Санкт-Петербургский государственный университет

Нафикова Лиана Ирековна

Выпускная квалификационная работа

Разработка прототипа системы определения тренда во временном ряде

Уровень образования: бакалавриат

Направление *02.03.03 «Математическое обеспечение и администрирование информационных систем»*

Основная образовательная программа *СВ.5162.2020 «Технологии программирования»*

Научный руководитель:
доктор физ.матем.наук, профессор О. Н. Граничин

Рецензент:
Deep Learning Engineer Brask Inc. В. Д. Панков

Санкт-Петербург
2024

Saint Petersburg State University

Liana Nafikova

Bachelor's Thesis

Development of a prototype system for determining a trend in a time series

Education level: bachelor

Speciality *02.03.03 «Software and Administration of Information Systems»*

Programme *CB.5162.2020 «Programming Technologies»*

Scientific supervisor:
Sc.D, prof. O. N. Granichin

Reviewer:
Deep Learning Engineer Brask Inc. V. D. Pankov

Saint Petersburg
2024

Оглавление

Введение

Термин «прогноз» происходит от греческого слова «prognosis», которое переводится как предвидение или предсказание. Под этим термином обычно понимают исследование перспектив развития некоторого явления. Чаще всего такие явления описываются временными рядами и включают в себя два обязательных элемента — отметку времени и значение показателя ряда.

Анализ временных рядов является мощным инструментом для извлечения ценной информации из последовательно поступающих данных. В нашем изменяющемся мире временные ряды применяются в различных областях — от экономики и финансов до задач медицины. Однако существуют сценарии, когда непрерывный мониторинг показателей может оказаться весьма дорогостоящим или решения необходимо принимать оперативно. В таких ситуациях объём данных для анализа чаще всего критически мал.

Примером может служить контроль надежности подшипников качения в машинах. Онлайн-мониторинг, применяемый для крупномасштабных и дорогостоящих установок, требует значительных затрат на оборудование. Другой подход контроля предполагает измерения с более длительными интервалами времени. Главным достоинством этой стратегии является ее более низкая стоимость, потому что все машины компании могут быть проконтролированы ограниченным числом устройств. Именно поэтому эта стратегия широко распространена для машин среднего и малого размера на промышленных предприятиях. В данном случае история каждой машины содержит значительно меньше данных. Также, например, в финансах или биржевой торговле рыночные условия меняются молниеносно. На финансовых рынках цены акций могут колебаться в зависимости от различных факторов, и для успешного трейдинга необходимо моментально анализировать данные и принимать решения о покупке или продаже. Помимо этого возможен случай, когда проведение измерений требует значительных затрат времени. Например, при контроле прочности динамических материа-

лов [22].

Временные ряды обладают различными статистическими характеристиками, которые влияют на результаты анализа. В первую очередь влияет модель представления членов временного ряда [7]. Одно из известных представлений выглядит следующим образом:

$$x_t = T_t + \varepsilon_t,$$

где T_t является компонентом тренда, а ε_t — случайным членом.

В некоторых временных рядах помимо этих двух компонент выделяют также сезонность. Для её учета обычно вводят фиктивные колебания. Следует отметить, что часто статистические данные публикуются с поправкой на сезонность, так что учитывать ее не нужно. Например, квартальные данные U.S.GDP[1] публикуются с исключением сезонности.

Второй член представления — это неопределённость, под которой обычно понимают случайную ошибку в измерениях, шум. Этот шум может быть вызван различными факторами, такими как ошибки измерений, аномальные события или непредсказуемые изменения в окружающей среде. Для корректных результатов необходимо учитывать, что многие временные ряды собраны в реальных условиях, где факторы шума неизбежны.

Тренд описывает направление распространения данных, поэтому обнаружение моментов изменения направлений тренда во временных рядах очень важно в принятии стратегических решений и адаптации к изменяющимся условиям. Эти моменты могут свидетельствовать о существенных изменениях, таких как экономические рецессии, изменения в потребительских предпочтениях или технологические инновации. Благодаря своевременному обнаружению этих изменений, организации могут адаптировать свои стратегии, минимизировать потери и максимизировать возможности для роста и развития. Точное определение моментов изменения тренда также помогает улучшить качество про-

[1]https://datatopics.worldbank.org/world-development-indicators/ — сайт Всемирного банка (дата обращения: 20.12.2023)

гнозирования, что необходимо для успешного управления ресурсами, финансами и производственными процессами. Важность этой задачи подчёркивается не только в области научных исследований, но и в практическом применении.

Эффективное управление данными шума и учет его влияния позволят разработать более точные и надежные модели определения тренда, способные адаптироваться к различным условиям и обеспечивать более достоверные результаты анализа. Такие алгоритмы, как метод знаковозмущённых сумм [4], могут быть полезны для построения достоверных прогнозов и доверительных интервалов. В условиях ограниченных ресурсов, когда необходимо оперативно принимать решения, создание системы, которая позволит выделять тренд, бороться с шумом во временных рядах, становится крайне важным.

