Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт космических и информационных технологий
Кафедра Информатики

УТВЕРЖДАЮ
Заведующий кафедрой

_____ Рубан А. И.
подпись инициалы, фамилия

"16" июня 2016 г.

БАКАЛАВРСКАЯ РАБОТА

27.03.03 «Системный анализ и управление»

Анализ временных рядов методом «сингулярного спектрального анализа»

Руководитель 13.06.16 доцент, К.Т.Н. Даничев А.А.
подпись, дата должность, ученая степень инициалы, фамилия

Студент КИ12-03Б 15.06.2016 Назаркин Н.С.
номер группы подпись, дата инициалы, фамилия

Красноярск 2016

РЕФЕРАТ

Выпускная квалификационная работа по теме "Анализ временных рядов методом «сингулярного спектрального анализа»" содержит 50 страницы текстового документа, 5 используемых источников, 41 иллюстрации.

ВРЕМЕННОЙ РЯД, ГУСЕНИЦА, СИНГУЛЯРНЫЙ ПЕКТРАЛЬНЫЙ АНАЛИЗ, МЕТОД ГЛАВНЫХ КОМПОНЕНТ.

Объект исследования: метод «Сингулярный спектральный анализ».

Цель исследования: описать методологию анализа временных рядов, используя метод сингулярного спектрального анализа(SSA).

В результате данной работы проведено детальной исследование метода SSA, изучена литература по данной теме, создана программа, реализующая данный метод.

В итоге был предложен алгоритм классификации собственных троек и реализован в программе.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время для изучения свойств сложных систем, в том числе и при экспериментальных исследованиях, широко используется подход, основанный на анализе сигналов, произведенных системой. Это очень актуально в тех случаях, когда математически описать изучаемый процесс практически невозможно, но в нашем распоряжении имеется некоторая характерная наблюдаемая величина. Например, в сейсмологии — запись колебаний земной коры, в метеорологии — данные метеонаблюдений. Обычно такой сигнал называется наблюдаемой, а метод исследования -реконструкцией динамических систем. Этот раздел теории динамических систем называется анализом временных рядов.

Проблема исследования заключается в отсутствии универсального метода анализа нестационарного временного ряда.

Исторически первыми были разработаны глобальные методы, в которых на основе статистического анализа предлагалось использовать авторегрессию, скользящее среднее и др. Позже в рамках нелинейной динамики были разработаны новые практические методики.

В настоящее время наиболее перспективным и сильно развивающимся методом является метод SSA (сингулярный спектральный анализ) в России больше известен под названием «Гусеница».

Объектом исследования Временной ряд и его сингулярное разложение.

Предметом исследования Метод анализа временных рядов SSA(сингулярный спектральный анализ).

В анализе временных рядов выделяются две основные задачи: задача идентификации и задача прогноза.

Задача идентификации при анализе наблюдаемых предполагает ответ на вопрос, каковы параметры системы, породившей данный временной ряд — размерность вложения, корреляционная размерность, энтропия.

Сейчас разработано и обосновано несколько различных методов прогноза. Однако все они подразделяются на два основных класса: локальные и глобальные. Такое деление проводится по области определения параметров аппроксимирующей функции, рекуррентно устанавливающей следующее значение временного ряда по нескольким предыдущим.

Цель исследования: описать методологию анализа временных рядов, используя метод сингулярного спектрального анализа(SSA).

Выдвижение данной цели обусловило постановку следующих **исследовательских задач**:

1. Исследования литературы по анализу временных рядов;
2. Программная реализация алгоритма SSA;
3. Автоматизация процесса группировки собственных троек;
4. Апробация программной реализации и алгоритма группировки;
5. Анализ качества решения;

Таким образом, теоретические исследования, основанные на анализе временных рядов, могут дать мощный инструмент для понимания многих явлений, особенно когда имеющихся данных для построения модели может быть недостаточно.

1 Теоретические сведенья

1.1 Временной ряд

При построении модели используются два типа данных:

- данные, характеризующие совокупность различных объектов в определенный момент времени;
- данные, характеризующие один объект за ряд последовательных моментов времени.

Модели, построенные по данным первого типа, называются пространственными моделями. Модели, построенные на основе второго типа данных, называются моделями временных рядов.

Временной ряд (ряд динамики) – это совокупность значений какого-либо показателя за несколько последовательных моментов или периодов времени. Временной ряд существенно отличается от простой выборки данных, так как при анализе учитывается взаимосвязь измерений со временем, а не только статистическое разнообразие и статистические характеристики выборки. Каждый уровень временного ряда формируется под воздействием большого числа факторов, которые условно можно подразделить на три группы:

- факторы, формирующие тенденцию ряда;
- факторы, формирующие циклические колебания ряда;
- случайные факторы.

Рассмотрим воздействие каждого фактора на временной ряд в отдельности.

Большинство временных рядов показателей имеют тенденцию, характеризующую совокупное долговременное воздействие множества факторов на динамику изучаемого показателя. Все эти факторы, взятые в отдельности, могут оказывать разнонаправленное воздействие на исследуемый

показатель. Однако в совокупности они формируют его возрастающую или убывающую тенденцию. На рисунке 1.1.1 показан гипотетический временной ряд, содержащий возрастающую тенденцию.

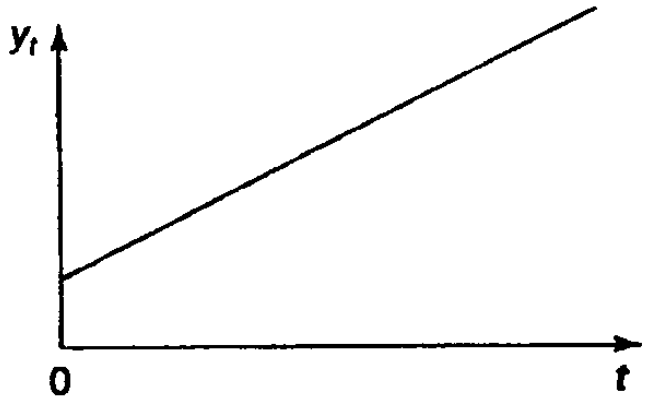

Рисунок 1.1.1 – Гипотетический временной ряд, содержащий возрастающую тенденцию.

Также изучаемый показатель может быть подвержен циклическим колебаниям. Эти колебания могут носить сезонный характер, поскольку экономическая деятельность ряда отраслей экономики зависит от времени года (например, цены на сельскохозяйственную продукцию в летний период выше, чем в зимний; уровень безработицы в курортных городах в зимний период выше по сравнению с летним). При наличии больших массивов данных за длительные промежутки времени можно выявить циклические колебания, связанные с общей динамикой конъюнктуры рынка. На рисунке 1.1.2 представлен гипотетический временной ряд, содержащий только сезонную компоненту.

Рисунок 1.1.2 – Гипотетический временной ряд, содержащий только сезонную компоненту.

Некоторые временные ряды не содержат тенденции и циклической компоненты, а каждый следующий их уровень образуется как сумма среднего уровня ряда и некоторой (положительной или отрицательной) случайной компоненты. Пример ряда, содержащего только случайную компоненту, приведен на рис. 3.

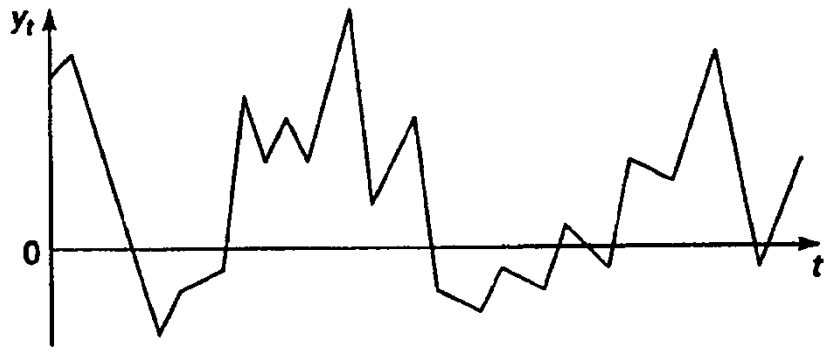

Рисунок 1.1.3 – Ряд, содержащей только случайную компоненту.

Очевидно, что реальные данные не следуют целиком и полностью из каких-либо описанных выше моделей. Чаще всего они содержат все три компоненты. Каждый их уровень формируется под воздействием тенденции, сезонных колебаний и случайной компоненты.

В большинстве случаев фактический уровень временного ряда можно представить как сумму или произведение трендовой, циклической и случайной компонент. Модель, в которой временной ряд представлен как сумма перечисленных компонент, называется аддитивной моделью временного ряда. Модель, в которой временной ряд представлен как произведение перечисленных компонент, называется мультипликативной моделью временного ряда. Основная задача исследования отдельного временного ряда – выявление и придание количественного выражения каждой из перечисленных выше компонент с тем, чтобы использовать полученную информацию для

прогнозирования будущих значений ряда или при построении моделей взаимосвязи двух или более временных рядов.

1.2 Анализ временного ряда

Существуют две основные цели анализа временных рядов: определение природы ряда и прогнозирование (предсказание будущих значений временного ряда по настоящим и прошлым значениям). Обе эти цели требуют, чтобы модель ряда была идентифицирована и, более или менее, формально описана. Как только модель определена, вы можете с ее помощью интерпретировать рассматриваемые данные (например, использовать в вашей теории для понимания сезонного изменения цен на товары, если занимаетесь экономикой). Не обращая внимания на глубину понимания и справедливость теории, вы можете экстраполировать затем ряд на основе найденной модели, т.е. предсказать его будущие значения.

Временные ряды исследуются с различными целями:

- Управление процессом, порождающим временной ряд - самые высокие требования к математической модели, которая должна описывать, в том числе, влияние управления на временной ряд;
- Предсказание будущего поведения временного ряда на основе истории - требования к математической модели полностью определяются требованиями к точности предсказания;
- Подбор статистической модели, описывающей временной ряд - Качество математических моделей в этом случае принято оценивать по количеству независимых параметров, использованных в них. Все специалисты знают, что, увеличивая размерность пространства параметров, ограниченный объем исходных данных можно подогнать под любую модель;
- Описание характерных особенностей ряда - Можно подумать, что математические методы в этом случае не нужны. На самом деле, на этом

уровне, подчас, применяются очень тонкие методы, например, проверка гипотезы случайности и др. Очень часто на этом уровне оказываются излишне амбициозные исследователи после последовательного спуска с трех предыдущих уровней, чтобы найти ту самую закономерность, которая позволит им улучшить прогноз.

В одном ряде случаях бывает достаточно получить описание характерных особенностей ряда, а в другом ряде случаев требуется не только предсказывать будущие значения временного ряда, но и управлять его поведением. Метод анализа временного ряда определяется, с одной стороны, целями анализа, а с другой стороны, вероятностной природой формирования его значений.

1.3 Распространненые методы анализа временных рядов

Спектральный анализ

Спектральный анализ является одним из самых мощных инструментов обработки эксперимента. В частности, он используется для анализа данных, выявления характерных частот, в целях подавления шума и т.д.

Спектром совокупности данных y(x) называют некоторую функцию другой координаты (или координат, если речь идет о многомерном спектре) F(ω), полученную в соответствии с определенным алгоритмом. Примерами спектров являются преобразование Фурье, спектр мощности, вейвлет-преобразование.

Спектральный анализ позволяет находить периодические составляющие временного ряда.

Корреляционный анализ

Корреляционный анализ позволяет выявить существенные свойства временных рядов. В том числе периодические зависимости и временные лаги

для единичного процесса (автокорреляция) или между несколькими процессами (кросскореляция).

Чем больше информации относительно величина Y содержится в исходных x1, x2,x3 … тем более тесную связь мы можем выявить между ними.

Установив характер взаимосвязи можно получить ожидаемое значение зависимой переменной при заданных значениях объясняющих переменных, то есть построить эконометрическую модель.

Добившись разбиения зависимой переменной на случайную и объясненную, мы можем, в частности, построить строить тренд.

Сам анализ состоит из нахождения взаимосвязей между значениями X(t) , нахождения тренда, между отклонением значений от линии тренда. Вычитая линию тренда из функции x(t) мы получим некоторые остатки, анализ этих остатков позволяет выявить существование периодичности и тенденции к смене тренда.

Модели авторегрессии и скользящего среднего

Модели ориентированы на описание процессов, проявляющих однородные колебания, возбуждаемые случайными воздействиями. Позволяют предсказывать будущие значения ряда.

Многоканальные модели авторегрессии и скользящего среднего

Модели применяются в тех случаях, когда имеется несколько коррелированных между собой временных рядов. В них имеются колебания, возбуждаемые одной причиной. Позволяют предсказывать будущие значения ряда.

Модель авторегрессии и скользящего среднего. Общая модель, предложенная Боксом и Дженкинсом (1976) включает как параметры авторегрессии, так и параметры скользящего среднего. Именно, имеется три

типа параметров модели: параметры авторегрессии (p), порядок разности (d), параметры скользящего среднего (q). В обозначениях Бокса и Дженкинса модель записывается как АРПСС (p, d, q). Например, модель (0, 1, 2) содержит 0 (нуль) параметров авторегрессии (p) и 2 параметра скользящего среднего (q), которые вычисляются для ряда после взятия разности с лагом 1.

1.4 Метод SSA(«Гусеница»)

Метод анализа временных рядов, основанный на преобразовании одномерного временного ряда в многомерный ряд с последующим применением к полученному многомерному временному ряду метода главных компонент.

Метод сочетает в себе элементы классического анализа временных рядов, многомерной статистики, многомерной геометрии, динамических систем и обработки сигналов. К источникам происхождения SSA можно отнести Метод главных компонент и классическую теорему Карунена-Лоэва для спектрального разложения временных рядом и цифровых изображений.

Краткое описание

Опишем кратко, как работает метод (строгое описание алгоритма содержится в разделе 1). Для анализа временного ряда выбирается целый параметр L; назовем его «длина окна». Параметр L может выбираться достаточно произвольно. При достаточно большой длине ряда и достаточно большом L результаты не будут зависеть от длины окна. Затем на основе ряда строится траекторная матрица, столбцами которой являются скользящие отрезки ряда длины L: с первой точки по L-ю, то второй по (L + 1)-ю и т. д. Следующий шаг — это сингулярное разложение траекторной матрицы в сумму элементарных матриц. Каждая элементарная матрица задается набором из собственного числа и двух сингулярных векторов — собственного и факторного.

Предположим, что исходный временной ряд является суммой нескольких рядов. Теоретические результаты позволяют при некоторых условиях определить по виду собственных чисел, собственных и факторных векторов, что это за слагаемые и какой набор элементарных матриц соответствует каждому из них. Суммируя элементарные матрицы внутри каждого набора и затем переходя от результирующих матриц к ряду, мы получаем разложение ряда на аддитивные слагаемые, например, на сумму тренда, периодики и шума или на сумму низкочастотной и высокочастотной составляющих. Возможность разбить совокупность элементарных матриц на группы, соответствующие интерпретируемым аддитивным составляющим ряда, тесно связана с понятием разделимости рядов, которое будет рассмотрено в разделе 2.

Таким образом, целью метода является разложение временного ряда на интерпретируемые аддитивные составляющие. При этом метод не требует стационарности ряда, знания модели тренда, а также сведений о наличии в ряде периодических составляющих и их периодах. При таких слабых предположениях метод «Гусеница»- SSA может решать различные задачи, такие как, например, выделение тренда, обнаружение периодик, сглаживание ряда, построение полного разложения ряда в сумму тренда, периодик и шума.

Платой за такой широкий спектр возможностей при достаточно слабых предположениях является, во-первых, существенно неавтоматическая группировка компонент сингулярного разложения траекторной матрицы ряда для получения составляющих исходного ряда. Во-вторых, отсутствие модели не позволяет проверять гипотезы о наличии в ряде той или иной составляющей (этот недостаток объективно присущ непараметрическим методам). Для проверки подобных гипотез требуется построение модели, которое, в свою очередь, может быть проведено на основе информации, получаемой с помощью метода Гусеница. Отметим также, что рассматриваемый непараметрический метод позволяет получить результаты, часто лишь незначительно менее

точные, чем многие параметрические методы при анализе ряда с известной моделью.

Базовый алгоритм

Пусть N > 2. Рассмотрим веществennозначный временной ряд F = (fo,..., fN-1) длины N. Будем предполагать, что ряд F — ненулевой, т. е. существует, по крайней мере, одно i, такое что $f_i \neq 0$. Обычно считается, что fi = f (iΔ) для некоторой функции f (t), где t — время, а Δ — некоторый временной интервал, однако это не будет играть особой роли в дальнейшем. Более того, числа 0,..., N — 1 могут быть интерпретированы не только как дискретные моменты времени, но и как некоторые метки, имеющие линейно-упорядоченную структуру.

Нумерация значений временного ряда начинается с i = 0, а не стандартно с i = 1 только из-за удобства обозначений.

Базовый алгоритм состоит из двух дополняющих друг друга этапов, разложения и восстановления.

Шаг 1. Вложение

Процедура вложения переводит исходный временной ряд в последовательность многомерных векторов.

Пусть L — некоторое целое число (длина окна), 1 < L < N. Процедура вложения образует K = N — L +1 векторов вложения

$$X_i = (f_{i-1}, \dots, f_{i+L-2})^T \quad 1 < i < K, \qquad (1)$$

имеющих размерность L. Если нам нужно будет подчеркнуть размерность X_i, то мы будем называть их векторами L-вложения.

L-Траекторная матрица (или просто траекторная матрица) ряда F

$$X = [X_1 : \ldots : X_k] \quad (2)$$

состоит из векторов вложения в качестве столбцов.

