**Предварительная обработка данных**

**1. Исходные данные**

Значения временного ряда y(t), t=0,1,2,…,m – годовой курс валюты (дирхама к российскому рублю)

**2. Цель работы**

Предварительная обработка и анализ данных.

**3. Задание**

1. Удаление выбросов. Выявить выбросы и восстановить значения (2 метода в зависимости от варианта). Методы:

* 2. Z-оценка;
* 7. Local Outlier Factor (LOF).

2. Фильтрация шума.

Вычислить статистику Бокса-Пирса для тестирования на наличие белого шума.

Очистить данные от шума алгоритмами:

* 5. Фильтр Чебышева 2-го рода;

3. Сглаживание значений ряда

* 3. Алгоритм экспоненциального скользящего среднего;

4. Проверить ряды на стационарность до предварительной обработки (исходные данные) и после выполнения п.1-3, используя тест Дикки-Фуллера (ADF-тест), тест Квятковского–Филлипса–Шмидта–Шина (KPSS- тест).

5. Проверить ряды на наличие тренда (1 метод в зависимости от варианта). Метод:

* 1. Метод проверки разностей средних уровней;

**4. Ход работы**

Временной ряд y(t):

|  |  |
| --- | --- |
|  |  |

**4.1. Удаление выбросов. Выявить выбросы и восстановить значения. Методы:**

Z-оценка

В статистике Z-оценка говорит нам о том, на сколько стандартных отклонений заданное значение отличается от среднего. Используется следующую формулу для расчета Z-оценки:

где:

* - это одно необработанное значение данных.
* - среднее значение
* - стандартное отклонение

Если количество элементов в данном наборе велико, то около 68% элементов имеют Z-оценку от -1 до 1; около 95% имеют Z-оценку от -2 до 2; около 99% имеют Z-оценку от -3 до 3. Это известно как эмпирическое правило, и оно определяет процент данных в пределах определенных стандартных отклонений от среднего значения при нормальном распределении. Так как мы знаем, что 99,73% наблюдений лежат в пределах трех стандартных отклонений от среднего, то можем предположить, что выбросами будут оставшиеся 0,27%.

Результаты: Во временном ряду выбросов не обнаружено

Local Outlier Factor (LOF)

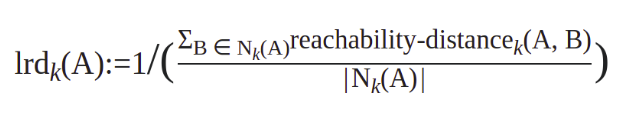
Алгоритм Local Outlier Factor (LOF) — это неконтролируемый метод обнаружения аномалий, который вычисляет отклонение локальной плотности данной точки данных по отношению к ее соседям. Он считает выбросами образцы, которые имеют значительно более низкую плотность, чем их соседи.

Сначала нужно определить K-расстояние точки. Интуитивно это расстояние от K-го ближайшего соседа к точке.

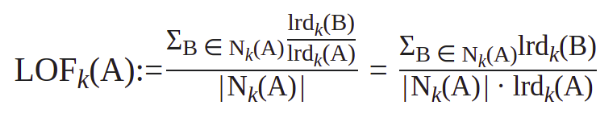
Окрестность точки K — это просто K ближайших точек к данной точке.

Расстояние достижимости определяет, насколько легко добраться до точки из другой точки. Если точка А находится в К-окрестности В, то легче достичь А, а расстояние достижимости меньше. Для точки А, которая не находится в К-окрестности В, расстояние достижимости выше.

Локальная плотность достижимости определяется как величина, обратная среднему расстоянию достижимости точек вокруг точки A до A:



Теперь, когда мы знаем плотность точки, локальный коэффициент достижимости — это просто среднее отношение плотностей соседних точек А к А.



Принимая отношение плотности к соседним точкам, а не сами фактические плотности, метод LOF может работать с наборами данных с областями разной плотности.

Результаты: Во временном ряду выбросов не обнаружено

Восстановление значений:

Метод замены выбросов зависит от задачи и типа данных. Наиболее распространенный метод замены выбросов - это замена на медиану или среднее значение в выборке или среди соседних значений.

Во временном ряду выбросов не обнаружено не по 1, не по 2 тесту, поэтому замена выбросов не требуется

**4.2. Фильтрация шума**

Q-статистика Бокса-Пирса — статистический критерий, предназначенный для нахождения автокорреляции временных рядов. Вместо тестирования на случайность каждого отдельного коэффициента, он проверяет на отличие от нуля сразу несколько коэффициентов автокорреляции. Статистика Бокса-Пирса вычисляется по следующей формуле

где – число наблюдений, – автокорреляция -го порядка, – число проверяемых лагов

При больших приближённо имеем . Нулевая гипотеза отвергается при выполнении неравенства . Нулевая гипотеза заключается в том, что рассматриваемый процесс является белым шумом

Статистика Бокса-Пирса и соответсвующее ей p-value для лагов от 1 до 10:

|  |
| --- |
|  |

Критическое значение , что меньше значений статистики. Их этого следует, что нулевая гипотеза отвергается, значит процесс не является белым шумом

Фильтр Чебышева 2-го рода

Фильтр Чебышева 2-го рода минимизирует абсолютную разницу между идеальной и фактической частотной характеристикой по всей полосе пропускания за счет включения равных пульсаций в полосу пропускания

АЧХ фильтра Чебышева второго рода описывается следующим образом:

Здесь - частота среза, — полином Чебышева -го порядка, - порядок фильтра, - параметр, определяющий величину пульсаций АЧХ в полосе задерживания.

