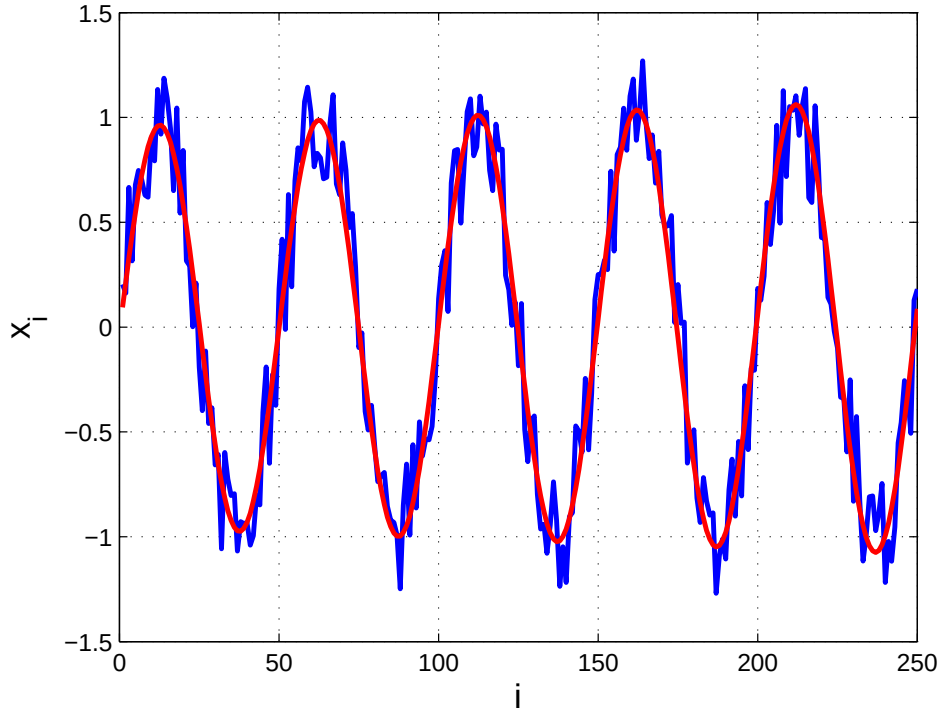

Рис. 1: Прогнозирование зашумленного синуса по двум первым компонентам

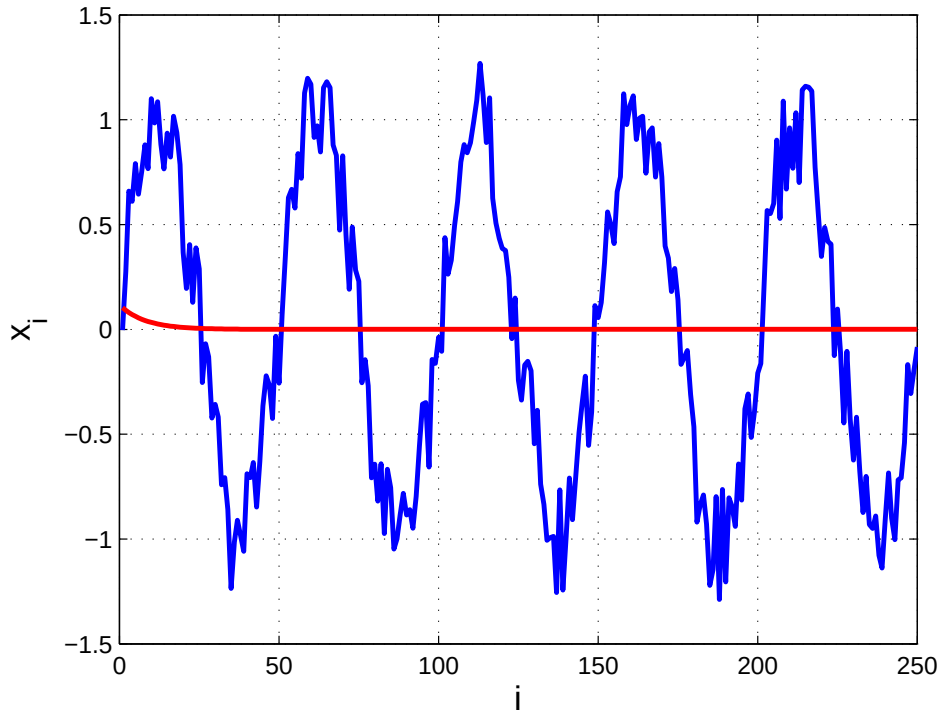

Рис. 2: Прогнозирование зашумленного синуса по первой компоненте

значительно увеличиваются. На рис. 4 показан результат работы алгоритма на том же ряде с той же длинной гусеницы, но число компонент выбрано равное 3.