Другими словами, траекторная матрица – матрица.

$$X = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} f_0 & \cdots & f_{K-1} \\ \vdots & \ddots & \vdots \\ f_{L-1} & \cdots & f_{N-1} \end{pmatrix} \quad (3)$$

Очевидно, что $x_{ij} = f_{i+j-2}$ и матрица X имеет одинаковые элементы на «диагоналях» i + j = const. Таким образом, траекторная матрица является ганкелевой. Существует взаимно-однозначное соответствие между ганкелевыми матрицами размерности $L \times K$ и рядами длины N = L + K — 1.

Шаг 2. Сингулярное разложение

Результатом этого шага является сингулярное разложение (SVD = Singular Value Decomposition) траекторией матрицы ряда.

Пусть S = XXT. Обозначим $\lambda_1, \ldots, \lambda_L$ собственные числа матрицы S, взятые в неубывающем порядке (λ_1 > ... > λ_L > 0) и U1,... ,UL — ортонормированную систему собственных векторов матрицы S, соответствующих собственным числам.

Пусть d = max{i : λ_i > 0}. Если обозначить $V_i = X^T U_i / \sqrt{\lambda_i}, i = 1, \ldots, d,$ то сингулярное разложение матрицы X может быть записано как

$$X = X_1 + \ldots + X_d, \quad (4)$$

где $X_i = \sqrt{\lambda_i} U_i V_i^T$. Каждая из матриц из матриц X_i имеет ранг 1. Поэтому их можно назвать элементарными матрицами.

Набор $(\sqrt{\lambda_i}, U_i, V_i)$ мы будем называть i-й собственной тройкой сингулярного разложения.

Шаг 3. Группировка

На основе разложения (4) процедура группировки делит все множество индексов {1,...,d} на m непересекающихся подмножеств $I_1, \ldots, I_m$

Пусть I = $\{i_1, \ldots, i_p\}$. Тогда результирующая матрица X_I, соответствующая группе I, определяется как

$$X_I = X_{i1} + \ldots + X_{ip} \quad (5)$$

Такие матрицы вычисляются для $I = I_1, \ldots, I_p$, тем самым разложение (4) может быть записано в сгруппированном виде

$$X = X_{I1} + \ldots + X_{Im} \quad (6)$$

Процедура выбора множеств $I = I_1, \ldots, I_p$, и называется группировкой собственных троек.

Шаг 4. Диагональное усреднение

На последнем шаге алгоритма каждая матрица сгруппированного разложения переводится в новый ряд длины N.

Пусть Y — некоторая L × K матрица с элементами y_{ij}, где 1 ≤ i ≤ L, 1 ≤ j ≤ K. Положим $L^* = \min(L, K)$, $K^* = \max(L, K)$, и N = L + K - 1. Пусть $y^*_{ij} = y_{ij}$, если L<K, и $y^*_{ij} = y_{ji}$ иначе.

Диагональное усреднение переводит матрицу Y в ряд $g_0, \ldots, g_{N-1}$ по формуле:

$$g_k = \begin{cases} \frac{1}{k+1}\sum_{m=1}^{k+1} y^*_{m,k-m+2} & \text{, для } 0 \le k \le L^* - 1 \\ \frac{1}{L^*}\sum_{m=1}^{L^*} y^*_{m,k-m+2} & \text{, для } L^* - 1 \le k \le K^* \\ \frac{1}{N-k}\sum_{m=k-K^*+2}^{N-K^*+1} y^*_{m,k-m+2} & \text{, для } K^* \le k \le N \end{cases} \quad (7)$$

Выражение (7) соответствует усреднению элементов матрицы вдоль «диагоналей» i + j = k + 2: выбор k = 0 дает $g_0 = y_{11}$, для k = 1 получаем $g_1 = (y_{12} + y_{21})/2$ и т. д. Заметим, что если матрица Y является траекторной матрицей некоторого ряда $(h_0, \ldots, h_{N-1})$ (другими словами, если матрица Y является ганкелевой), то $g_i = h_i$ для всех i.

Примечание к шагу 2

Сингуля́рное разложе́ние (singular value decomposition, SVD) — это разложение прямоугольной вещественной или комплексной матрицы, имеющее широкое применение, в силу своей наглядной геометрической интерпретации, при решении многих прикладных задач.

Так вот, математики утверждают, что любую матрицу можно разложить в произведение трех матриц:

$$A = U*W*V^T \quad (8)$$

где U и V – ортогональные матрицы, а W – диагональная матрица. Причем, математики предпочитают когда все отсортировано, поэтому диагональные элементы матрицы W располагаются в порядке убывания. Это и есть сингулярное разложение или SVD.

Это очень важное свойство. Теперь, удаляя наименьшие (т.е. последние) сингулярные числа и проводя обратное преобразование, мы можем эффективно убирать шумы из наших данных. На этом построены, например, некоторые способы улучшения и сжатия изображений.

Геометрический смысл

Пусть матрице A поставлен в соответствие линейный оператор. Сингулярное разложение можно переформулировать в геометрических терминах. Линейный оператор, отображающий элементы пространства $\mathbb{R}^n$ в себя представим в виде последовательно выполняемых линейных операторов вращения и растяжения. Поэтому компоненты сингулярного разложения наглядно показывают геометрические изменения при отображении линейным оператором A множества векторов из векторного пространства в себя или в векторное пространство другой размерности[1].

Для более визуального представления рассмотрим сферу S единичного радиуса в пространстве $\mathbb{R}^n$. Линейное отображение T отображает эту сферу в эллипсоидпространства $\mathbb{R}^m$. Тогда ненулевые сингулярные значения диагонали матрицы Σ являются длинами полуосей этого эллипсоида. В случае когда $n = m$ и все сингулярные величины различны и отличны от нуля, сингулярное разложение линейного отображения T может быть легко проанализировано как последствие трех действий: рассмотрим эллипсоид $T(S)$ и его оси; затем рассмотрим направления в $\mathbb{R}^n$, которые отображение T

переводит в эти оси. Эти направления ортогональны. Вначале применим изометрию $\mathbf{v}^*$, отобразив эти направления на координатные оси $\mathbb{R}^n$. Вторым шагом применим эндоморфизм $\mathbf{d}$, диагонализированный вдоль координатных осей и расширяющий/сжимающий эти направления, используя длины полуосей $T(S)$ как коэффициенты растяжения. Тогда произведение $\mathbf{d} \otimes \mathbf{v}^*$ отображает единичную сферу на изометричный эллипсоид $T(S)$. Для определения последнего шага u, просто применим изометрию к этому эллипсоиду так, чтобы перевести его в $T(S)$. Как можно легко проверить, произведение $\mathbf{u} \otimes \mathbf{d} \otimes \mathbf{v}^*$ совпадает с T.

Прогноз

На этом этапе строится ряд $(x_i)_{i=1}^{N+p}$. При этом прогноз на p точек вперед осуществляется как применение p раз операции прогноза на одну точку.

Базовая идея нахождения значения x_{N+1} (9) состоит в следующем. Пусть имеется набор значений $x_1, x_2, \ldots, x_{N+1}$. Теперь построим выборку в виде матрицы X. В качестве базиса поверхности, содержащей эту выборку, можно взять отобранные ранее собственные векторы $V^1, V^2, \ldots, V^L$ (11) матрицы C.

$$x_{N+1} = \frac{V_L V_*^T Q}{1 - V_L V_*^T} \tag{9}$$

$$V_* = \begin{pmatrix} V_1^1 & \cdots & V_1^L \\ \vdots & \ddots & \vdots \\ V_{L-1}^1 & \cdots & V_{L-1}^L \end{pmatrix} \tag{10}$$

$$V_L = (V_L^1, V_L^2, \cdots, V_L^L) \tag{11}$$

2 Анализ временных рядов методом SSA

На этапе группировки собственных троек необходимо сгруппировать собственные вектора на аддитивные составляющие: тренд, периодика (гармоника) и шум. Собственный вектор, принадлежащий к тренду, будет медленно изменятся (смотри главу 3 рисунок 3.3).

Гармоническая составляющая имеет синусоидальные собственные вектора, так как имеет регулярное периодическое поведение. Однако, поскольку гармоника с периодом большим, чем 2, порождает две собственные тройки (косинус с периодом 2 порождает только одну собственную тройку, сингулярные вектора которой имеют пилообразный вид, а таких на одномерных диаграммах в данном случае не обнаружено), то искать относящиеся к гармоникам нужно пары собственных троек.

Отдельно остановимся на вопросе отделения компонент, относящихся к сигналу, от шумовых компонент. Во-первых, нерегулярное поведение сингулярных векторов может говорить о принадлежности их к набору, порожденному шумовой компонентой (эти случаи нужно не путать с перемешиванием компонент, порожденным отсутствием сильной разделимости рядов). Во-вторых, медленное, почти без скачков, убывание собственных значений с некоторого номера также говорит об этом. В-третьих, большой набор собственных троек, порождающих коррелирующие друг с другом восстановленные компоненты, скорее всего относится к шуму(смотри главу 3).

2.1 Классификация

Автоматизируя процесс группировки собственных троек, предложено ориентироваться на собственные числа (рисунок 2.1.1). Из исследования стало ясно, что собственные числа, принадлежащие к тренду, имеют наибольшие веса, т.е. находятся в начале списка собственных троек (т.к. сингулярное разложение сортирует по убыванию весов собственных троек). Гармоникам соответствуют пары соседних векторов, которые находятся на одной

«ступеньки», т.е. имеют схожие по значению собственные вектора. Шум же определяется, как медленное убывание собственных чисел.

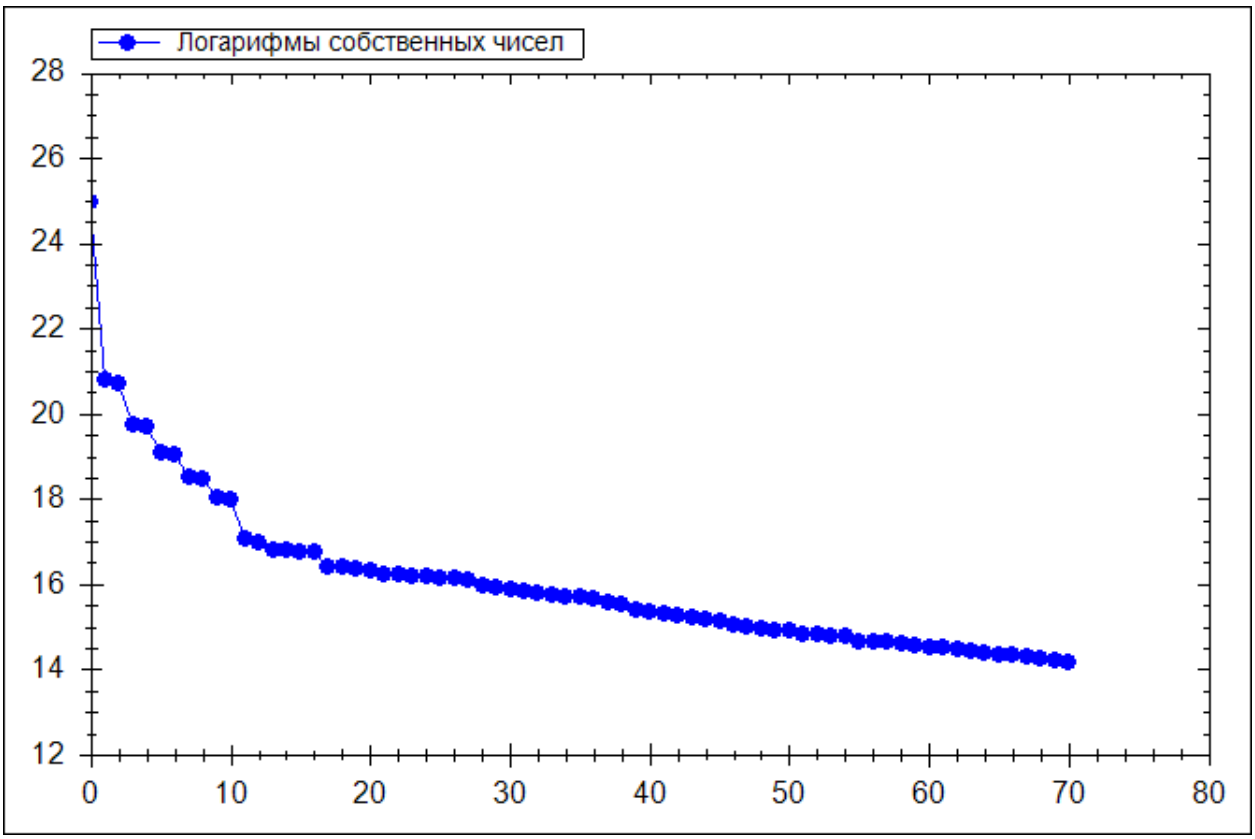

Рисунок 2.1.1 – Собственные числа

Предложено классифицировать собственные тройки по собственным числам. Исследованные научные методы кластеризации «K-средних», «K-ближайших соседей» подразумевают изначальную известность количества классов. При применении метода SSA заранее невозможно сказать, сколько будет собственных пар чисел относящихся к гармоническим составляющим.

Также было замечено, что расстояние между парами собственных чисел взаимозависимых гармоник во много больше, чем расстояния в самих парах между собственными числами. Также известно, что существует последовательность тренд-гармоника-шум. Шумовой вектор может оказаться между двумя собственными числами, но, не имея пары, мы причислим его к шуму. Тренд, как правило, описывается одним собственным вектором.

Используя эти наблюдения, была осуществлена классификация на тренд, гармоническую и шумовую оставляющую по следующему алгоритму:

1) Прологарифмируем собственные числа.
2) Подсчитаем расстояния между каждыми соседними точками и переведем их в процентное соотношение от суммы всех расстояний.
3) Введем коэффициент a%.
4) Берутся последовательно две точки. Если расстояния между ними больше a%, то создаем новый класс. Если меньше a, то добавляем вторую точку в предыдущий класс.
5) Если образовался класс с 1 собственным числом, относим его к шуму в последний класс.
6) Как замечено выше первый класс относим к тренду если он состоит из одного собственного вектора, а последний к шуму. Следовательно, что между ними - это гармоники. Объединим их в один класс.

На рисунке 2.1.1 показаны логарифмы собственных чисел. Алгоритм классификации разбил нам их на 7 классов (рисунок 2.1.2). После чего нулевой (тренд) т.к. состоит из одного вектора и шестой класс (шум) остается нетронутым, а остальные классы (гармоники) объединили в один класс. Номера собственных векторов разбитых на классы тренд, гармоники, шум можно видеть на рисунках 2.1.3 2.1.4 2.1.5 соответственно.

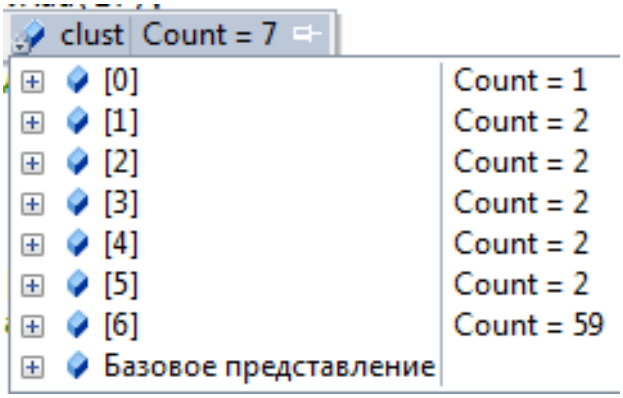

Рисунок 2.1.2 – Классы

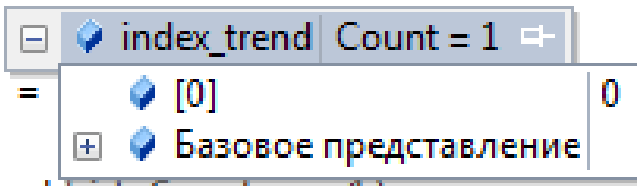

Рисунок 2.1.3 – Класс тренд

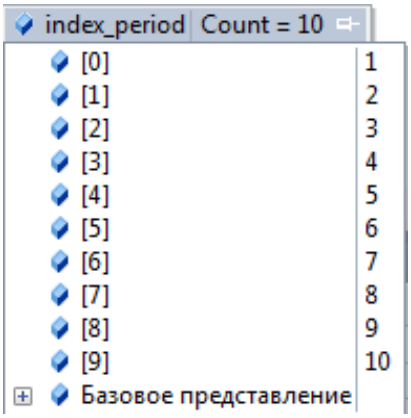

Рисунок 2.1.4 – Класс гармоник

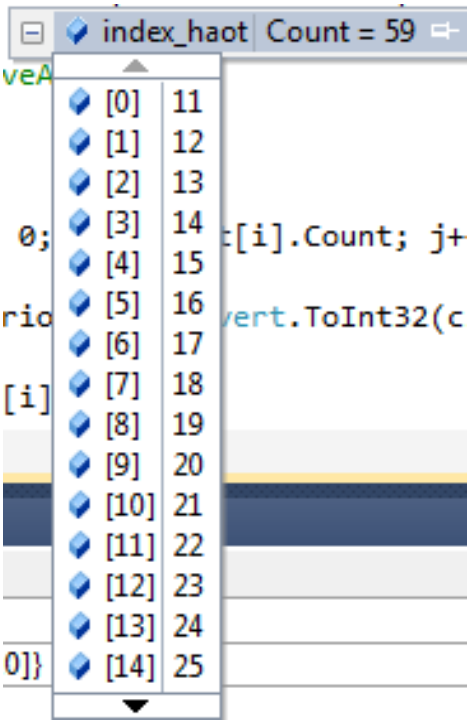

Рисунок 2.1.5 – Класс шум

2.2 Анализ выявленных гармоник

Удобнее проводить анализ выявленных пар зависимых собственных векторов на двумерных диаграммах (скаттеграммах). Так как при достаточно большой длине ряда соответствующая пара собственных чисел имеет близкие значения, то достаточно рассматривать двумерные диаграммы собственных векторов из соседних, упорядоченных по собственным значениям, собственных троек.