Значение параметра и уровень пульсаций (в децибелах) связаны следующим образом:

Очистка данных от шума



**4.3. Сглаживание значений ряда**

Алгоритм экспоненциального скользящего среднего

Сглаживание временного ряда – это замена фактических уровней расчетными значениями, имеющими меньший разброс, чем исходные

Формула экспоненциального сглаживания:

где – среднее прогнозное значение будущего периода, рассчитываемое с учетом экспоненциального сглаживания; - фактические данные текущего периода; - прогнозное значение показателя текущего периода; - постоянная сглаживания.

Хотя в принципе может принимать любые значения из диапазона , обычно ограничиваются интервалом от 0.2 до 0.5.

Сглаживание данных



**4.4. Проверка рядов на стационарность**

Временной ряд считается стационарным, если его среднее значение, дисперсия и структура автокорреляции не меняются со временем. Другими словами, стационарный временной ряд будет иметь устойчивый статистический тренд во времени

1) Расширенный тест Дики-Фуллера - статистический тест, который оценивает наличие единичного корня в одномерном временном ряду, что является ключевым показателем нестационарности

Гипотезы для расширенного теста Дики-Фуллера (ADF):

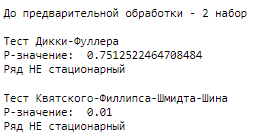
* Нулевая гипотеза (H0): временной ряд нестационарен, потому что существует единичный корень (если p-значение > 0,05).
* Альтернативная гипотеза (H1): временной ряд является стационарным, так как отсутствует единичный корень (если p-значение ≤ 0,05).

2) Тест Квятковского-Филлипса-Шмидта-Шина - еще один тест на стационарность, дополняющий тест ADF. Он используется для определения того, является ли временной ряд стационарным относительно детерминированного тренда.

Гипотезы для теста Квятковского-Филлипса-Шмидта-Шина (KPSS) таковы :

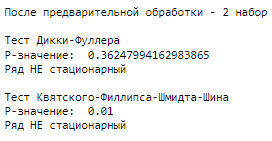
* Нулевая гипотеза (H0): временной ряд стационарен, так как отсутствует единичный корень (если значение p > 0,05).
* Альтернативная гипотеза (H1): временной ряд не является стационарным, поскольку существует единичный корень (если p-значение ≤ 0,05).

Тесты на стационарность до предварительной обработки:



По результатам тестирования на стационарность можно сделать вывод, что до предварительной обработки временной нестационарный

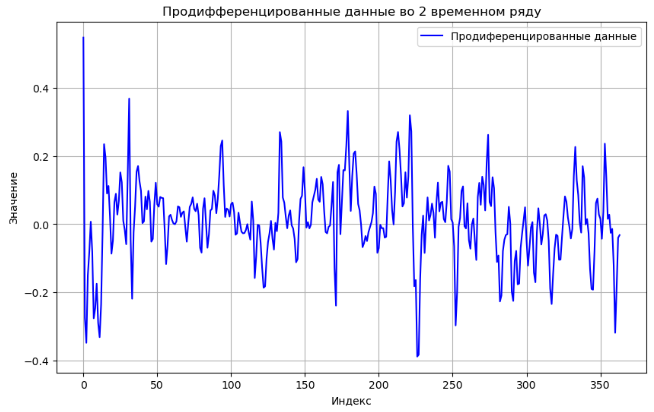
Тесты на стационарность после предварительной обработки:



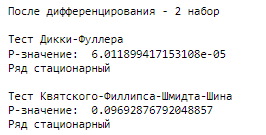
После предварительной обработки можно сделать вывод, что временной ряд нестационарный

Некоторые модели прогнозирования временных рядов требуют стационарных временных рядов, потому что их легче моделировать из-за их постоянных статистических свойств. Таким образом, нужно сделать временной ряд стационарным, если это не так

Для того, чтобы перейти к стационарному ряду часто применяют операцию дифференцирования. Операция дифференцирования или первая конечная разность ряда обозначается как



Результат тестов на стационарность:



После дифференцирования тесты показали, что временной ряд стал стационарным

После выбора соответствующей модели для прогнозирования, требуется обратно преобразовать прогнозные значения, проведя обратное дифференцирование с учетом исходных данных. Если ряд был дифференцирован один раз, обратное дифференцирование будет представлять собой кумулятивное суммирование разностей (пошаговое суммирование)

**4.5. Проверка рядов на наличие тренда**

Метод сравнения средних применяется дли выявления монотонно возрастающей или монотонно убывающей тенденции. Временной ряд разбивается на две примерно одинаковые части и и вычисляются средние и выборочные дисперсии для обеих частей соответственно

Далее расчитывается значение критерия Стьюдента по формуле:

если предполагается, что значения дисперсий на этих участках не равны между собой, и по формуле:

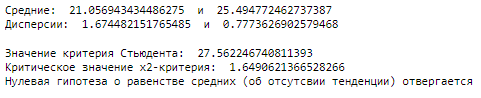
где – общая выборочная дисперсия ряда, если предполагается, что дисперсии одинаковы

Нулевая гипотеза о равенстве средних (об отсутсвии тенденции) отвергается, если выполняется условие , где – табличное значение -критерия Стьюдента при уровне значимости и числе степеней свободы

Проверка на наличие тренда обычно проводится перед принятием решения о дифференцировании ряда.