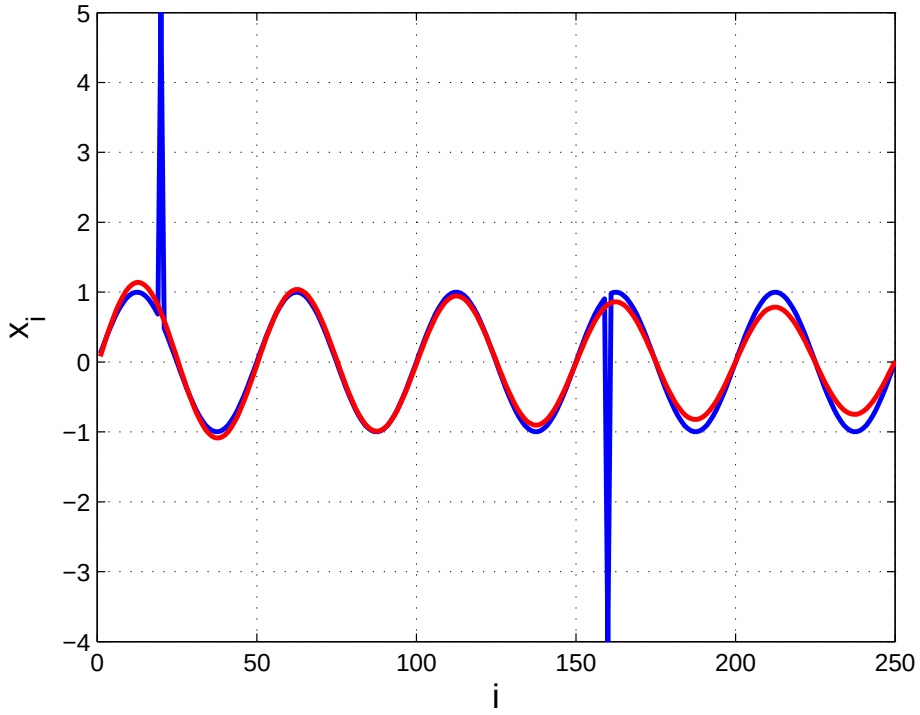

Рис. 3: Прогнозирование синуса с двумя сильными выбросами по двум компонентам

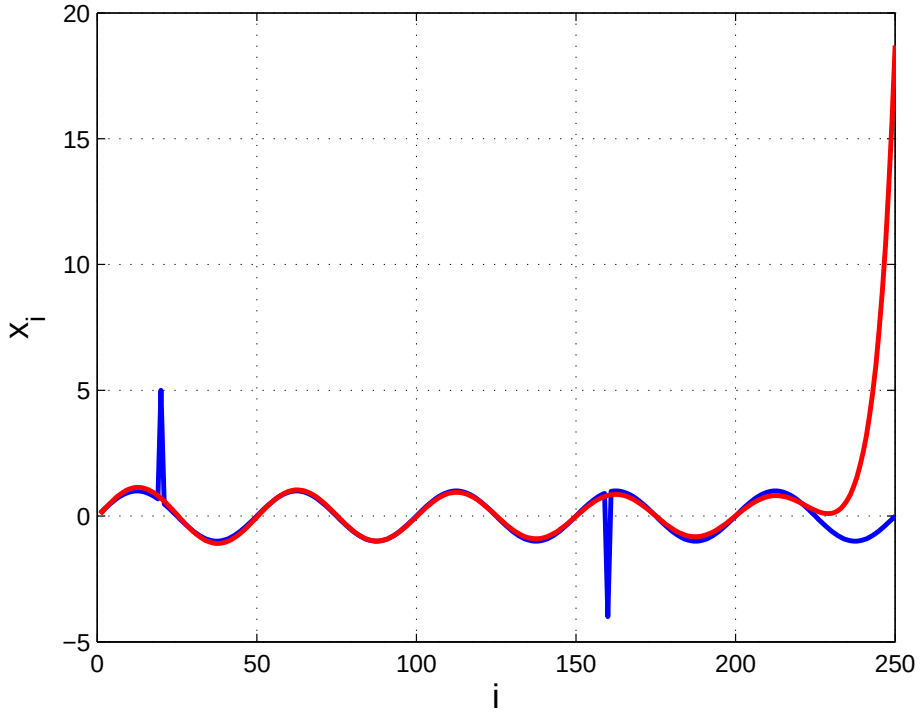

Рис. 4: Прогнозирование синуса с двумя сильными выбросами по трем компонентам

Далее рассмотри ряд являющийся суммой двух периодических зашумленных рядов с периодами 57 и 70. То есть суммарный ряд имеет два периода, которые сильно отличаются

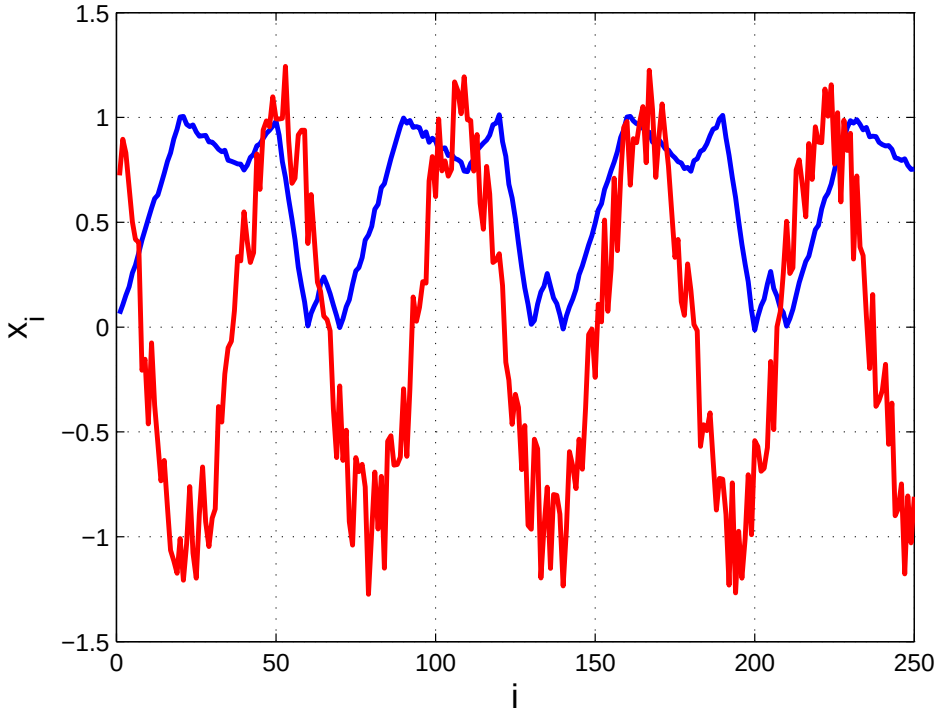

Рис. 5: Два зашумленных несинфазных ряда с разными периодоми

по длине, не синфазны и не кратны. Чтобы их корректно восстановить мы взяли первые четыре компоненты и длину гусеницы равную 80.

Для проведения следующего эксперимента были использованы данные почасовой погоды в Москве в апреле 2011 года. Мы прогнозируем погоду на ближайший час по трем предыдущим дням. Как видно ошибка на обучении мало зависит от длины гусеницы, но резко уменьшается при увеличении числа компонент.