Каждая точка скаттеграммы образуется значениями собственных векторов по осям X и Y в один и тот же момент времени.

Анализируя скаттеграммы можно получить много полезной информации. Например, на рисунке 2.2.1 можно видеть, что есть гармоническая составляющая с периодом 4, на рисунке 2.2.2 с периодом 8. На рисунке 2.2.3 гармоническая составляющая с периодом 12(годовым) с затухающей со временем амплитудой. А на рисунке 2.2.4 нет, какой либо периодической зависимости, поэтому это шум.

Также бывают случаи, когда на скаттеграмме можно определить составляющую, за счет которой происходит изменение частоты в ряде (рисунок 2.2.5).

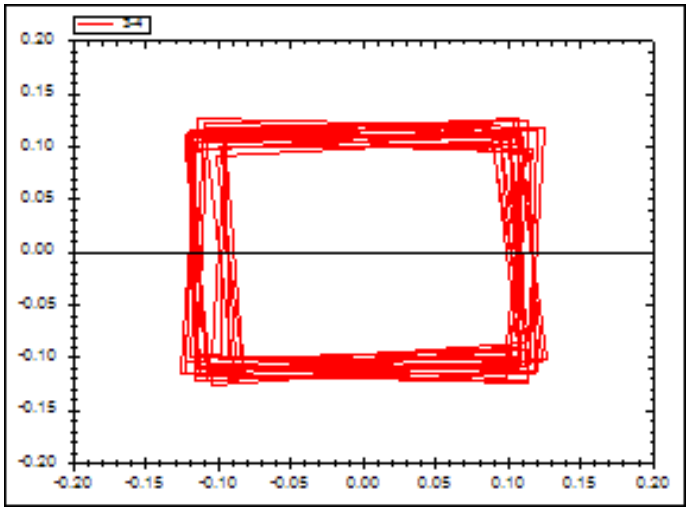

Рисунок 2.2.1 Период 4

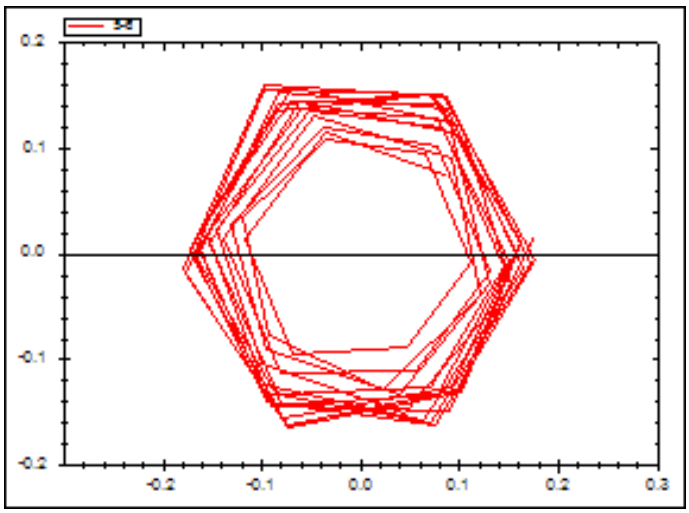

Рисунок 2.2.2 – Период 8

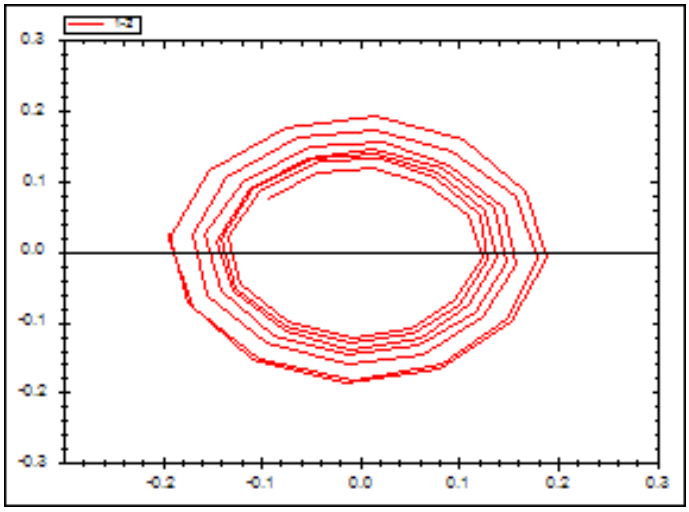

Рисунок 2.2.3 – Период 12

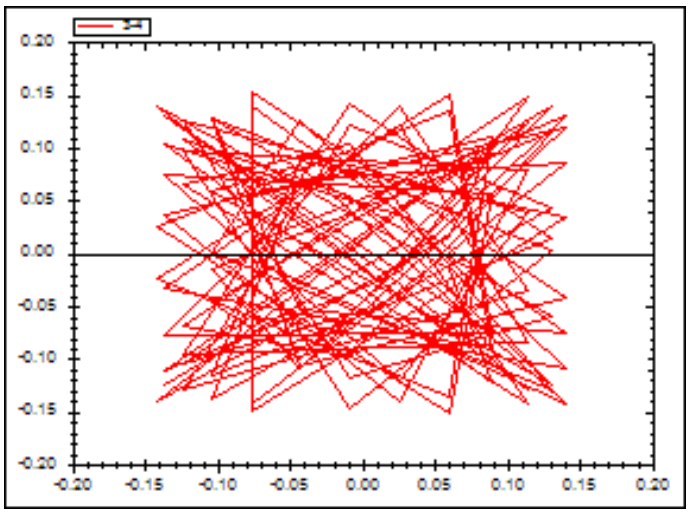

Рисунок 2.2.4 – Шум

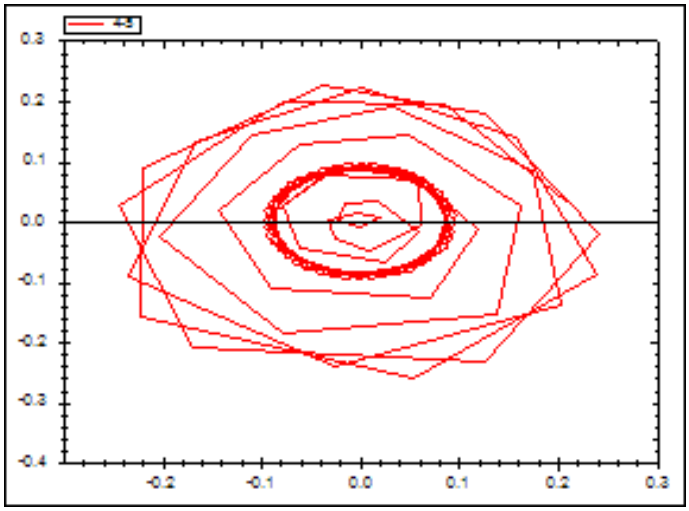

Рисунок 2.2.5 – Изменение частоты

2.3 Выявление псевдогармоник

Бывают случаи, когда метод SSA не может справиться со своими задачами. Дан временной ряд (рисунок 2.3.1). Как видно на рисунке прогноз скорей всего сделан не правильно. По виду собственных чисел (рисунок 2.3.2) можно утверждать что ET0,1; ET2,3; ET4,5; ET6,7; ET8,9 являются гармоническими составляющими ряда. Также это подтверждают и скаттеграммы собственных векторов(рисунок 2.3.3). Но прогноз получается не верным. Это происходит из-за того, что выявленные гармоники не являются

периодическими составляющими ряда, т.е. являются псевдогармониками. На протяжении всего временного ряда собственные вектора его аппроксимировали но далее компенсируясь и накладываясь друг на друга (из за разных частот и фаз) выдали неправильный результат.

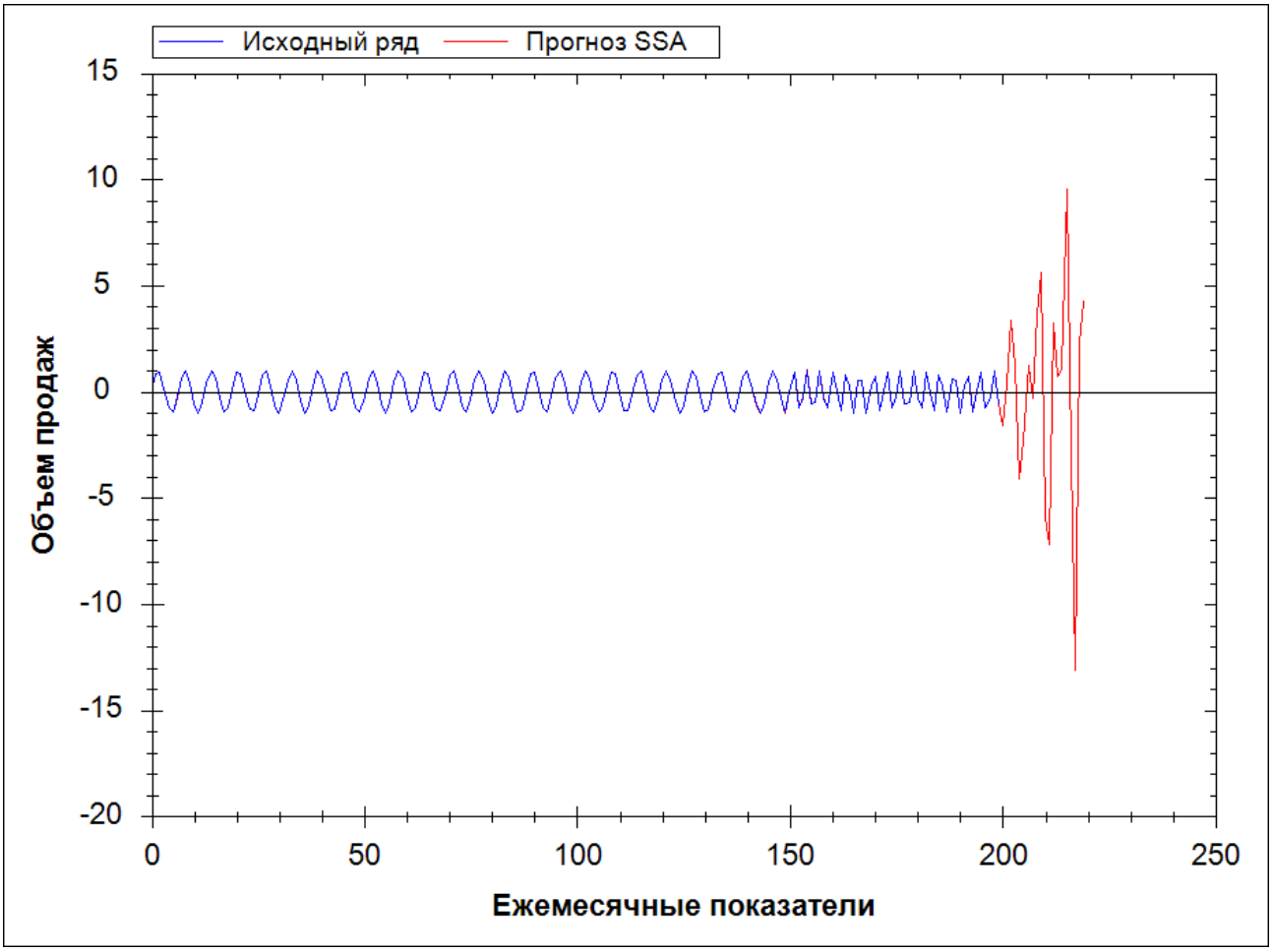

Рисунок 2.3.1 – Временной ряд

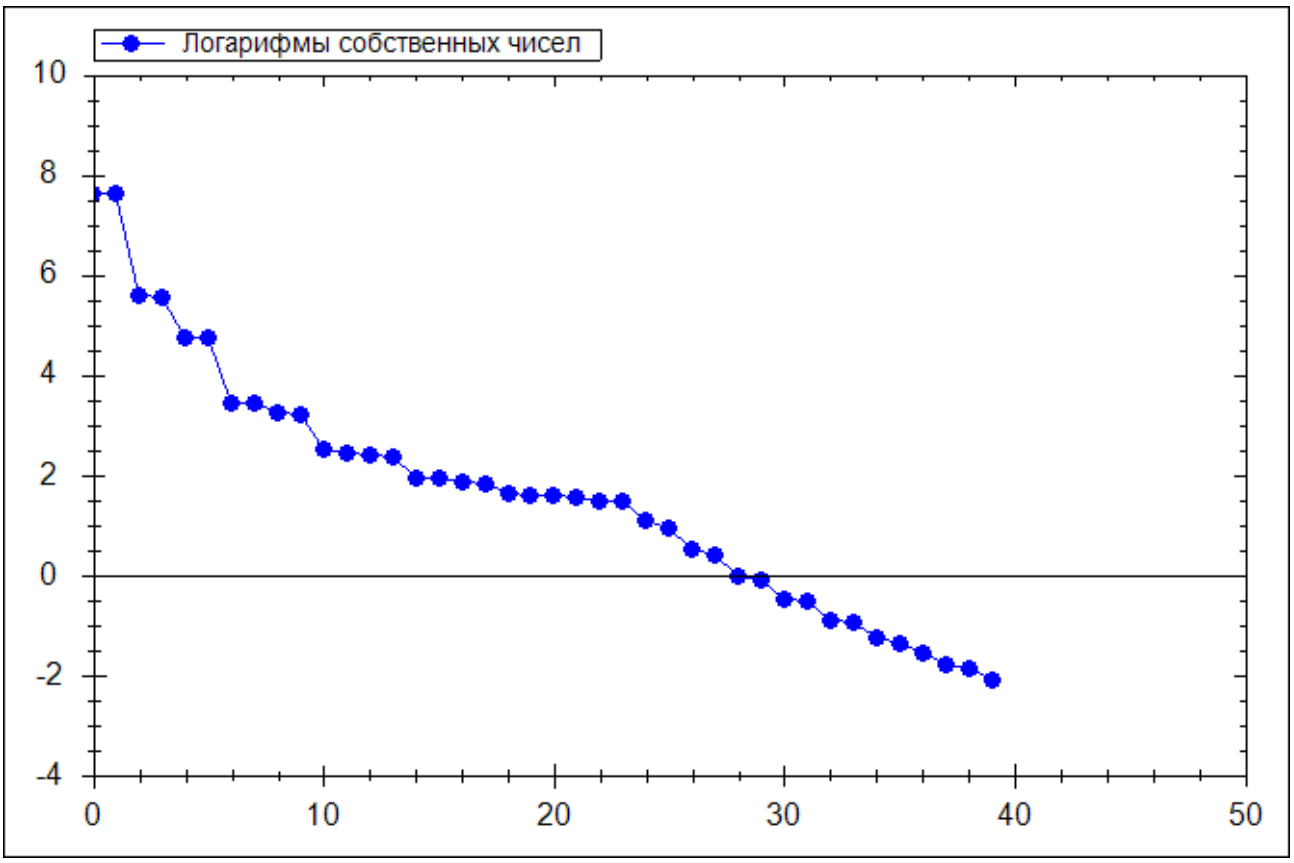

Рисунок 2.3.2 – Собственные числа

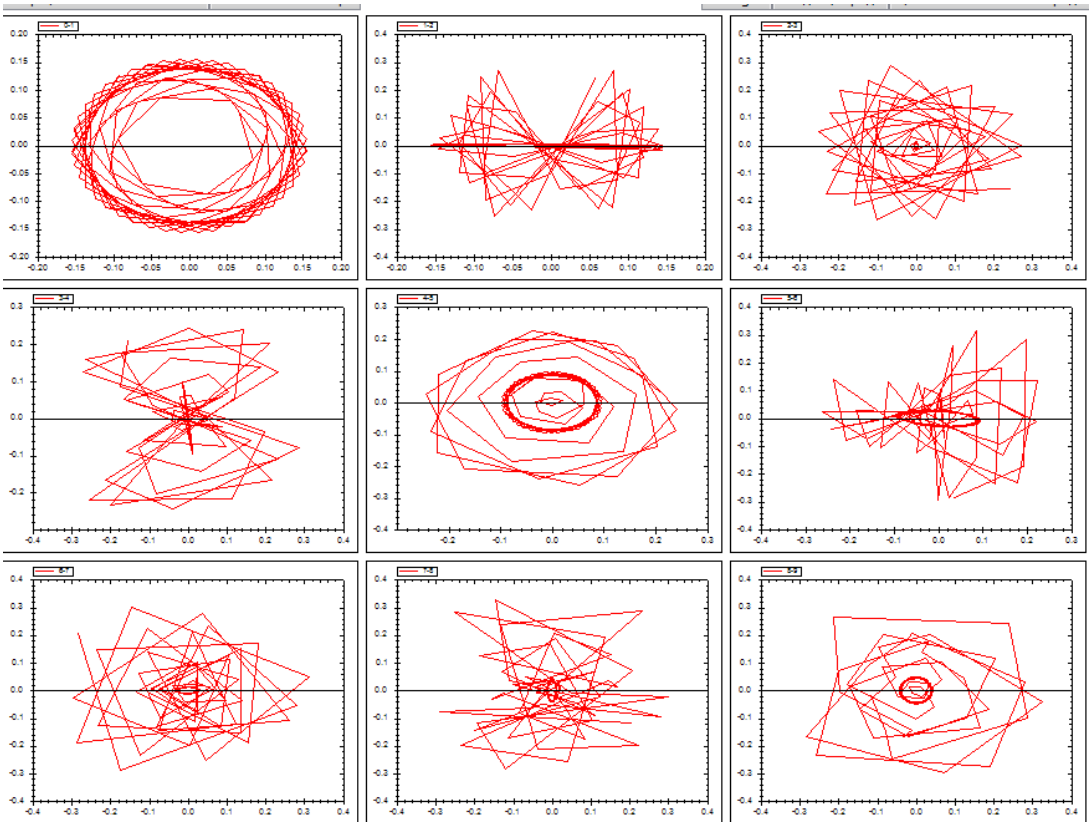

Рисунок 2.3.3 – Скаттеграммы собственных векторов

3 Программный продукт и апробация алгоритма.