Если у временного ряда есть тренд, то это может означать, что в данных присутствует систематическое изменение с течением времени, и это может создать проблемы при моделировании ряда. Дифференцирование ряда может помочь устранить тренд и сделать временной ряд стационарным

В связи с этим проведём проверку на наличие тренда на данных до дифференцирования:



Во 2 временном ряду тренд присутсвует

**5. Вывод**

Результаты предварительной обработки:



В ходе работы были выполнены предварительная обработка и анализ данных.

В первую очередь были выявлены выбросы и восстановлены их значения. Поиск выбросов был выполнен двумя методами: Z-оценка и Local Outlier Factor (LOF).

Статистические методы могут быть полезны при работе с данными, в которых присутствует нормальное распределение и выбросы соответствуют нормальному закону распределения. Методы, основанные на моделях машинного обучения, могут быть эффективны в обнаружении выбросов в данных, в которых присутствуют сложные и нелинейные закономерности. Однако эти методы могут быть сложны в использовании и требуют дополнительной настройки параметров.

В целом, выбор между статистическими методами и методами машинного обучения для выявления выбросов в данных зависит от конкретной задачи, характеристик данных и уровня сложности выбросов. Часто используется комбинированный подход, при котором различные методы применяются вместе для более точного определения выбросов.

Дальше была произведена фильтрафия данных от шума. Для этого была вычислена статистика Бокса-Пирса для определения наличия белого шума, а после была произведена очистка данных с помощью фильтра Чебышева 2-го рода.

Было проведено сглаживание значений ряда алгоритмом скользящего среднего

При проверке рядов на стационарность двумя тестами (ADF и KPSS) было выявлено, что временной ряд нестационарный, поэтому была применена операция дифференцирования

Проверка на наличие тренда методов проверки разностей средних уровней показала, что во временном ряду есть тренд (проверка до дифференцирования)

**Построение регрессионных, авторегрессионных моделей и моделей в пространстве состояний**

**1. Исходные данные**

Значения временного ряда y(t), t=0,1,2,…,m – годовой курс валюты (дирхама к российскому рублю)

**2. Задание**

1. Построить и исследовать точность интегрированной модели авторегрессии скользящего среднего (ARIMA). Модель использует три основных параметра (p, d, q), которые выражаются целыми числами. Потому модель также записывается как ARIMA(p, d, q).

* p – порядок авторегрессии (AR), который позволяет добавить предыдущие значения временного ряда.
* d – порядок интегрирования (порядок разностей исходного временного ряда). Он добавляет в модель понятия разности временных рядов (определяет количество прошлых временных точек, которые нужно вычесть из текущего значения).

Для нестационарного временного ряда устанавливается параметр d=1, для стационарного d=0.

* q – порядок скользящего среднего, который позволяет установить погрешность модели как линейную комбинацию наблюдавшихся ранее значений ошибок.

Исследовать различнyю параметризацию модели ARIMA (параметры определить на основе критерия Акаике), установить оптимальные значения параметров для различных длин мерных интервалов. Результаты представить в виде таблицы.

2. Построить и исследовать точность модели структурного временного ряда (BSTS). Исследовать различнyю параметризацию модели, установить оптимальные значения параметров для различных длин мерных интервалов. Результаты представить в виде таблицы из п.1.

3. Разложение ряда на компоненты

3.1. Выделить

* трендовую составляющую на основе полиномиальной регрессионной модели (можно использовать линейную, квадратичную, экспоненциальную функцию, линейную комбинацию многочленов (в т.ч. многочленов Чебышева));
* сезонную (гармоническую) составляющую на основе Фурье-анализа;
* остаточную составляющую (проверить имеют ли остатки нормальное распределение).

3.2. Сравнить полученные в п.3 результаты с аддитивной нелинейной регрессионной моделью (пакет Prophet).

Исследовать точность моделей из п.3.1 и 3.2 для различных длин мерных интервалов. Результаты представить в виде таблицы из п.1.

**3. Ход работы**

1) Модель ARIMA

Модель ARIMA (AutoregRessive Integrated Moving Average) – один из наиболее распространённых методов анализа и прогнозирования временных рядов. Эта модель позволяет обработать данные временного ряда, чтобы лучше понять этот ряд или предсказать его развитие.

Информационный критерий Акаике (AIC) - критерий для выбора лучшей из нескольких статистических моделей, построенных на одном и том же наборе данных и использующих логарифмическую функцию правдоподобия. Критерий позволяет найти компромисс между сложностью модели (числом параметров) и ее точностью. В общем случае AIC вычисляется по формуле:

где - число параметров модели, - максимизированное значение функции правдоподобия модели. Лучшей признается та модель, для которой значение AIC минимально.