Наименьшая ошибка наблюдается на длине гусеницы от 22 до 26 и числе компонент от 13 до 18. При числе компонент порядка длинны гусеницы (то есть при почти полном наборе компонент) ошибка резко возрастает, тем самым иллюстрируется переобучение нашего алгоритма. Чем меньше длина гусеницы, тем раньше (при меньшем числе компонент) начинается переобучение. Таким образом, MAPE также как и SSE на обучении мало зависит от длины гусеницы и резко уменьшается при увеличении числа компонент.

Из рис. 10 можно заключить, что MAPE почти всюду, кроме областей переобучения, принимает значения меньше 10%. За счет переобучения при выборе числа компонент близкого по величине к длине гусеницы значение MAPE достигает $40 - 50\%$.

Заключение

В данной работе была исследована зависимость SSE и MAPE при прогнозировании временных рядов методом “Гусеница” в зависимости от значений входных параметров, то есть длины и числа компонент гусеницы. Результаты вычислительного эксперимента показали, что для минимизации SSE необходимо выбирать длину гусеницы кратной (или почти

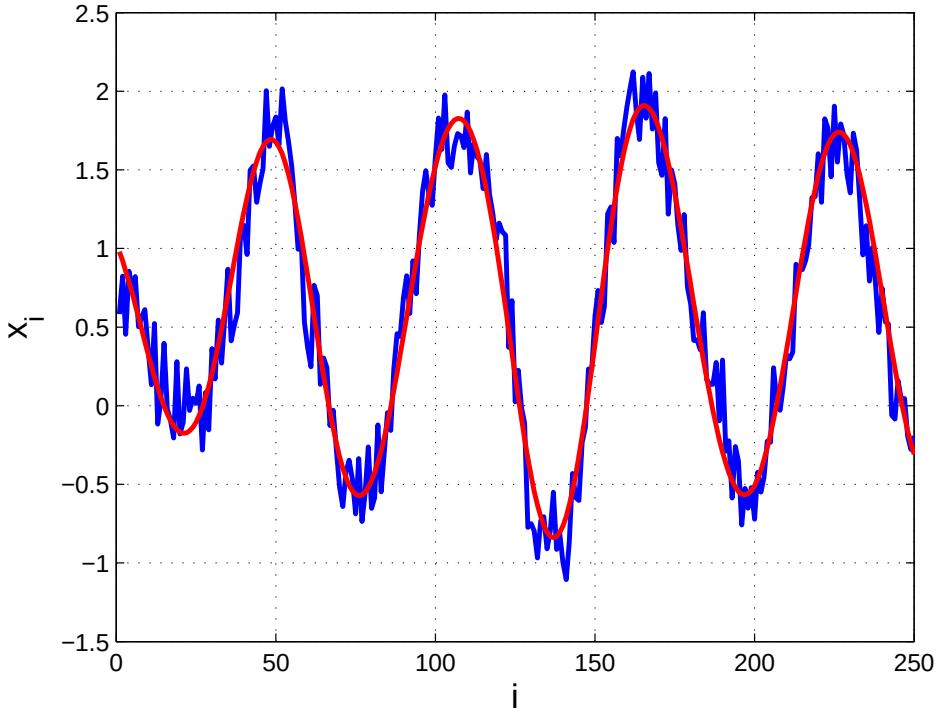

Рис. 6: Прогнозирование суммы двух периодических рядов

кратной) периоду ряда. Переобучение, связанное с большим числом компонент по которым строится прогноз, наступает (для рядов значений почасовой температуры) примерно на 10 компонентах. Помимо этого была исследована эффективность работы алгоритма на зашумленных рядах и рядах с выбросами. Зашумленные ряды метод гусеницы сглаживает и строит хороший прогноз. При наличии в рядах выбросов метод неустойчив к выбору числа компонент.