3.1 Описание программы

В целях исследования алгоритма была написана программа на языке C# в среде программирования Microsoft Visual Studio 2010. Для реализации сингулярного разложения была использована функция «SVD» языка программирования R в библиотеке RDOTNET. На рисунке 3.1.1 показан вид программы.

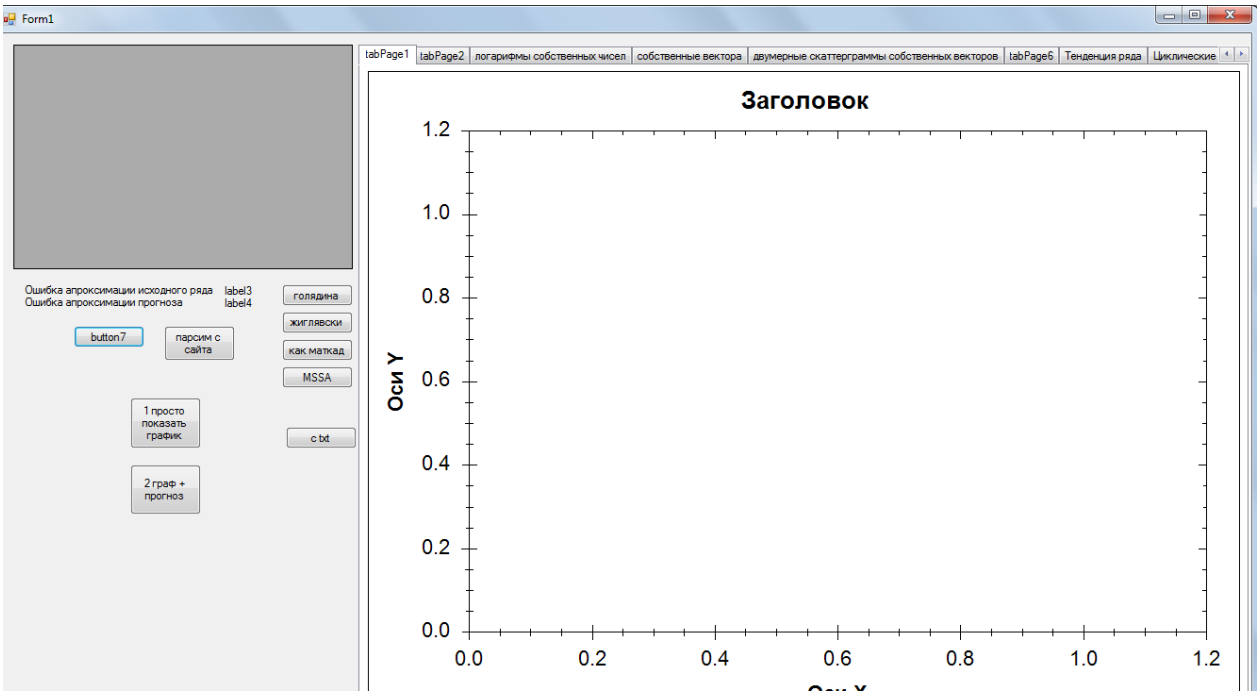

Рисунок 3.1.1 – Вид программы

Временные ряды загружаются с файлов формата txt, которые находятся в корневой папке программы. Для удобства исследования алгоритма в программе было реализовано создание тестовых временных рядов.

3.2 Описание работы программы на примере временной ряд «Ford»

Продемонстрируем методику применения метода «Гусеница»-SSA на примере анализа временного ряда «Объемы месячных продаж крепленых вин в Австралии с января 1984 года по июнь 1994 года». Так как мы будем рассматривать как сам ряд длины N = 174.

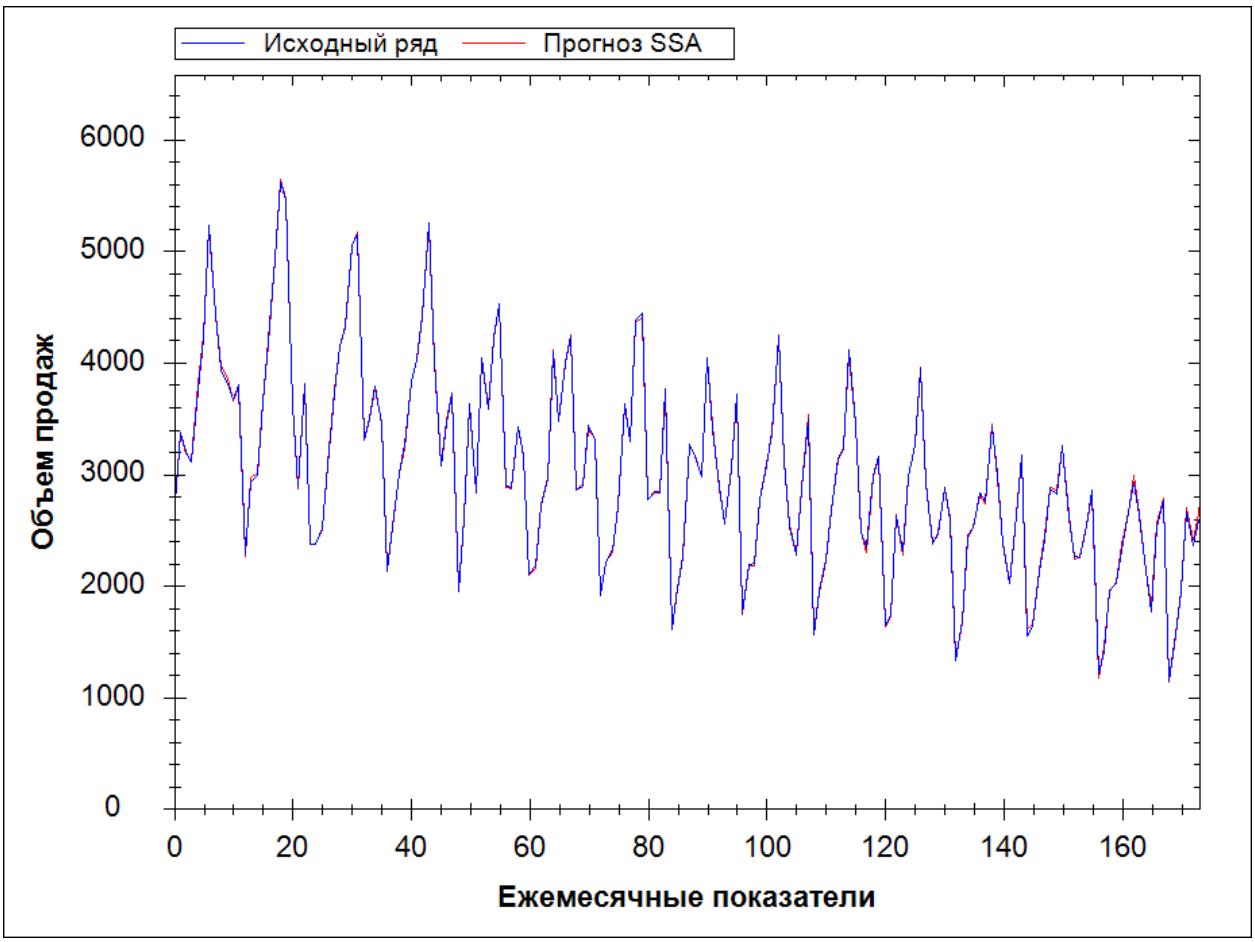

Рисунок 3.2.1 – «Объемы месячных продаж крепленых вин в Австралии с января 1984 года по июнь 1994 года»

Визуальный анализ ряда, изображенного на рисунке 3.2.1, говорит о том, что ряд имеет тренд, который должен хорошо описываться либо линейной функцией, либо убывающей экспонентой, а также сезонное поведение довольно сложной меняющейся формы.

Поставим задачу поиска разложения данного ряда на три компоненты — тренд, сезонную компоненту и шум.

Выберем длину окна равной L = 84.

Подход к идентификации собственных троек. Рассмотрим результат сингулярного разложения траекторной матрицы ряда при таком выборе длины окна. На рисунке 3.2.2 приведено изображение собственных векторов из 71

собственных троек сингулярного разложения, которые несут в себе 99.99% информации о сигнале.

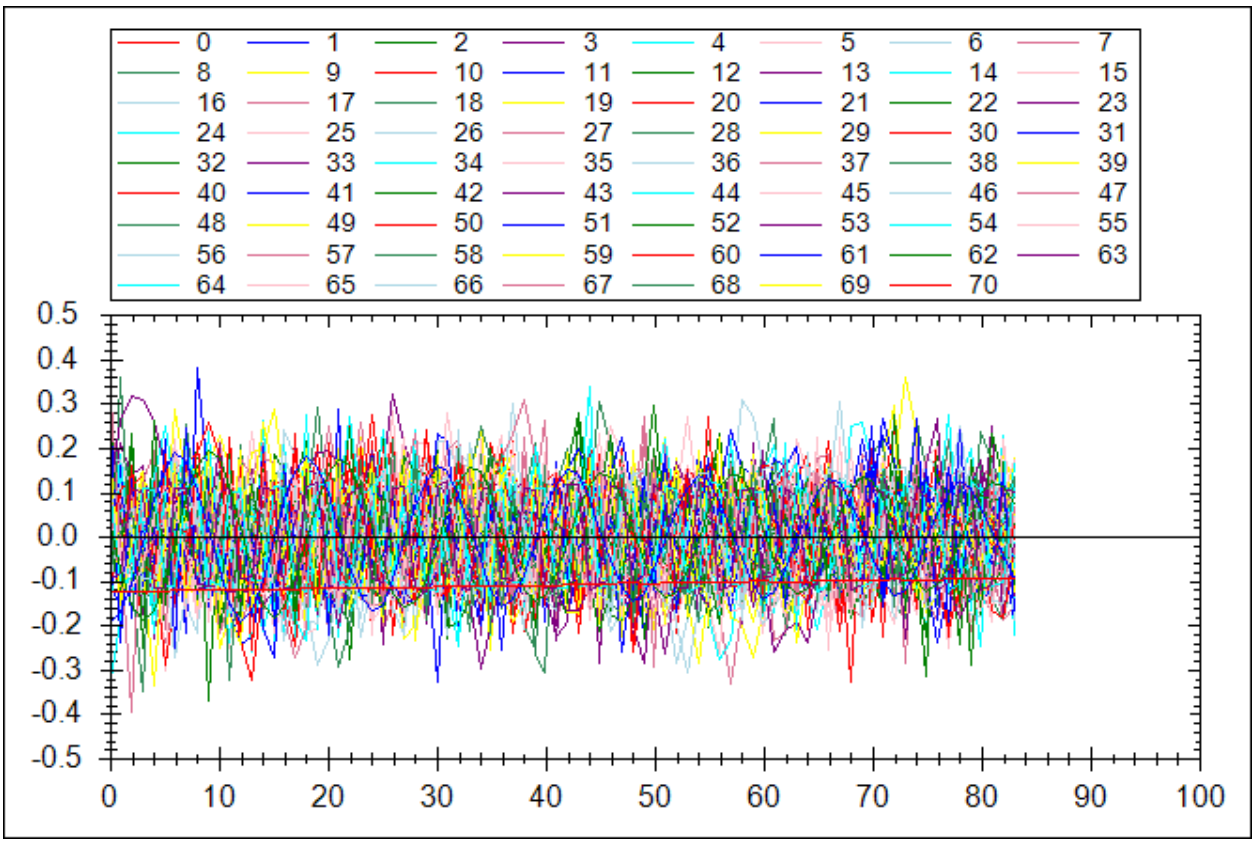

Рисунок 3.2.2 – Собственные вектора

Для идентификации собственных троек воспользуемся результатами о виде собственных векторов, соответствующих тренду и гармоникам при условии из приближенной разделимости.

Для большей наглядности первых 7 собственных векторов собственных векторов покажем рисунке 3.2.3.

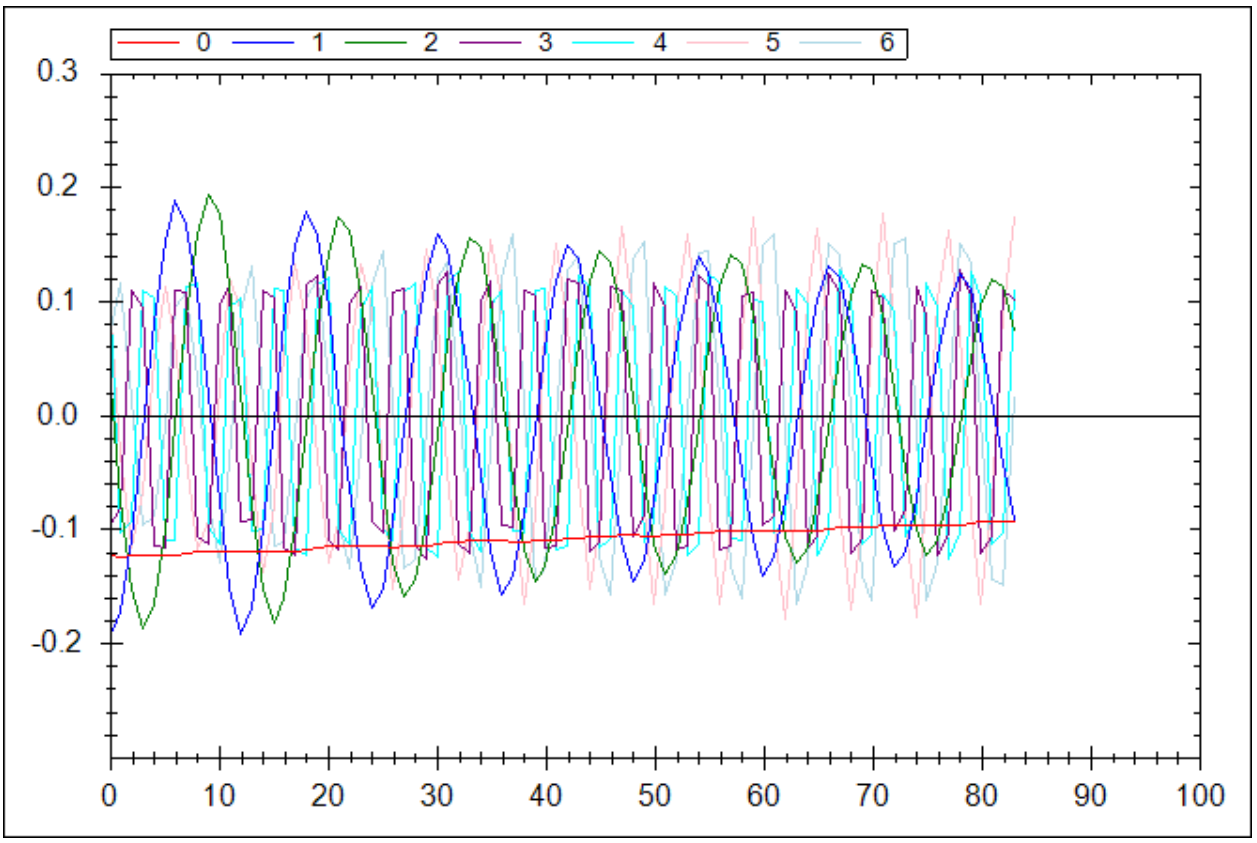

Рисунок 3.2.3 – Собственные вектора

Идентификация тренда. Начнем с идентификации тренда. Мы знаем, что сингулярные (в частности, собственные) вектора имеют в целом такой же вид, как и компонента исходного ряда, которой они соответствуют. Поэтому на одномерных диаграммах собственных векторов нужно найти медленно меняющиеся собственные вектора. В данном случае только один, а именно, нулевой собственный вектор имеет требуемый вид. Так как получилось, что в данном случае тренд описывается единственной собственной тройкой, то это означает, что тренд аппроксимируется экспонентой. Чем сложнее форма тренда, тем больше его (приближенная) размерность и тем большее число собственных троек ему соответствует.

Идентификация гармоник. Займемся теперь идентификацией гармонических (возможно, с меняющейся амплитудой) компонент, порожденных сезонной компонентой исходного ряда. На рисунке 3.2.2 видно,

что собственные тройки с номерами 2—6 возможно соответствуют каким-либо гармоникам, так как имеют регулярное периодическое поведение. Искать относящиеся к гармоникам пары собственных троек удобнее на двумерных диаграммах (скаттеграммах). На рисунке 3.2.4 можно различить регулярные двумерные изображения, образующие двумерные траектории с вершинами, лежащими на кривой, имеющей спиралеобразную форму. Это означает, что соответствующая пара собственных векторов порождена модулированной гармонической компонентой исходного ряда. Таким образом, получаем, что собственные тройки (используем аббревиатуру ET — eigentriples) ET2,3 соответствуют периоду 12, ET4,5 — периоду 4, КТ6.7 — периоду 6, КТ8.9 — периоду 2.4 (говоря о дробном периоде, мы имеем в виду гармонику с частотой, обратной этому периоду, в данном случае — 5/12), ET10Д1 — периоду 3.