В 1 части работе был проведён тест на стационарность, который показал, что временной ряд НЕ является стационарным, поэтому зададим порядок интегрирования d = 1

Размерность 2 временного ряда равна 365

|  |  |  |  |
| --- | --- | --- | --- |
| Длина мерного интервала | p | q | Среднеквадратическое отклонение |
| 20 | 1 | 2 | 0.154578 |
| 2 | 2 | 0.151684 |
| 1 | 3 | 0.149484 |
| 2 | 1 | 0.309309 |
| 50 | 1 | 3 | 0.388169 |
| 2 | 3 | 0.325324 |
| 2 | 2 | 0.360564 |
| 3 | 2 | 0.162241 |
| 100 | 2 | 3 | 0.3796 |
| 3 | 3 | 0.363401 |
| 3 | 2 | 0.384406 |
| 2 | 2 | 0.881398 |
| 200 | 2 | 3 | 0.434202 |
| 3 | 2 | 0.32582 |
| 3 | 3 | 0.447849 |
| 2 | 2 | 0.560127 |
| 300 | 2 | 3 | 0.392566 |
| 3 | 3 | 0.472607 |
| 3 | 2 | 0.329042 |
| 2 | 2 | 0.505012 |
| 355 | 2 | 3 | 0.390307 |
| 3 | 3 | 0.281221 |
| 3 | 2 | 0.281996 |
| 2 | 2 | 0.285348 |

Анализ мерных интервалов показал, что самый лучший прогноз даёт минимальная длина мерного интервала

2) Модель структурного временного ряда (BSTS)

Байесовская модель структурных временных рядов (BSTS) — это статистический метод, используемый для выбора признаков, прогнозирования временных рядов, прогнозирования текущих значений, вывода причинно-следственных связей и других приложений. Модель предназначена для работы с данными временных рядов.

Структурные модели временных рядов — это одно из семейств моделей пространства состояний. В структурных моделях временной ряд представлен в виде суммы ненаблюдаемых компонент, которые можно интерпретировать как тренд, сезонность, эффекты предикторов и т.д.

num\_warmup, num\_samples, num\_chains – параметры модели BSTS.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Длина мерного интервала | num\_warmup | num\_samples | num\_chains | Среднеквадратическое отклонение |
| 20 | 500 | 500 | 2 | 4.057462 |
| 500 | 500 | 3 | 3.969551 |
| 250 | 250 | 2 | 3.322375 |
| 50 | 500 | 500 | 2 | 8.015036 |
| 500 | 500 | 3 | 6.919893 |
| 250 | 250 | 2 | 7.156304 |
| 100 | 500 | 500 | 2 | 1.282715 |
| 500 | 500 | 3 | 1.136338 |
| 250 | 250 | 2 | 2.558352 |
| 200 | 500 | 500 | 2 | 2.097891 |
| 500 | 500 | 3 | 2.74065 |
| 250 | 250 | 2 | 1.965168 |
| 300 | 500 | 500 | 2 | 1.909097 |
| 500 | 500 | 3 | 2.409703 |
| 250 | 250 | 2 | 2.319264 |
| 355 | 500 | 500 | 2 | 3.53315 |
| 500 | 500 | 3 | 7.537256 |
| 250 | 250 | 2 | 3.089138 |

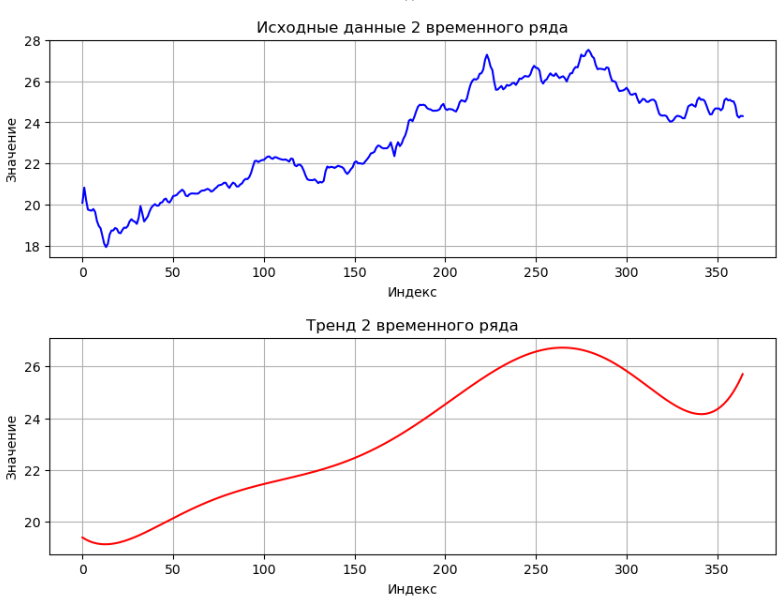
Во временном ряду оптимальным значением мерного интервала оказалось 100.