Код, необходимый для повторения вычислительного эксперимента расположен на сайте: `https://mlalgorithms.svn.sourceforge.net/svnroot/mlalgorithms/GaterpillarLearning/`

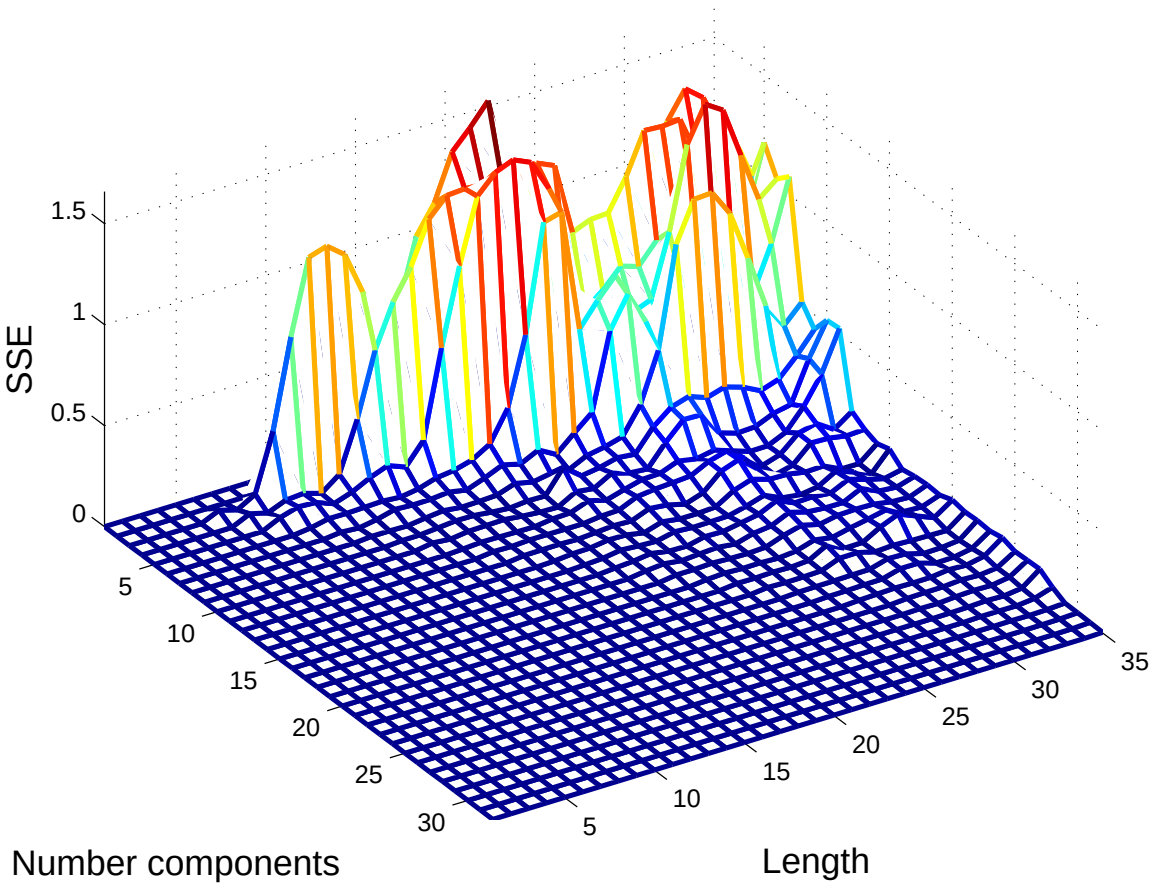

Рис. 7: График зависимости SSE от длины и числа компонент на обучении

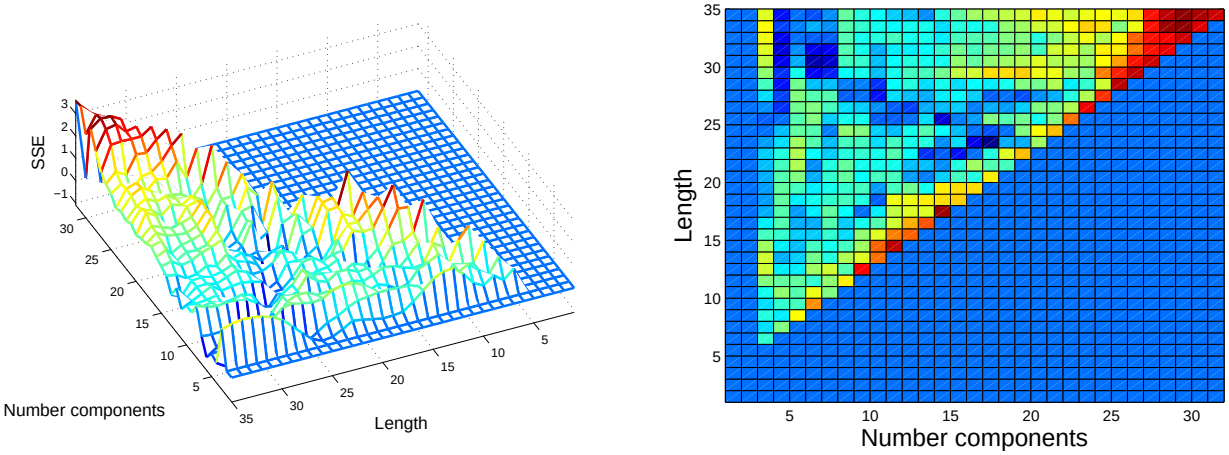

Рис. 8: График зависимости SSE от длины и числа компонент на прогнозе в логарифмическом масштабе