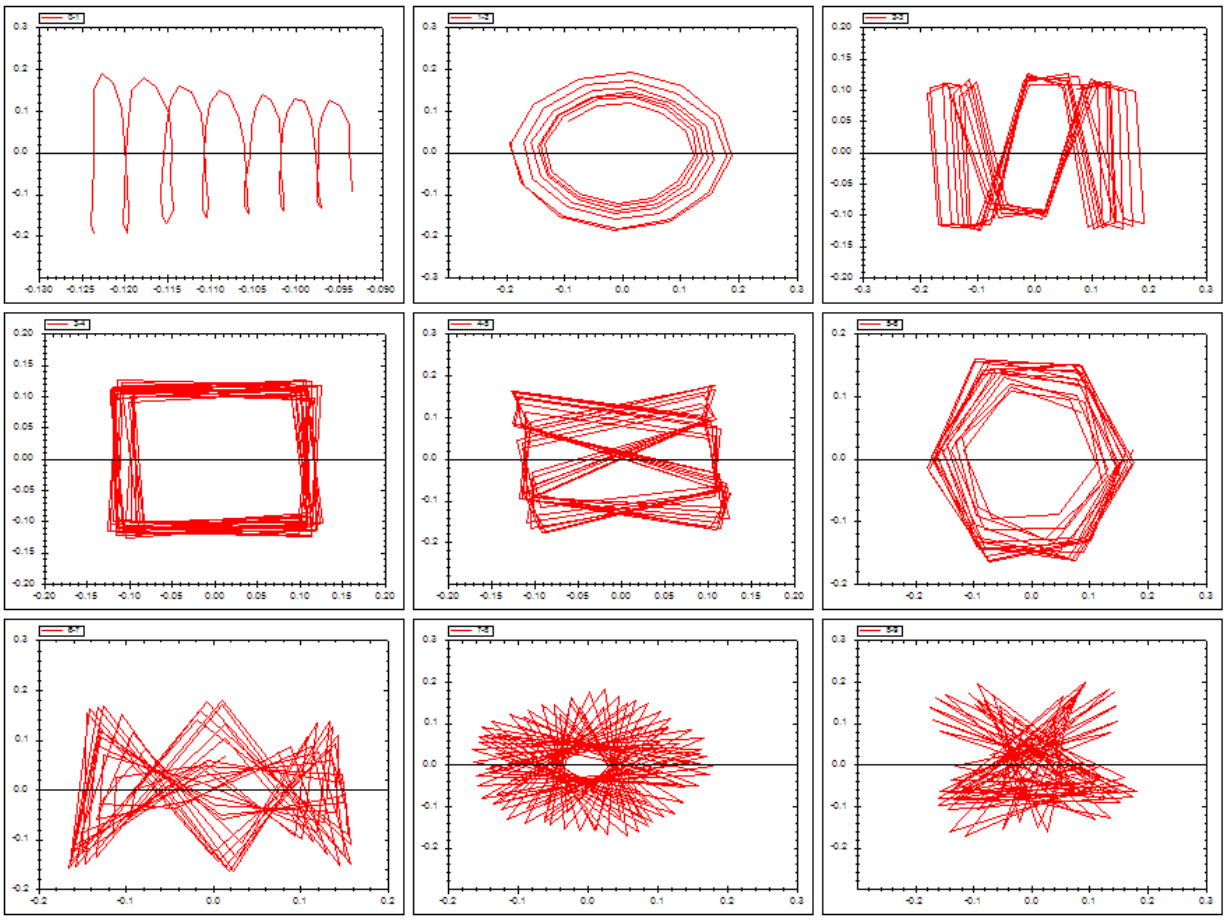

Рисунок 3.2.4 – Скаттеграммы пар собственных векторов

Вспомогательные характеристики. Рассмотрим, какая дополнительная информация может помочь для идентификации собственных троек (или подтвердить то, что получено). На рисунке 3.2.5, на котором изображены логарифмы собственных значений, подтверждает найденные пары собственных троек (каждой паре соответствует “ступенька”).

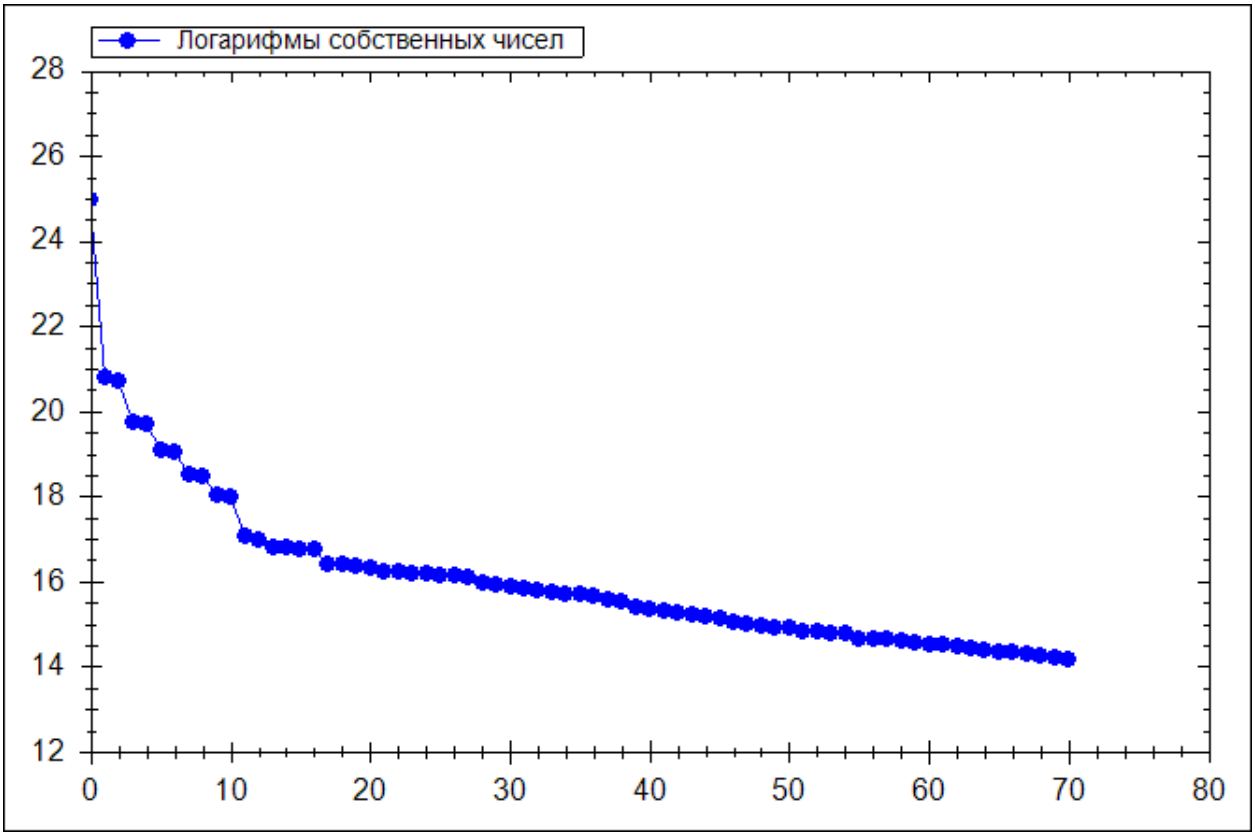

Рисунок 3.2.5 – Логарифмы собственных чисел

Отделение сигнала от шума. Медленное, почти без скачков, убывание собственных значений с некоторого номера может говорить о том, что эти компоненты составляют шум. Рисунок 3.2.4 показывает, что собственные числа с номерами 14—84 как раз образуют такой блок. Так же можно отметить, что веса у данных компонент настолько малы, что можно принять их за шум.

Подтверждением правильности разделения сигнала и шума является проверка на принадлежность к шуму компоненты ряда, полученной с помощью восстановления по так сказать “шумовым” собственным тройкам. На рисунках

3.2.6, 3.2.7, 3.2.8 изображено разложение исходного ряда на три компоненты — тренд (ET1; на фоне исходного ряда), периодику (ET2-11) и шум (ET12-84). Третья компонента является реализацией белого шума.

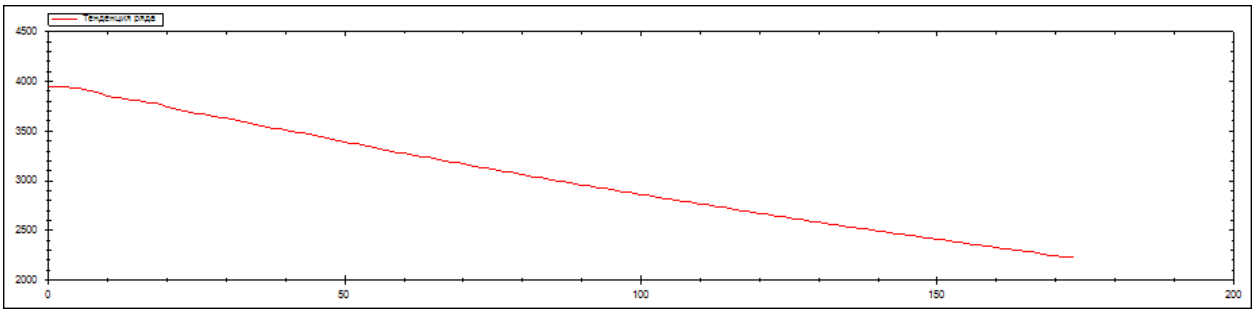

Рисунок 3.2.6 – Тренд

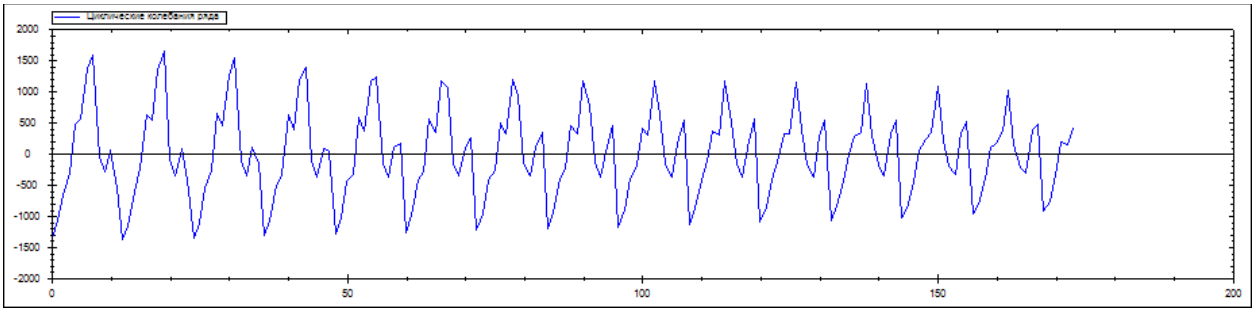

Рисунок 3.2.7 – Гармоника

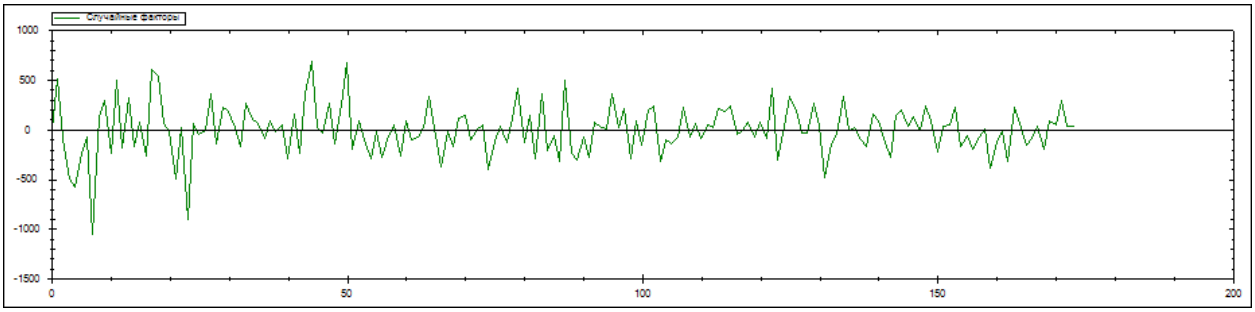

Рисунок 3.2.8 – Белый шум

На рисунках 3.2.9, 3.2.10, 3.2.11 показаны восстановленные компоненты ряда в сравнении с исходным рядом.