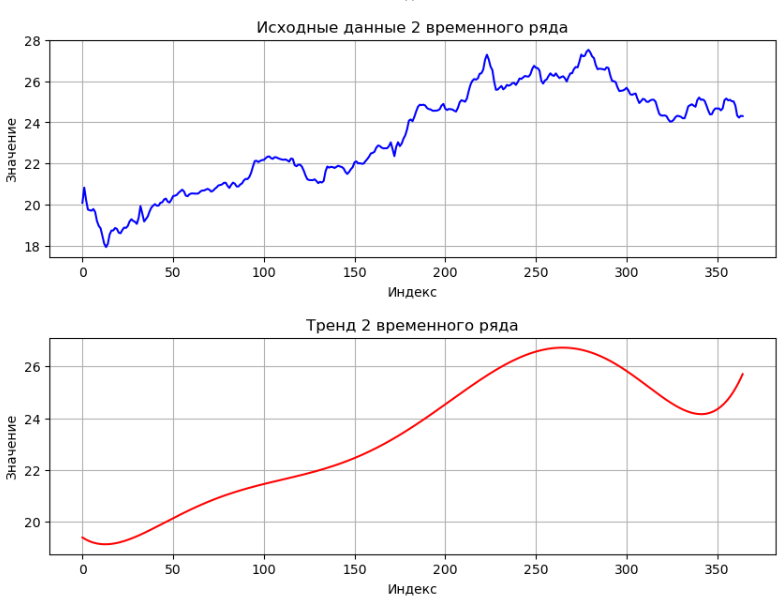
Как видно, модель BSTS работает хуже для временного ряда. Можно предположить, что это из-за того, что ряд является нестационарным. Нестационарность временных рядов приводит к изменению их среднего и дисперсии со временем, что делает сложным построение точных и надежных прогнозов.

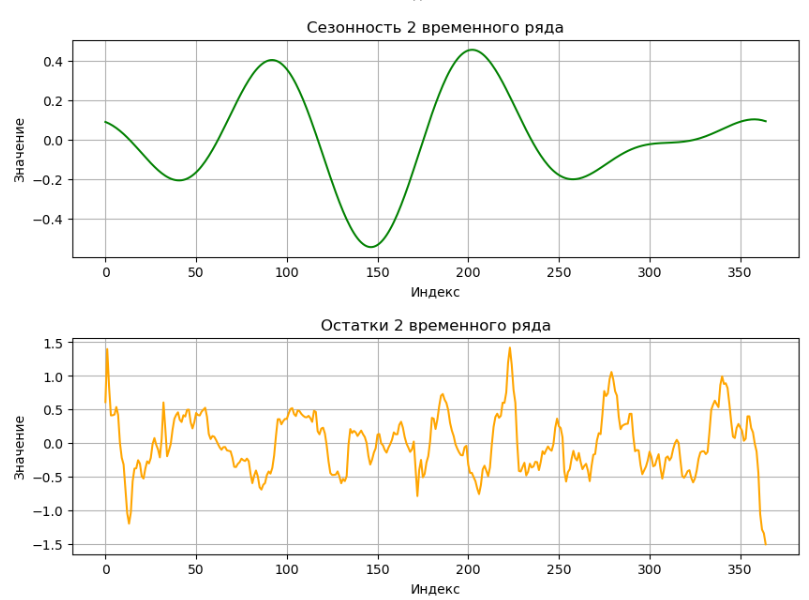
3) Разложение ряда на компоненты

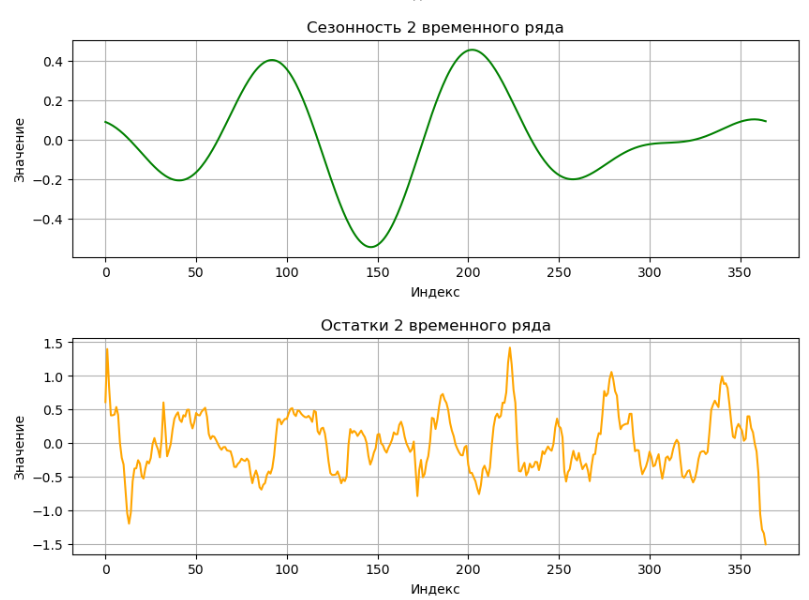
3.1) Выделение

* трендовой составляющей на основе полиномиальной регрессионной модели
* сезонной составляющей на основе Фурье-анализа;
* остаточной составляющей (проверить имеют ли остатки нормальное распределение).