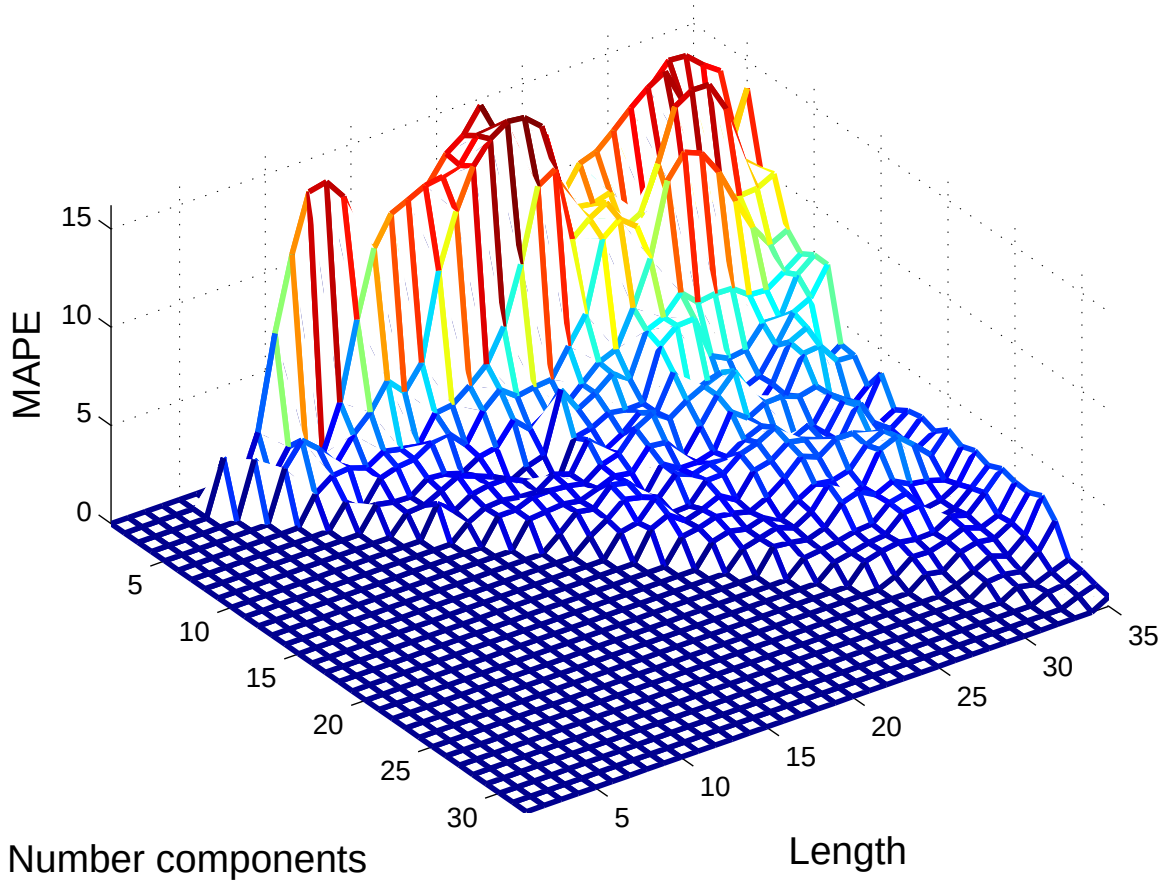

Рис. 9: График зависимости MAPE от длины и числа компонент на обучении

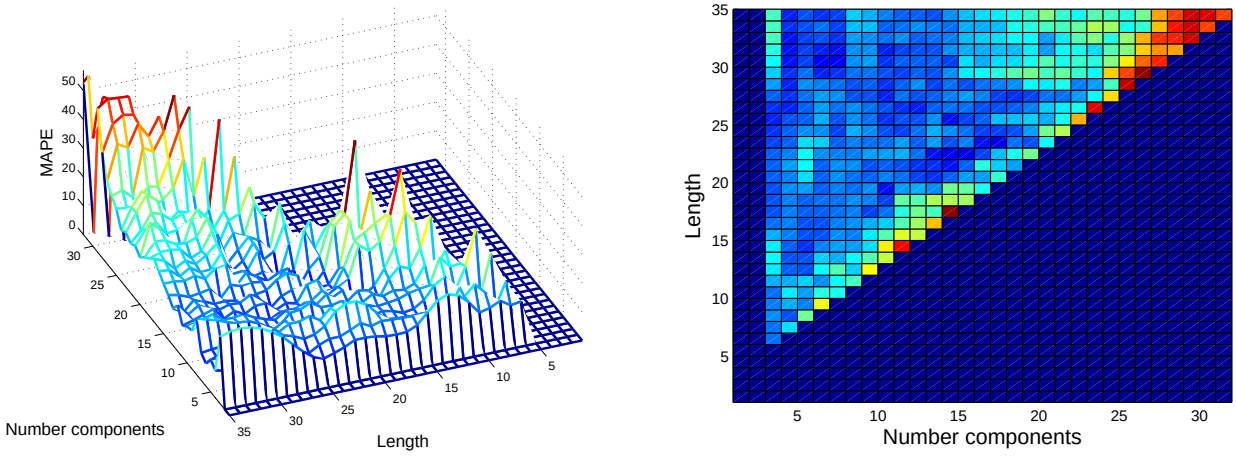

Рис. 10: График зависимости MAPE от длины и числа компонент на прогнозе

Литература

[1] В. Н.Солнцев, Д. Л. Данилов, А. А. Жиглявский. *Главные Компоненты Временных Рядов: Метод "Гусеница"*, С.-Петербургский государственный университет, 1997.

[2] Н. Э. Голяндина. *Метод "Гусеница"-SSA: анализ временных рядов*, С.-Петербургский государсвенный университет, 2004.

[3] Н. Э. Голяндина. *Метод "Гусеница"-SSA: прогноз временных рядов*, С.-Петербургский государсвенный университет, 2004.

[4] N. Golyadina, V. Nekrutkin. *Analysis of Time Series Structure: SSA and Related Techniques*, Chapman&Hall/CRC, 2001.

[5] Nina Golyandina. *On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods*, Statistics and Its Interface, 2010, №3: 259-279.

[6] Ф. И. Александров. *Выделение аддитивных компанент временного ряда на основе метода "Гусеница"*, С.-Петербургский государсвенный университет, 2003.

[7] Н. Э. Галядина, Е. В. Осипов. *Метод "Гусеница"-SSA для анализа временных рядов с пропусками, С.-Петербургский государсвенный университет.*

[8] Hossein Hassani. *Singular Spectrum Analysis: A Relatively New and Powerful Technique for Time series Analysis and Forecasting*, The 31st Annual International Symposium on Forecasting, 2011.