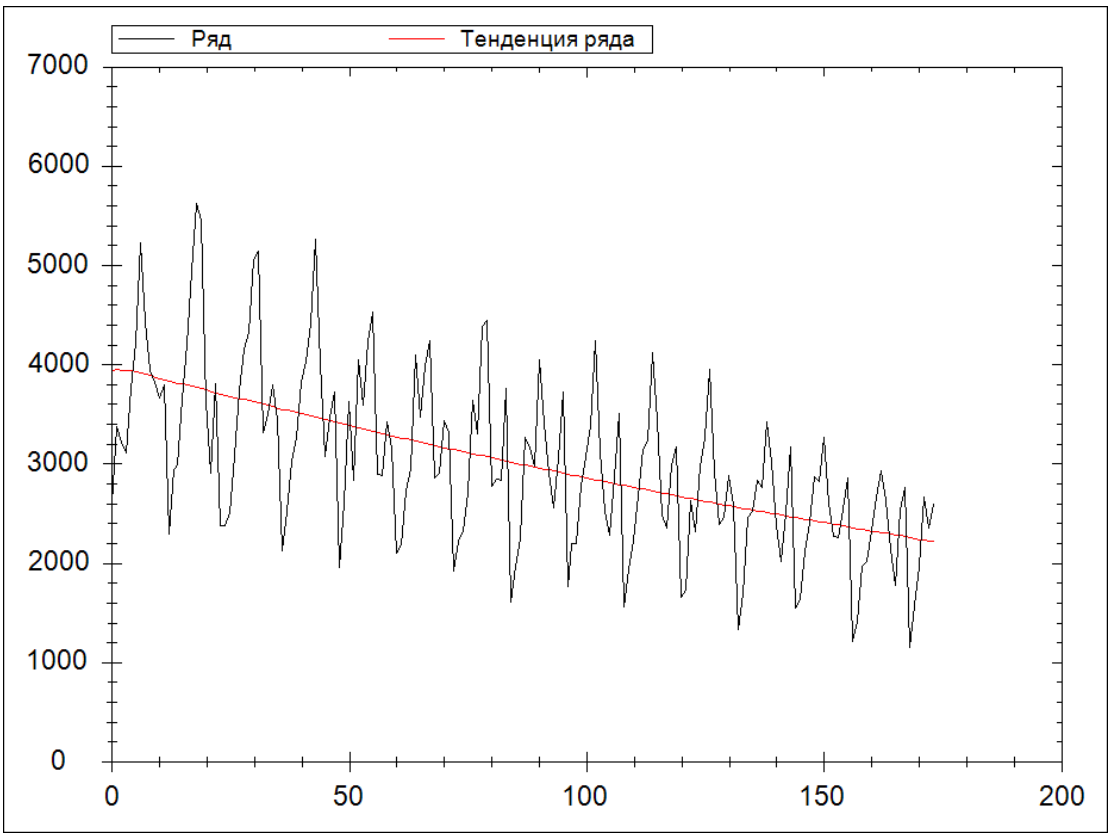

Рисунок 3.2.9 – Тренд и исходный ряд

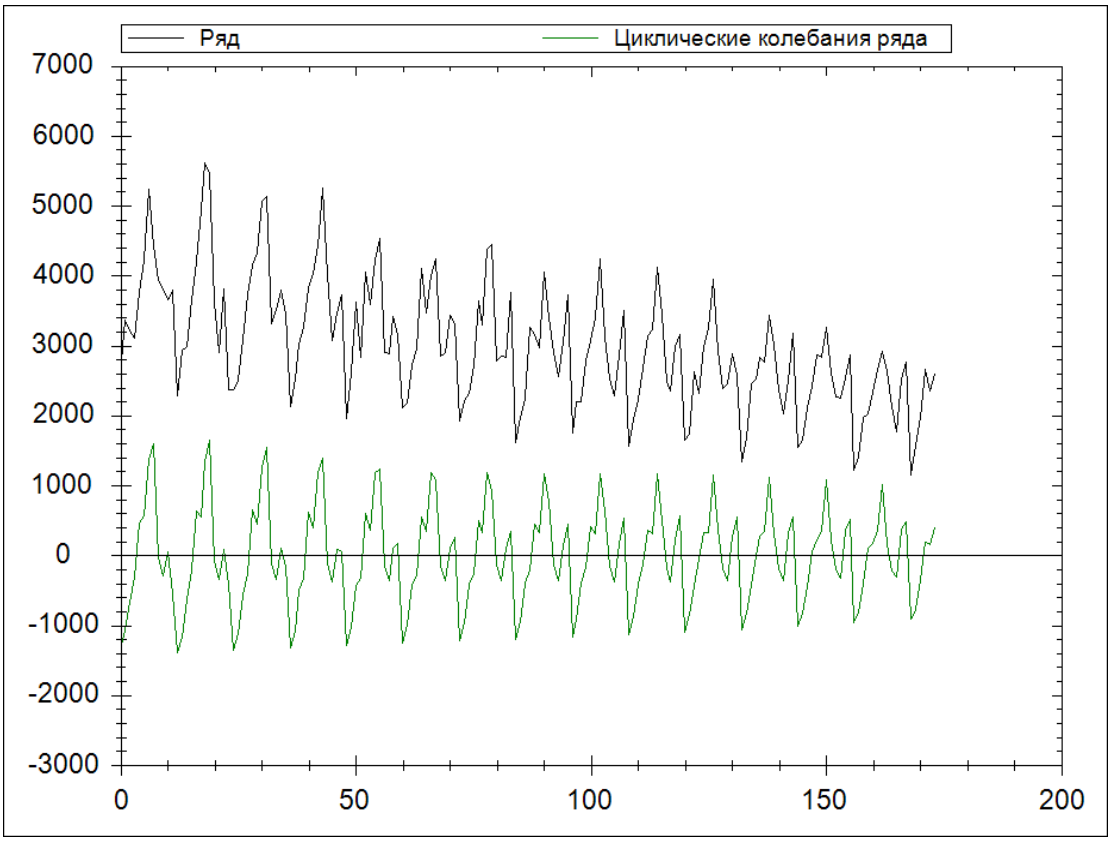

Рисунок 3.2.10 – Гармоника и исходный ряд

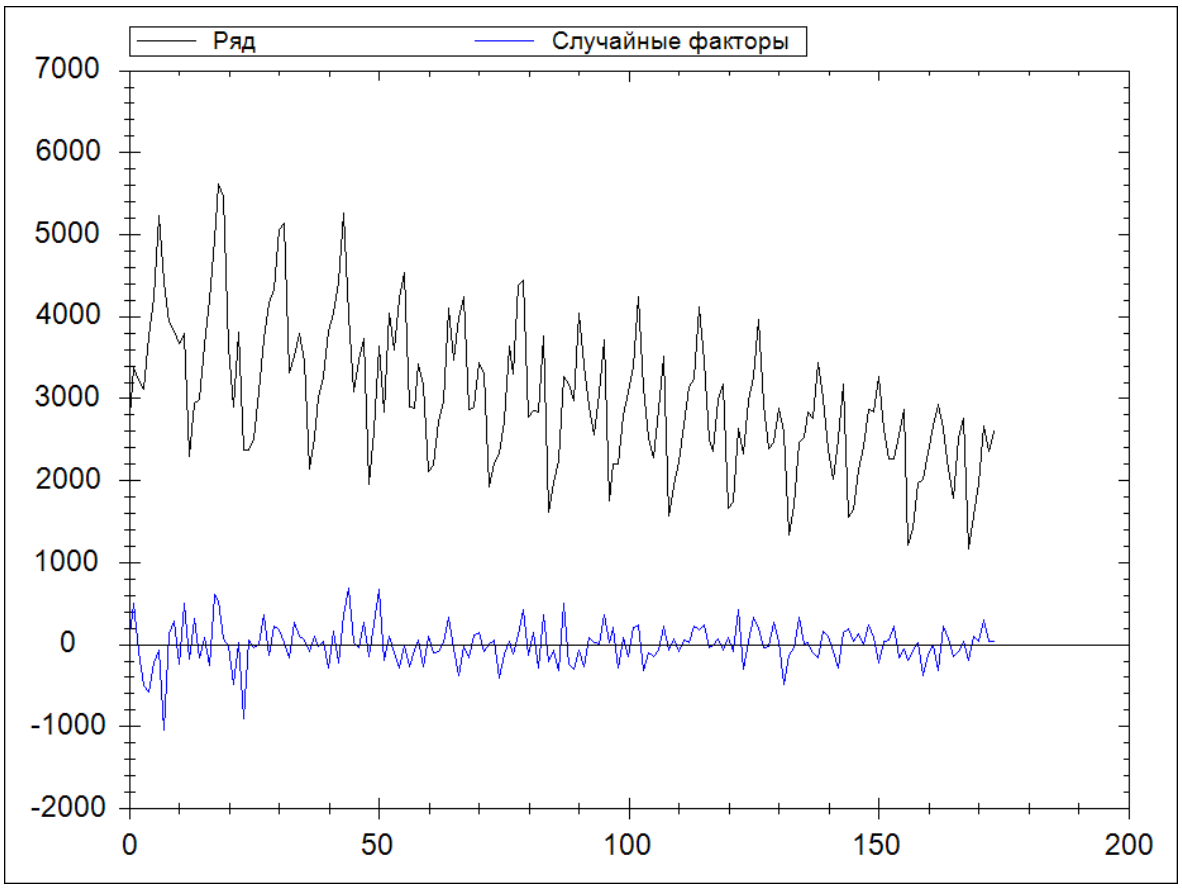

Рисунок 3.2.11 – Шум и исходный ряд

Далее исключая шумовую составляющую спрогнозируем временной ряд на 24 шага вперед, т.е. на 2 года. Полученный результат на рисунке 3.2.12. Также на рисунке можно наблюдать как произошла апроксимация исходного временного ряда трендовой и гармонической состовляющей(красная линия).

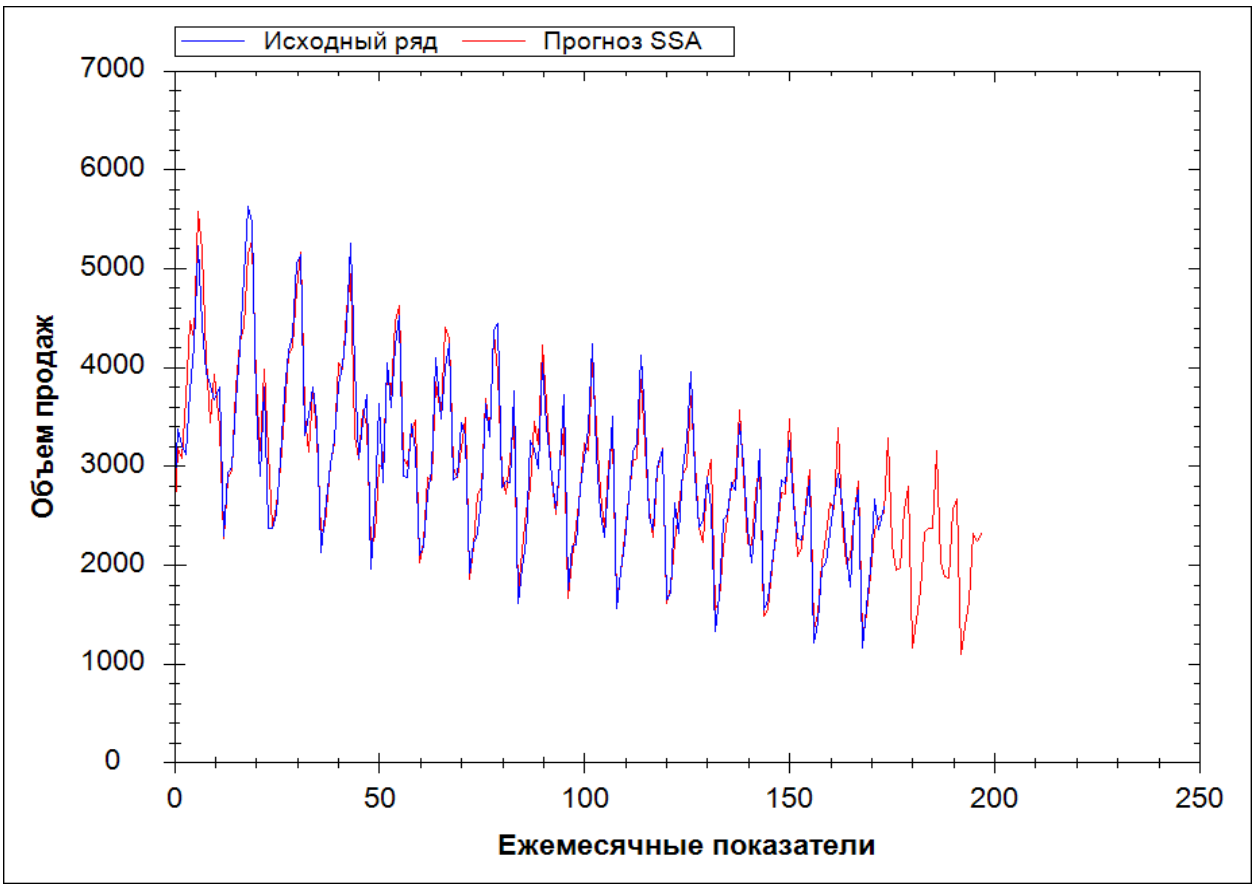

Рисунок 3.2.12 – Прогноз

Данный метод может решать задачи сглаживания ряда за счет метода главных компонент. На рисунке 3.2.2 показано, что 99.99% информации о ряде несут в себе 71 собственный вектор. На рисунке 3.2.13 видно, что исходный ряд аппроксимирован полностью, но прогноз будет не верен.

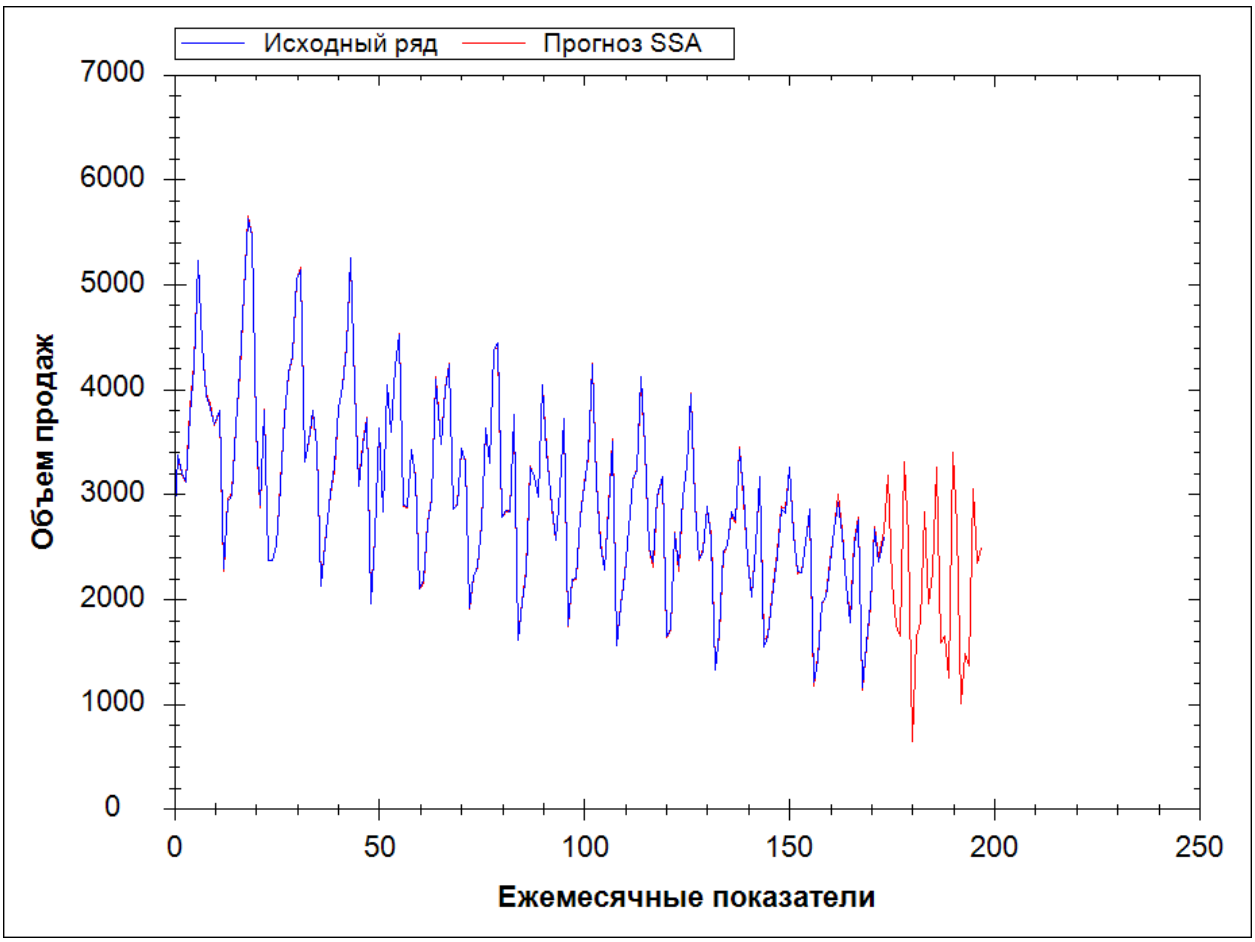

Рисунок 3.2.13 – Прогноз, используя 71 собственный вектор

На рисунке 3.2.14 показано, что 98.5% информации о ряд несут в себе 5 собственных векторов. На рисунке 3.2.15 показано как эти 5 векторов восстановили ряд и сделали прогноз. Т.е. размерность ряда уменьшилась во много раз с минимальными потерями информации.

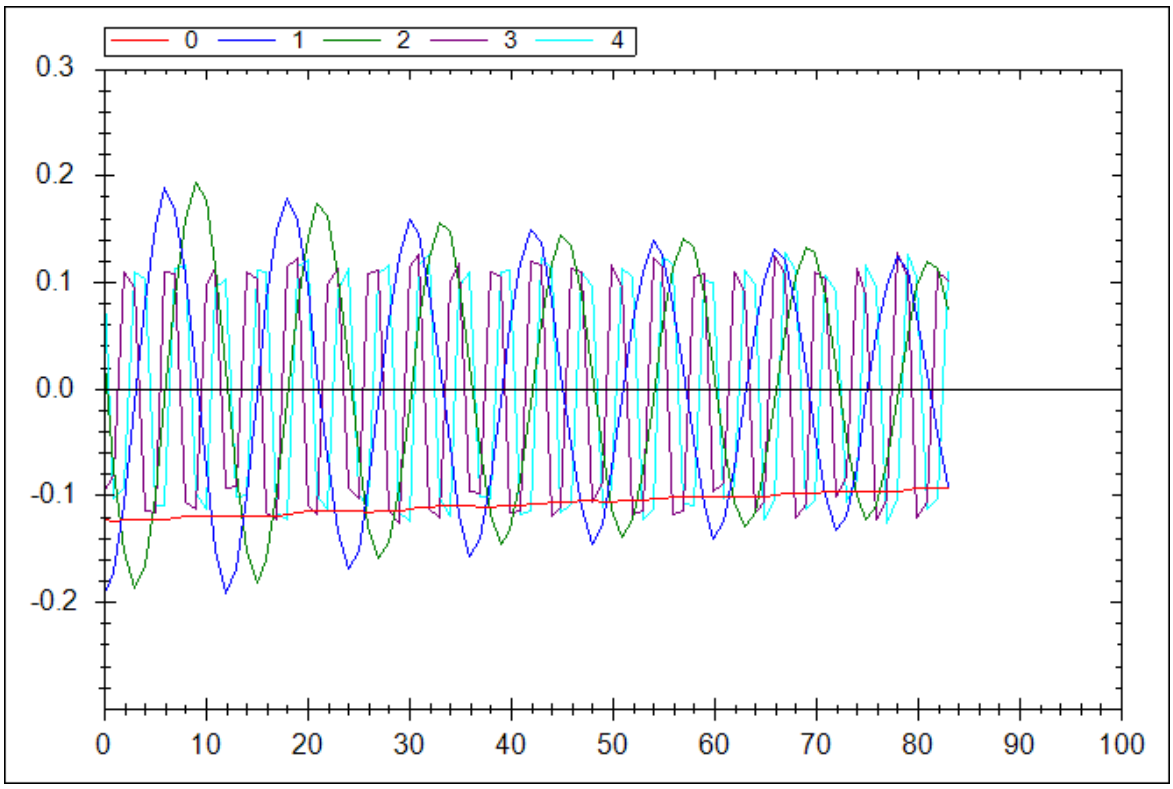

Рисунок 3.2.14 – Собственные вектора(98.5% информации)

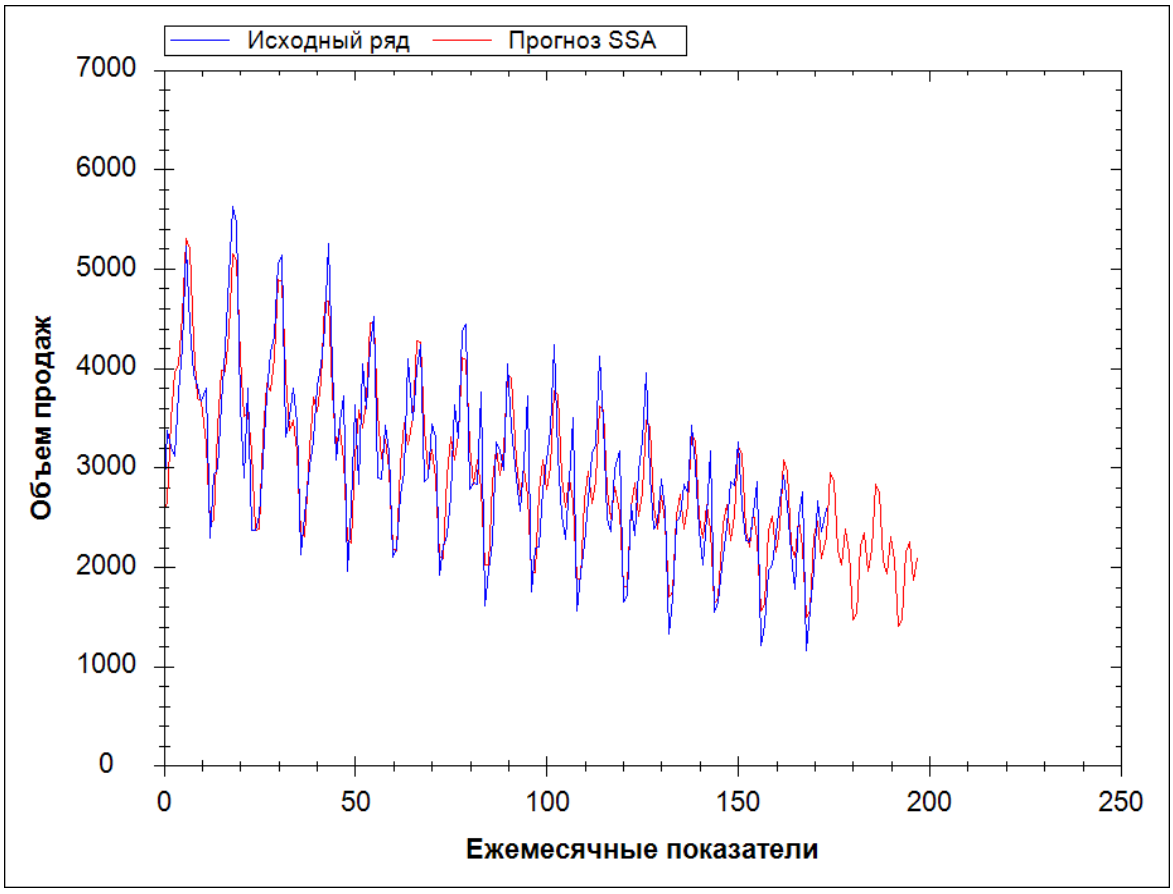

Рисунок 3.2.15 – Восстановленный временной ряд по 5 собственным векторам

3.3 Временной ряд «синус» и «зашумленный синус» с постоянной частотой

На рисунке 3.3.1 показан синус и восстановленный временной ряд. Метод определил только одну собственную тройку сингулярного разложения поэтому аппроксимировать и сделать адекватный прогноз не представляется возможным.