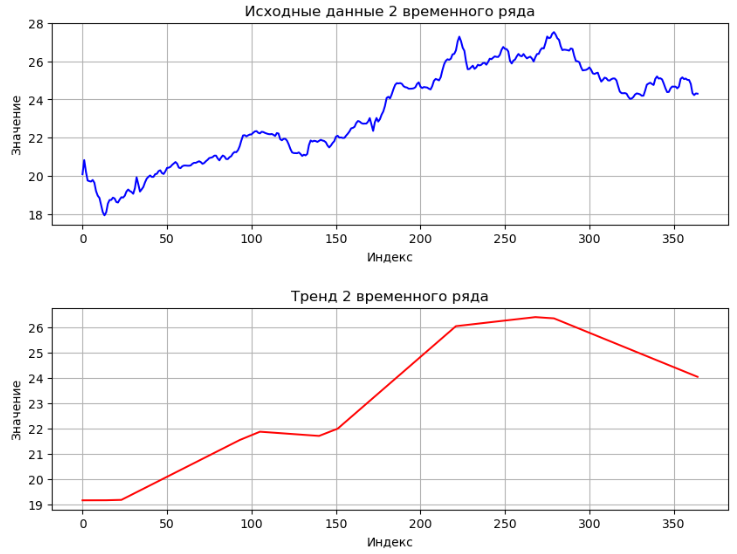




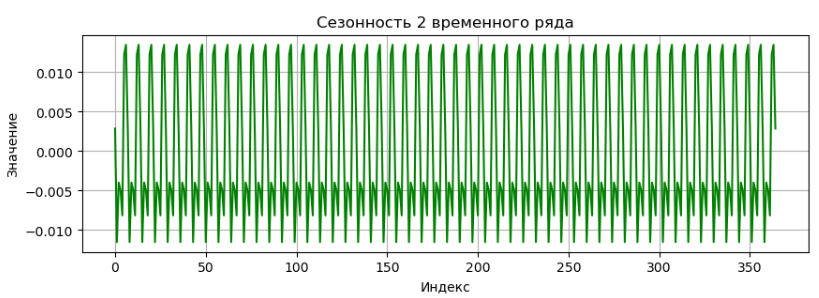


Остатки имеют нормальное распределение.

3.2) Сравнение полученных в п.3.1 результатов с аддитивной нелинейной регрессионной моделью (пакет Prophet).









Остатки имеют нормальное распределение.

3.3) Исследовать точность моделей из п.3.1 и 3.2 для различных длин мерных интервалов. Результаты представить в виде таблицы из п.1.

Точность модели на основе разложения ряда на компоненты:

|  |  |
| --- | --- |
| Длина мерного интервала | Среднеквадратическое отклонение |
| 20 | 3661.913629 |
| 50 | 162.226109 |
| 100 | 0.100933 |
| 200 | 2.084075 |
| 300 | 1.767713 |
| 355 | 2.734927 |

Во временном ряду оптимальным значением мерного интервала оказалось 100.

Точность модели на основе пакета Prophet:

|  |  |
| --- | --- |
| Длина мерного интервала | Среднеквадратическое отклонение |
| 20 | 0.86555 |
| 50 | 0.143701 |
| 100 | 0.192307 |
| 200 | 0.150945 |
| 300 | 0.737172 |
| 355 | 0.833029 |

Во временном ряду оптимальным значением мерного интервала оказалось 50.

**4. Вывод**

В работе было рассмотрено несколько моделей для прогнозирования временных рядов, в таблицах ниже выписаны самые оптимальные значения для различных моделей.

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Мерный интервал | Параметры (при наличии) | Среднеквадратическое отклонение |
| Модель ARIMA | 20 | (p, d, q) = (1, 1, 3) | 0.149484 |
| Модель BSTS | 100 | (num\_warmup, num\_samples, num\_chains) = (500, 500, 3) | 1.136338 |
| Разложение ряда на компоненты | 100 | - | 0.100933 |
| Пакет Prophet | 50 | - | 0.143701 |

Модель BSTS дала самый плохой результат, а разложения ряда на компоненты (тренд, сезонность и остатки) с мерным интервалом 100 дала очень хороший результат.

**Прогноз временного ряда методом SSA**

**1. Исходные данные**

Значения временного ряда y(t), t=0,1,2,…,m – годовой курс валюты (дирхама к российскому рублю)

**2. Задание**

Выполнить прогноз временного ряда рекуррентным методом SSA (Singular Spectrum Analysis, "Гусеница") на 10 значений вперед. В качестве обучающей выборки использовать предыдущие данные N значений временного ряда. Для реализации метода SSA необходимо задать 2 параметра: L - ширина окна для построения траекторного пространства ряда, r - количество компонент (r<L<N)

**3. Описание алгоритма**

Сингулярный спектральный анализ (SSA) – это метод анализа временных рядов. Суть метода заключается в том, что временной ряд разбивается на компоненты сингулярного разложения и затем эти компоненты используются для анализа и прогнозирования.

SSA может быть использован без предварительного задания модели ряда для анализа произвольных, в том числе, нестационарных, рядов. Основная цель SSA — разложить ряд в сумму интерпретируемых компонент, таких как тренд, периодические компоненты, шум.

Базовый алгоритм SSA

1) Формирование траекторной матрицы

где – длина окна (обычно размер окна ), – количество компонент. Наш ряд расположен в первом столбце и продолжается в последней строке

2) Сингулярное разложение траекторной матрицы , где - матрица левых сингулярных векторов, - диагональная матрица сингулярных чисел, - матрица правых сингулярных векторов

3) Группировка сингулярных компонент.

Упорядочим сингулярные числа в порядке убывания

Выберем наиболее значимых сингулярных компонент, основываясь на некотором критерии (например, по доле объясненной дисперсии)

4.1) Восстановление исходного ряда

Сформируем матрицу восстановленных компонент размера :

Усредним диагональные элементы матрицы для получения восстановленного временного ряда:

4.2) Прогнозирование с использованием SSA

Выделим последнюю строку матрицы , содержащую последних наблюдений и спрогнозируем следующее значение временного ряда, используя линейную комбинацию последних значений:

Повторим этот процесс для получения требуемого количества прогнозных значений. Коэффициенты можно найти, решив систему линейных уравнений, основанную на последних значениях восстановленного ряда .