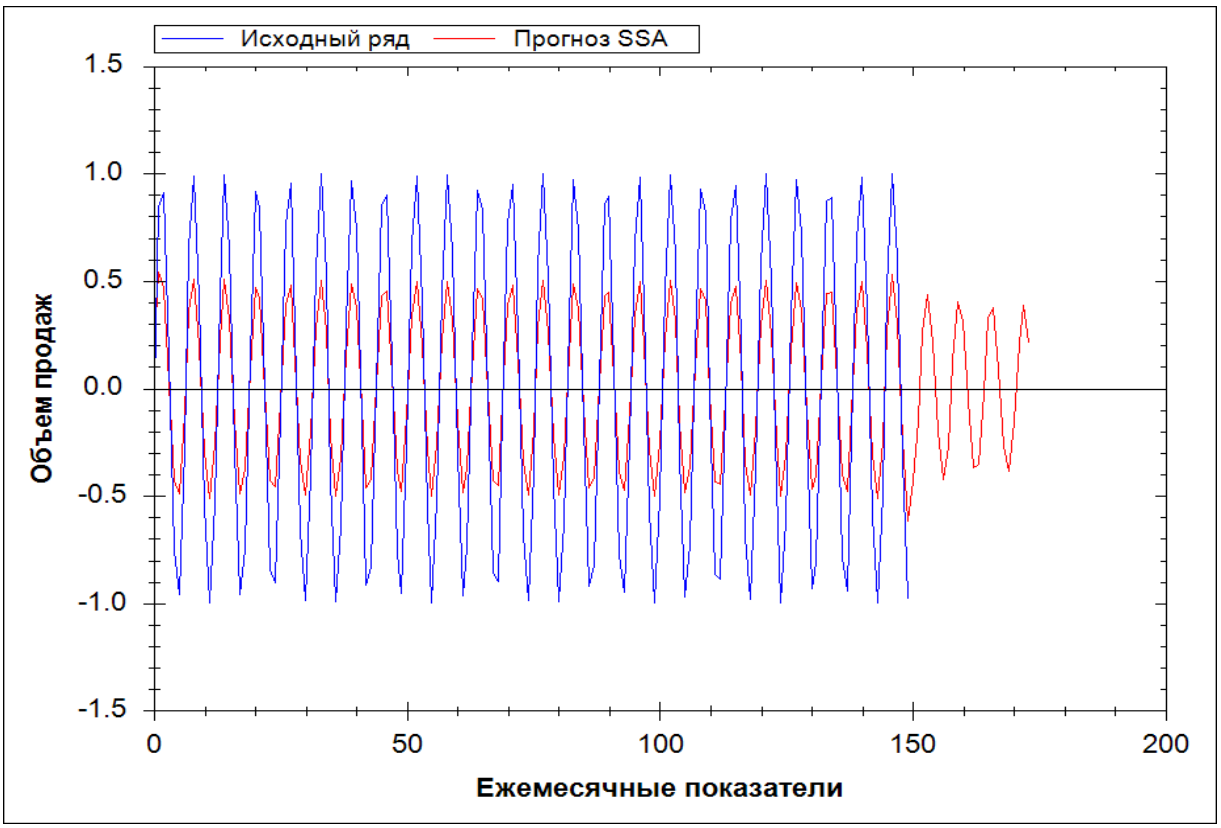

Рисунок 3.3.1 –Синус

Далее зашумим синус. На рисунке 3.3.2 показан восстановленный временной ряд зашумленного синуса по двум собственным векторам(рисунок 3.3.3). Зависимость векторов показана на скатеграмме (рисунок 3.3.4).

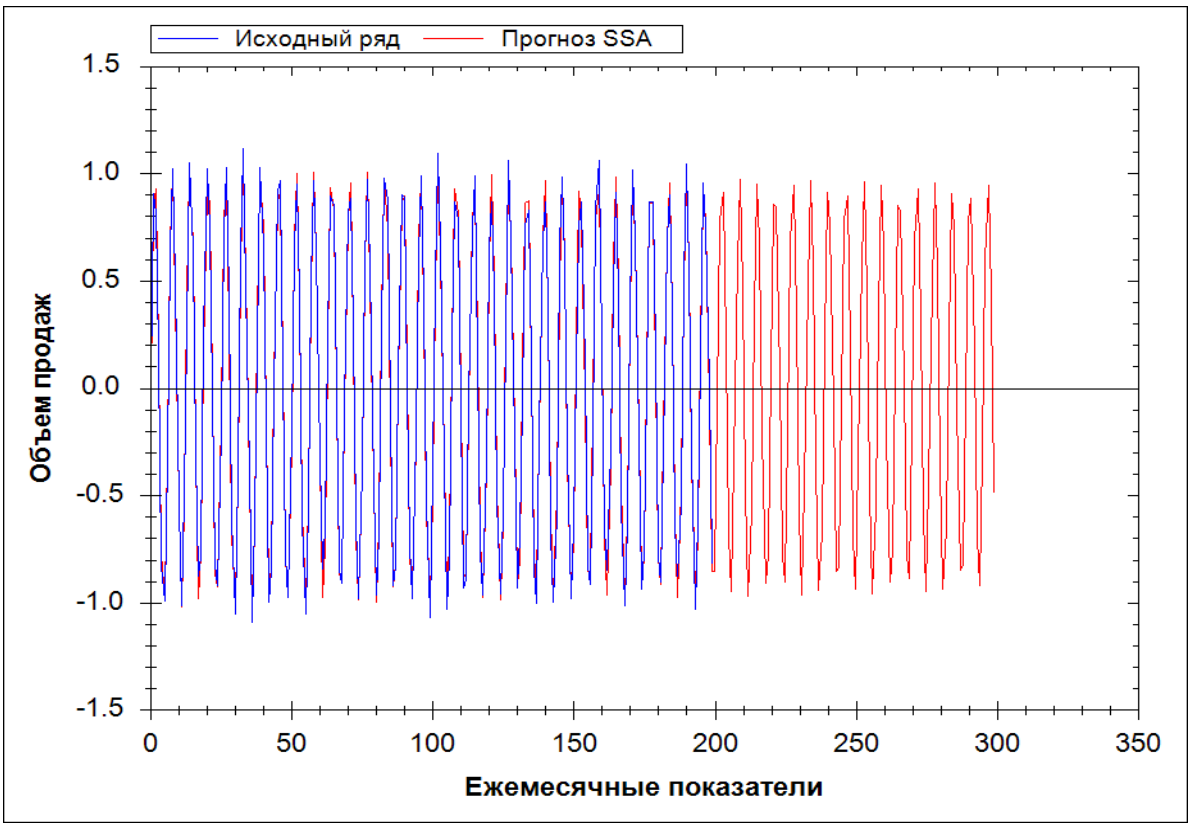

Рисунок 3.3.2 – Зашумленный синус

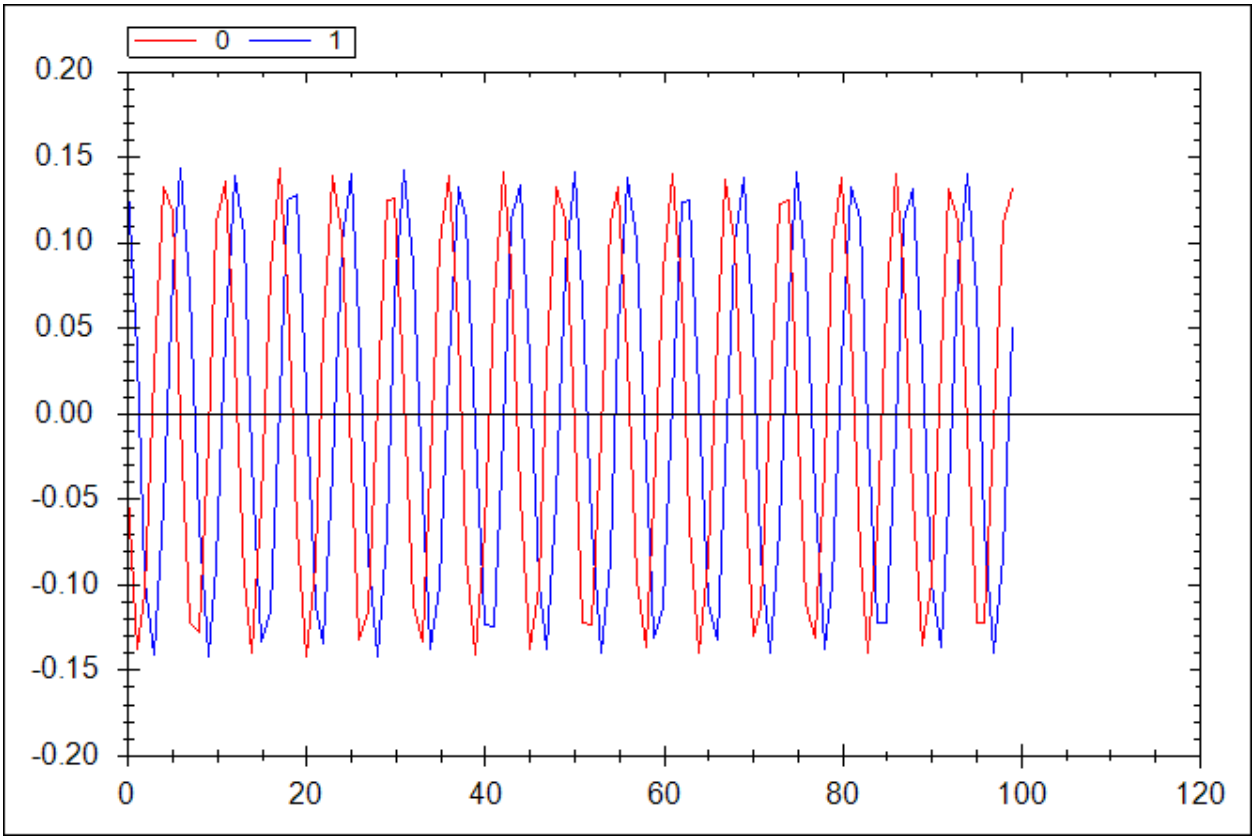

Рисунок 3.3.3 – Собственные вектора соответствующие гармонической составляющей

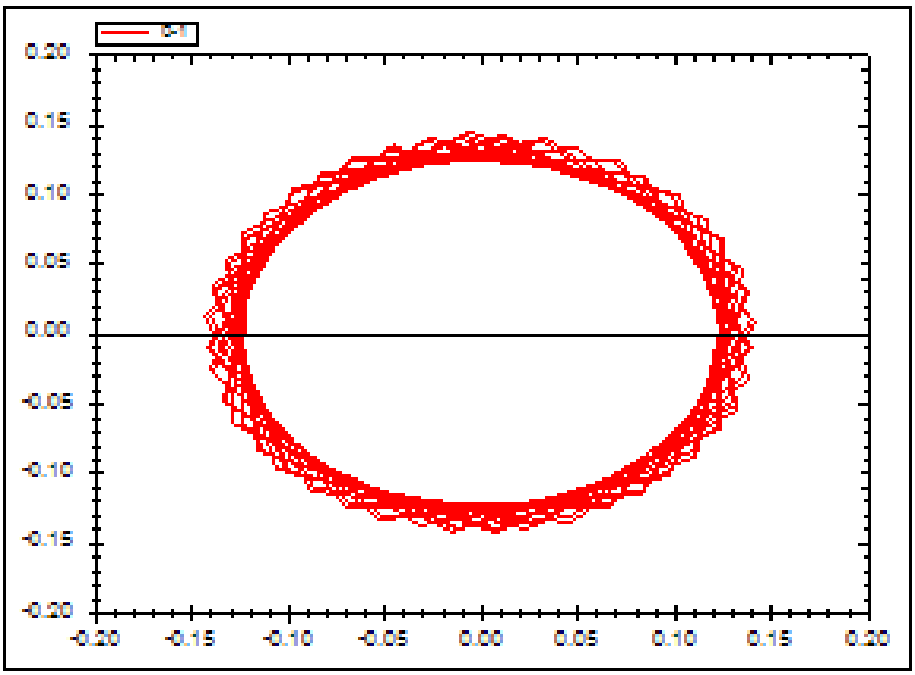

Рисунок 3.3.4 – Скаттеграмма собственных векторов

3.4 Временной ряд «синус» с разной частотой

Сложной задачей для методов анализа временных рядов является анализ рядов с изменяющейся частотой. На рисунке 3.4.1 дан ряд с изменяющейся частотой.

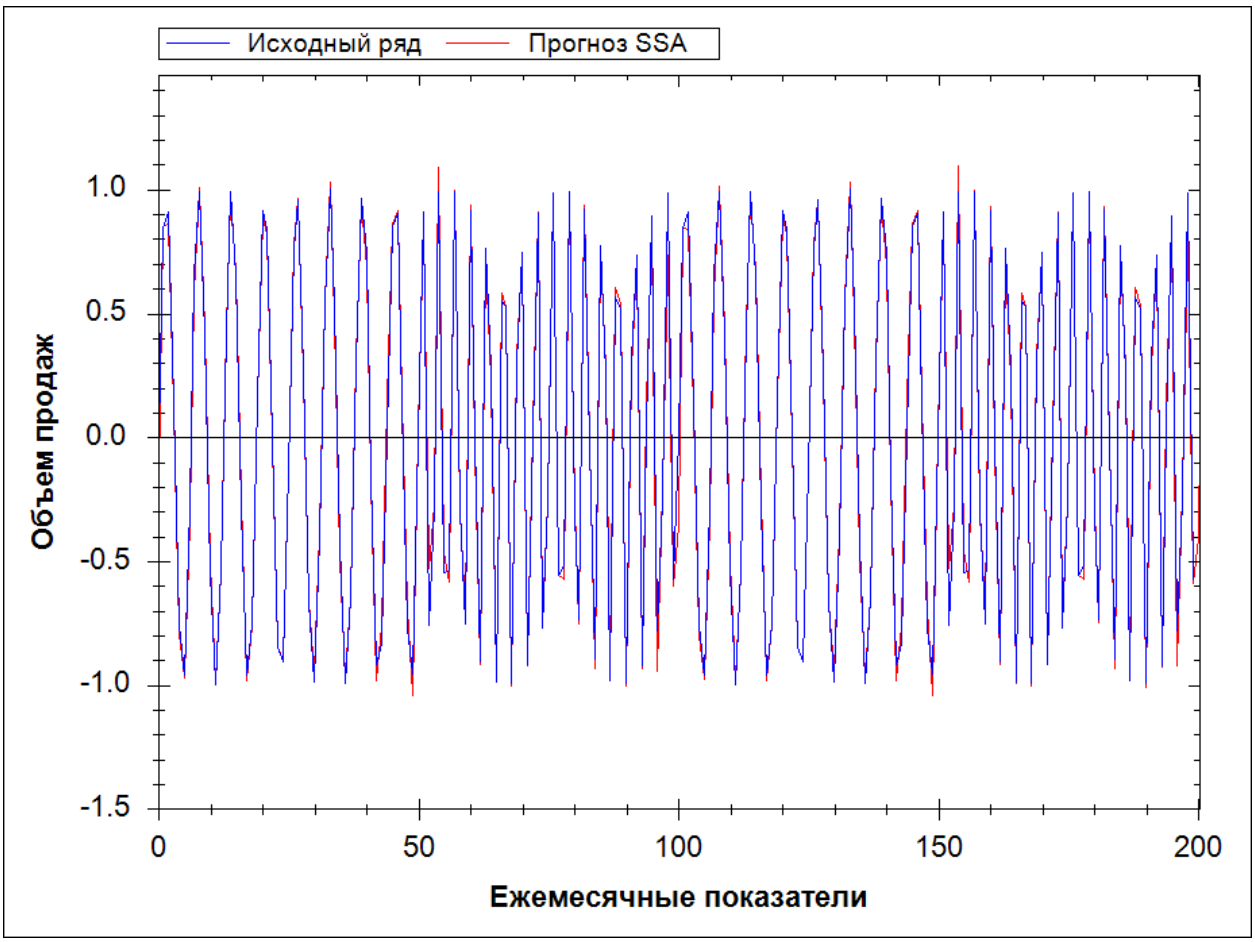

Рисунок 3.4.1 – Синус с изменяющейся частотой

По виду собственных чисел (рисунок3.4.2) и скаттеграмм (рисунок 3.4.3) делаем вывод о том, что тренда нет, а ET0,1,2,3; ET4,5; ET6,7; ET8,9 являются гармониками, остальные вектора примем за шум.