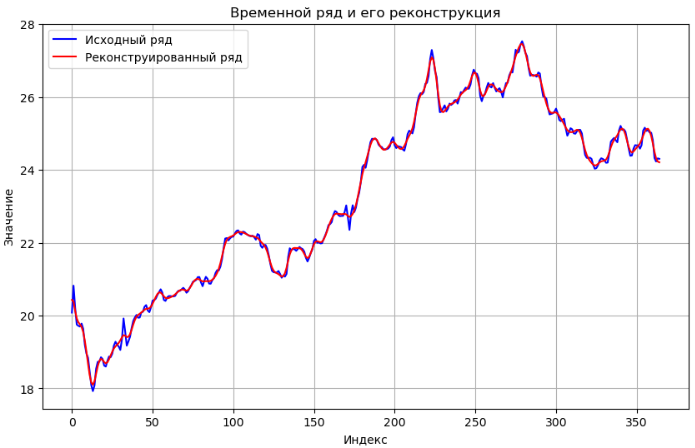
Основные преимущества SSA:

* Способность выявлять и разделять различные компоненты временного ряда (тренд, сезонность, шум).
* Возможность восстановления исходного ряда с удалением ненужных компонент.
* Применимость к нелинейным и нестационарным временным рядам.
* Отсутствие необходимости в предположениях о модели временного ряда.

Таким образом, SSA представляет собой мощный метод анализа и обработки временных рядов, позволяющий эффективно выделять и изучать различные структурные компоненты.

**4. Ход работы**

С помощью метода SSA нам удалось разложить временной ряд курса валют на компоненты, после чего реконструировать это ряд. Результаты расположены на графике ниже:



Ряд был разложен на некоторое количество компонент. С помощью графика энтропии были выбраны те компоненты, которые сильно влияли на энтропию (компоненты, добавление которых не давало существенного изменения энтропии, были отброшены). Т.е. метод выбора оптимального количества компонент на основе энтропии позволяет отделить полезные компоненты (тренд, сезонность) от шумовых компонент в разложении временного ряда методом SSA

Полезные компоненты:

|  |  |
| --- | --- |
|  |  |
|  |  |

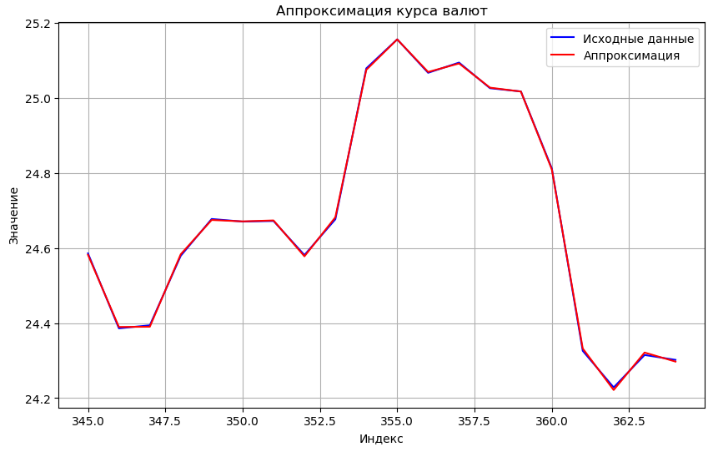
Компонента, имеющая плавное, монотонное изменение во времени, скорее всего, отражает тренд. Компоненты с явно выраженной периодичностью (например, годовая или квартальная) соответствуют сезонной составляющей

**5 Исследования**

1) Проведём исследования для поиска оптимальных значений мерного интервала , ширины окна и количество компонент , необходимых для аппроксимации курса валют:

|  |  |  |  |
| --- | --- | --- | --- |
| Длина мерного интервала | Ширина окна | Количество компонент | Среднеквадратическое отклонение |
| 20 | 5 | 2 | 0.00673 |
| 3 | 0.001497 |
| 4 | 0.000148 |
| 7 | 2 | 0.01182 |
| 4 | 0.001581 |
| 6 | 1.5e-05 |
| 10 | 3 | 0.00682 |
| 6 | 0.000826 |
| 9 | 2e-06 |
| 50 | 15 | 4 | 0.007506 |
| 9 | 0.000393 |
| 12 | 2.4e-05 |
| 20 | 5 | 0.006221 |
| 10 | 0.000759 |
| 15 | 5.4e-05 |
| 25 | 8 | 0.002599 |
| 14 | 0.000456 |
| 20 | 2e-05 |
| 100 | 25 | 8 | 0.005829 |
| 14 | 0.00075 |
| 20 | 3.4e-05 |
| 35 | 8 | 0.008504 |
| 16 | 0.001333 |
| 24 | 0.000102 |
| 50 | 10 | 0.008032 |
| 25 | 0.000539 |
| 40 | 6e-06 |
| 200 | 50 | 10 | 0.01276 |
| 25 | 0.001358 |
| 40 | 4.6e-05 |
| 75 | 20 | 0.00769 |
| 40 | 0.000973 |
| 60 | 3.8e-05 |
| 100 | 30 | 0.004337 |
| 50 | 0.000819 |
| 80 | 1.9e-05 |
| 300 | 75 | 20 | 0.006586 |
| 40 | 0.001016 |
| 60 | 4.2e-05 |
| 100 | 20 | 0.010631 |
| 50 | 0.001169 |
| 80 | 3.8e-05 |
| 150 | 25 | 0.012385 |
| 50 | 0.002908 |
| 100 | 0.000109 |
| 355 | 85 | 20 | 0.009476 |
| 40 | 0.001728 |
| 60 | 0.000189 |
| 125 | 50 | 0.002468 |
| 75 | 0.000382 |
| 100 | 3.6e-05 |
| 175 | 10 | 0.065064 |
| 25 | 0.016774 |
| 50 | 0.004526 |
| 365 | 90 | 25 | 0.007636 |
| 45 | 0.001407 |
| 65 | 0.000152 |
| 135 | 30 | 0.009594 |
| 70 | 0.000859 |
| 100 | 7e-05 |
| 180 | 15 | 0.035624 |
| 30 | 0.014062 |
| 55 | 0.004396 |