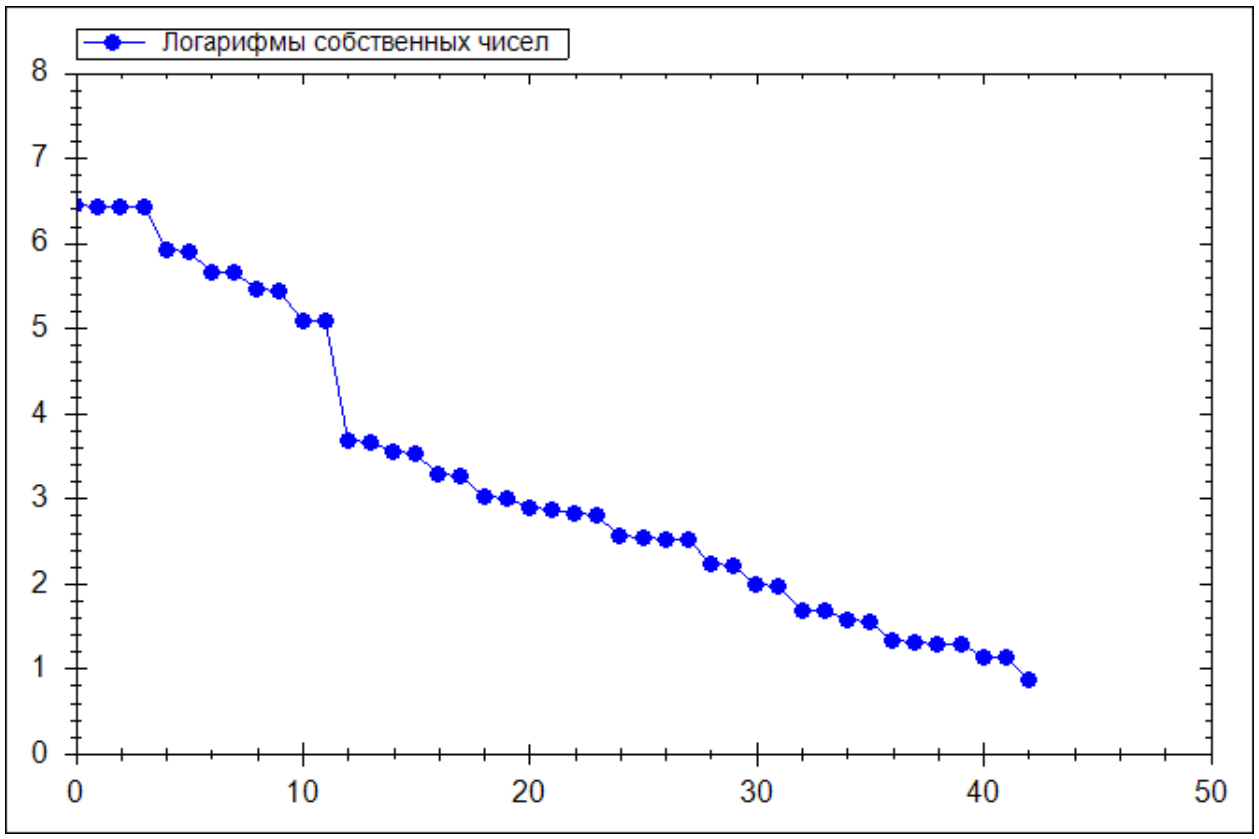

Рисунок 3.4.2 – Собственные числа

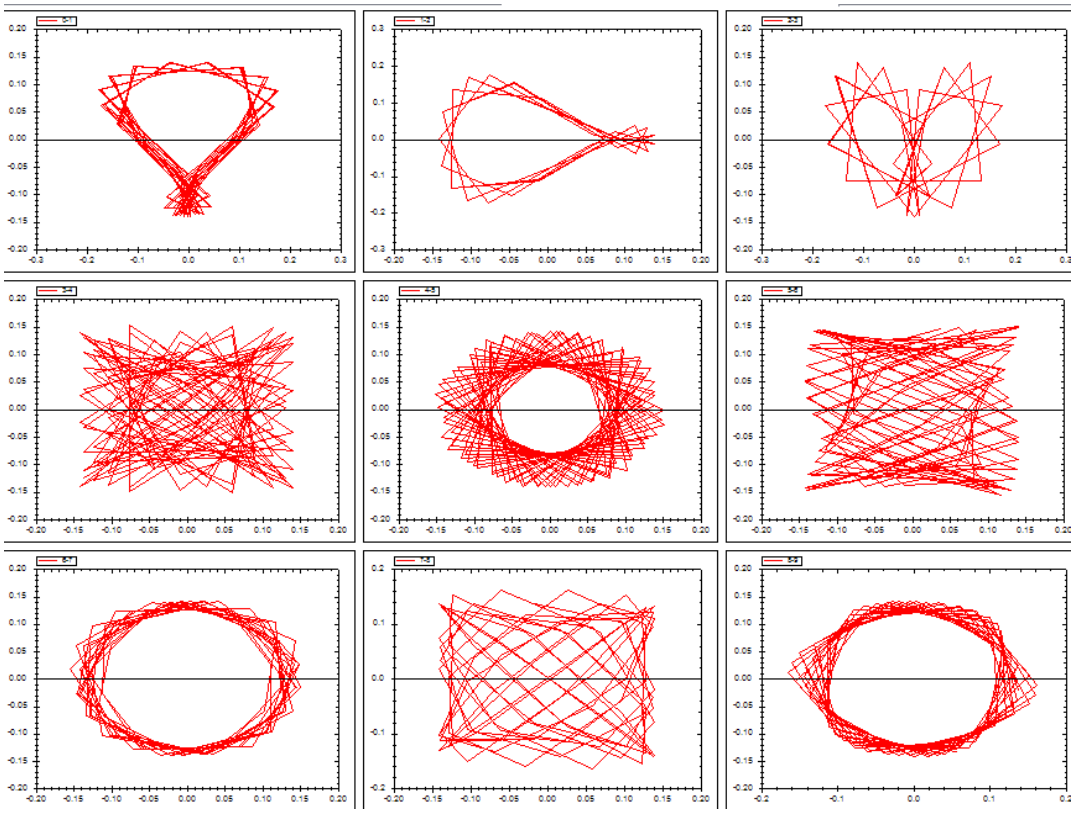

Рисунок 3.4.3 – Скаттеграммы

Прогноз сделанный по взятым составляющим свидетельствует о том что метод SSA хорошо справляется с данными типами временных рядов (рисунок 3.4.4).

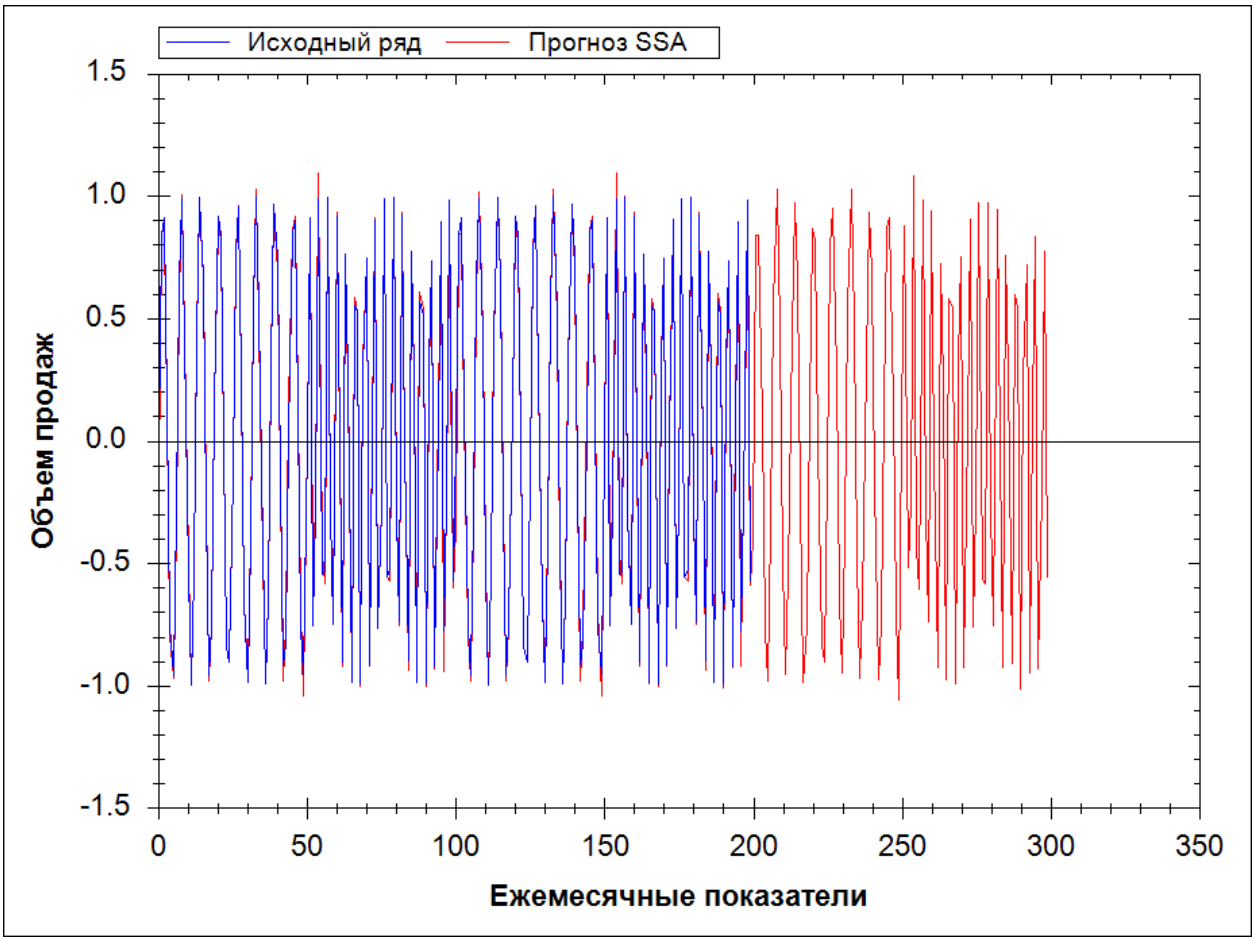

Рисунок 3.4.4 – Прогноз на 100 точек

3.5 Проверка построения прогноза

Был взят исходный ряд «Ford» из которого удалили данные за последние 24 месяца, т.е. 24 точки. Построили прогноз по оставшейся выборки. Результат можно видеть на рисунке 3.5.1. Ошибка прогноза составила 15%.

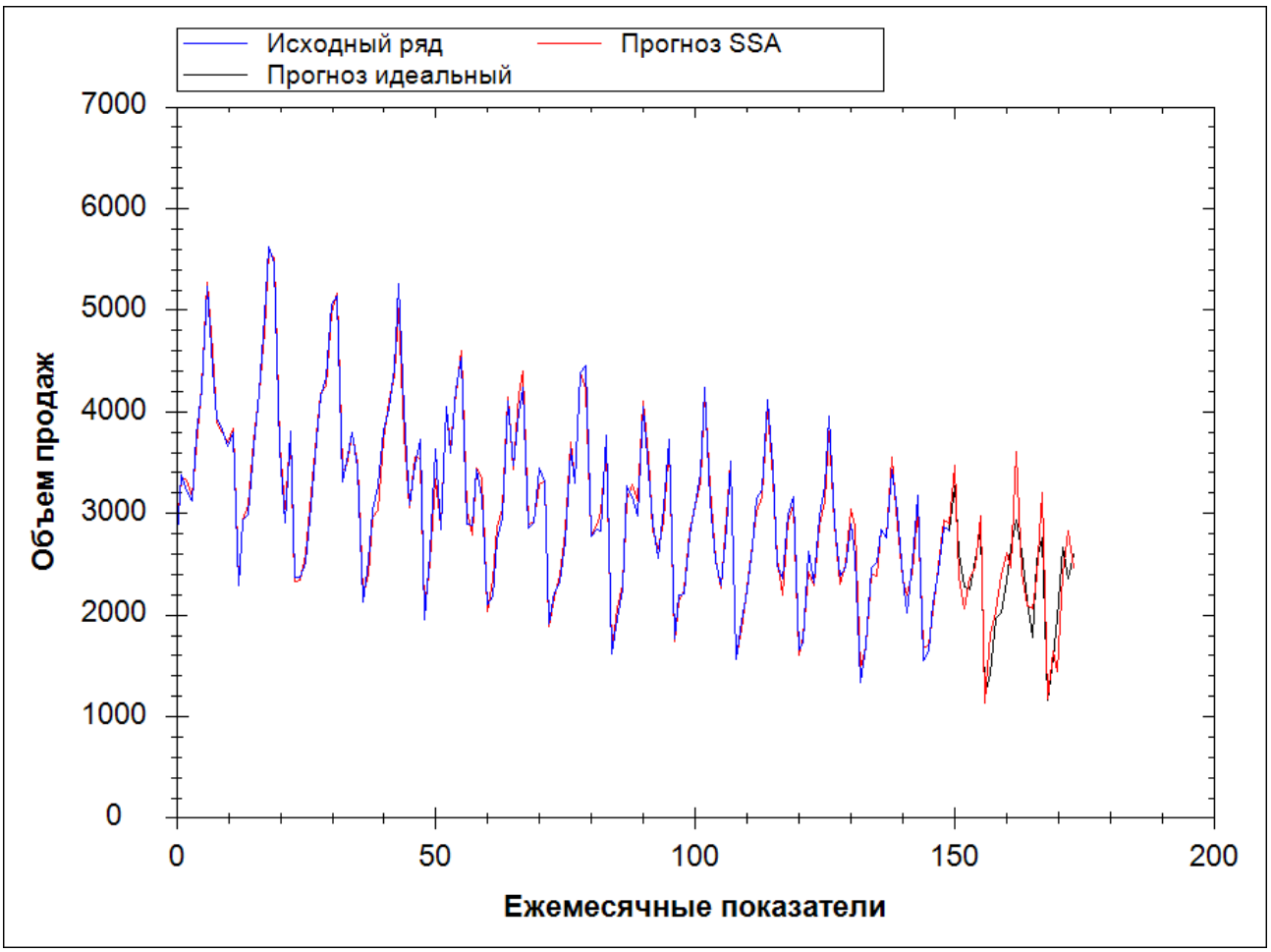

Рисунок 3.5.1 – Ошибка прогноза составила 15%

ЗАКЛЮЧЕНИЕ

В ходе данной бакалаврской работы был изучен большой объем информации по анализу временных рядов, в частности метода SSA.

В ходе изучения материала был сделан вывод, что данный метод имеет некоторые преимущества перед другими методами:

1) Может быть использован без предварительного задания модели ряда;
2) Может работать с нестационарными рядами;
3) В отличие от анализа Фурье, где рассматривается фиксированный базис из синусов и косинусов, SSA использует адаптивный базис, порождаемый самим рядом. В результате, SSA может выделять амплитудно-модулированные синусы и косинусы с частотами;
4) SSA не требует предварительного задания параметрической модели, что может дать значительное преимущество, когда нет очевидной модели. В частности, SSA позволяет выделять периодичности без знания значений периодов;

Диапазон областей знаний, где SSA может быть применен, очень широк: климатология, океанология, геофизика, техника, обработка изображений, медицина, эконометрика и многие другие.

Метод SSA позволяет:

1) различать составляющие временного ряда, полученные из последовательности значений какой-либо величины, взятой через равные промежутки времени;
2) находить заранее неизвестные периодичности ряда;
3) сглаживать исходные данные на основе отобранных составляющих;
4) наилучшим образом выделять компоненту с заранее известным периодом;

5) предсказывать дальнейшее поведение наблюдаемой зависимости.

Для автоматизации процесса группировки сингулярных троек по группам тренд, сезонная составляющая, шум был разработан алгоритм классификации, используя при этом собственные числа сингулярных троек.

Также для исследования алгоритма была осуществлена программная реализация данного метода с возможностью создания тестовых выборок.

СПИСОК ИСТОЧНИКОВ

1) Голяндина, Н. Э. Метод «Гусеница»-SSA: анализ временных рядов: учебное пособие / Н. Э. Голяндина. – Санкт-Петербург: 2004.

2) Жиглявский, А. А. Главные компоненты временных рядов: метод «Гусеница»: учебное пособие/ А.А. Жиглявский. – Санкт-Петербург: 1997.

3) Ефимов В. М. Анализ и прогноз временных рядов методом главных компонент / Ю. К. Галактионов, Н. Ф. Шушпанова - Новосибирск: Наука, 1988. −70с.

4) Ефимов В. М. О возможности прогнозирования циклических изменений численности млекопитающих / Галактионов Ю. К. 1983. № 3, с.343-352.

5) Голяндина, Н. Э. Метод «Гусеница»-SSA: прогноз временных рядов: учебное пособие / Н.Э. Голяндина. – Санкт-Петербург : Изд-во СПбГУ – 2004.