Можно сделать вывод, что оптимальными значениями являются = 20, = 7 и = 6:



2) Проведём исследования для поиска оптимальных значений мерного интервала , ширины окна и количество компонент , необходимых для прогнозирования 10 значений курса валют:

|  |  |  |  |
| --- | --- | --- | --- |
| Длина мерного интервала | Ширина окна | Количество компонент | Среднеквадратическое отклонение |
| 20 | 5 | 2 | 0.038114 |
| 3 | 0.491012 |
| 4 | 0.047551 |
| 7 | 2 | 0.070601 |
| 4 | 0.14039 |
| 6 | 0.107963 |
| 10 | 3 | 3.996326 |
| 6 | 1.083964 |
| 9 | 0.312541 |
| 50 | 15 | 4 | 1.015815 |
| 9 | 1.11344 |
| 12 | 0.067765 |
| 20 | 5 | 0.175221 |
| 10 | 7.393156 |
| 15 | 2.2936 |
| 25 | 8 | 0.26062 |
| 14 | 2.131455 |
| 20 | 0.998846 |
| 100 | 25 | 8 | 1.500448 |
| 14 | 0.613422 |
| 20 | 1.471223 |
| 35 | 8 | 2.18119 |
| 16 | 2.267076 |
| 24 | 2.812871 |
| 50 | 10 | 3.255322 |
| 25 | 2.508683 |
| 40 | 2.776649 |
| 200 | 50 | 10 | 0.478012 |
| 25 | 0.609944 |
| 40 | 0.814127 |
| 75 | 20 | 0.284223 |
| 40 | 0.165201 |
| 60 | 0.255125 |
| 100 | 30 | 0.307209 |
| 50 | 0.645174 |
| 80 | 0.865981 |
| 300 | 75 | 20 | 1.035321 |
| 40 | 0.972886 |
| 60 | 0.561286 |
| 100 | 20 | 0.551419 |
| 50 | 0.375227 |
| 80 | 0.414207 |
| 150 | 25 | 6.95732 |
| 50 | 10.622679 |
| 100 | 23.302918 |
| 355 | 85 | 20 | 0.470801 |
| 40 | 0.503049 |
| 60 | 0.255462 |
| 125 | 50 | 0.467715 |
| 75 | 0.618494 |
| 100 | 0.204509 |
| 175 | 10 | 0.859566 |
| 25 | 6.260588 |
| 50 | 9.562279 |

Можно сделать вывод, что оптимальными значениями являются = 20, = 5 и = 2:



**6. Вывод**

1) В предыдущих частях работы были получены следующие оптимальные значения параметров для АППРОКСИМАЦИИ ряда для различных моделей:

|  |  |
| --- | --- |
| Модель | Среднеквадратическое отклонение |
| Модель ARIMA | 1.120807724 |
| Модель BSTS | 4e-06 |
| Разложение ряда на компоненты | 0.191411 |
| Пакет Prophet | 0.200685 |
|  |  |
| Метод SSA | 1.5e-05 |

Т.е. наилучшей моделью из предыдущих лабораторных работ стала модель BSTS с = 4e-06. Сингулярный спектральный анализ же дал лучший результат с параметрами = 20, = 7 и = 6 и = 1.5e-05

2) В предыдущих лабораторных работах были получены такие оптимальные значения параметров для ПРОГНОЗИРОВАНИЯ 10 значений для различных моделей:

|  |  |
| --- | --- |
| Модель | Среднеквадратическое отклонение |
| Модель ARIMA | 0.149484 |
| Модель BSTS | 1.136338 |
| Разложение ряда на компоненты | 0.100933 |
| Пакет Prophet | 0.143701 |
|  |  |
| Метод SSA | 0.038114 |

Т.е. наилучшей моделью из предыдущих лабораторных работ стало разложение ряда на компоненты с = 0.100933. Сингулярный спектральный анализ же дал лучший результат с параметрами = 20, = 5 и = 2 и = 0.038114

В ходе работы был проведен сравнительный анализ эффективности различных методов анализа и прогнозирования временных рядов на примере реального набора данных. Рассматривались следующие подходы:

1. Модели ARIMA
2. Модели BSTS
3. Ручное разложение временного ряда на компоненты
4. Использование пакета Prophet
5. Сингулярный спектральный анализ (SSA)

Результаты проведенного исследования показали, что метод сингулярного спектрального анализа (SSA) продемонстрировал наиболее высокую эффективность в задаче анализа и прогнозирования временного ряда по сравнению с другими рассмотренными методами.