1. Постановка задачи

Целью работы является создание функционального прототипа системы для анализа временных рядов, определения их трендов и моментов существенных изменений трендов. Для достижения этой цели были сформулированы следующие задачи.

- Сформулировать требования к системе.
- Провести обзор существующих решений и исследовать алгоритмы, позволяющие определять тренд по малому числу входных данных с шумом.
- Разработать архитектуру системы.
- Реализовать прототип.
- Провести эксперименты.

2. Требования к системе

Для системы были сформулированы следующие требования.

- Возможность изменения алгоритмов.
- Обработка исходных данных.
- Возможность моделирования шума.
- Представление результатов работы алгоритмов.
- Требования к пользовательскому интерфейсу.

Рассмотрим эти требования детально.

2.1. Возможность изменения алгоритмов

Каждый из алгоритмов имеет свои входные параметры. Например, SPS-метод в качестве входных параметров принимает значения q и M, которые задают доверительную вероятность $p = 1 - \frac{q}{M}$. Система должна быть спроектирована таким образом, чтобы была возможность гибко настраивать каждый из алгоритмов. То есть, пользователю должна быть предоставлена возможность изменять параметры алгоритмов в соответствии с требованиями конкретной задачи или ситуации.

2.2. Обработка исходных данных

Система должна обрабатывать как готовые (в том числе сгенерированные) наборы данных, так и те, что поступают в режиме реального времени. Сгенерированные данные необходимы для тестирования системы.

Для обработки готовых наборов данных необходимо реализовать интерфейс для загрузки. Пользователь должен иметь возможность выбрать нужные данные из датасета и сохранить их для дальнейшей работы. При этом система должна корректно обрабатывать разные типы

входных данных, так как временные ряды могут представлены, например, в JSON- или CSV-файлах.

Для работы с потоковыми данными система должна поддерживать подключение к различным источникам данных.

2.3. Возможность моделирования шума

В системе должна быть функциональность для добавления шума при генерации данных. Этот шум должен обладать разными статистическими свойствами. Фильтр Калмана, например, требует, чтобы шум был нормально распределённым с нулевым матожиданием, в то время как метод SPS требует лишь симметричности функции шума.

2.4. Представление результатов работы алгоритмов

Результаты работы алгоритмов система должна представлять в виде графиков. Также должна быть возможность загружать полученные результаты на компьютер пользователя. В результатах необходимо представлять подсчитанные метрики качества.

2.5. Требования к пользовательскому интерфейсу

Интерфейс взаимодействия должен быть интуитивно понятным и удобным для пользователей для того, чтобы предоставлять возможность тестировать алгоритмы, абстрагируясь от их реализации. Необходимо создать веб-приложение, чтобы пользователи смогли получать прогнозы и анализировать данные в удобном для них формате.

3. Обзор

3.1. Обзор алгоритмов

Далее будем придерживаться следующих обозначений.

- $\hat{x}_{t|t-1}$ — прогноз значения, соответствующего времени t по $t-1$ предыдущим измерениям.
- $x_1, \ldots, x_t$ — значения ряда в разные моменты времени.
- $\varepsilon_1, \ldots, \varepsilon_t$ — случайные ошибки в разные моменты времени.

Тренд является компонентом, который отражает основное направление изменений данных во времени. Перед тем, как фиксировать изменения тренда, необходимо выделить его из данных. Считается, что тренд является некоторой сглаженной версией исходного ряда, то есть представляет собой также временной ряд.

Задачу определения тренда можно связать с задачей прогнозирования временного ряда. В ходе прогнозирования необходимо создать модель, способную предсказывать будущие значения временного ряда на основе предыдущих:

$$\hat{x}_{t|t-1} \approx F(x_{t-1}, ..., x_1),$$

Для каждого значения временного ряда в качестве модели для определения тренда можно использовать модель, которая учитывает шумы, из задачи предсказания. Поэтому в качестве алгоритмов для определения тренда в этой главе будут рассметрены алгоритмы для предсказания будущих значений.

3.1.1. Метод Хольта

Метод Хольта является модификацией экспоненциального сглаживания. Основная идея обоих методов заключается в том, чтобы взвешивать текущее значение временного ряда с предыдущими прогнозами, уменьшая веса по мере удаления в прошлое.

Экспоненциальное сглаживание имеет параметр сглаживания α, который определяет вес текущего значения по отношению к предыдущим прогнозам. Выбор оптимального значения α зависит от характера временного ряда и требуемой чувствительности к изменениям.

Общая формула для прогноза в методе экспоненциального сглаживания с учётом одного последнего значения выглядит следующим образом:

$$\hat{x}_{t|2} = \alpha \cdot x_t + \alpha(1-\alpha) \cdot \hat{x}_{t-1|2}$$

Хольт расширил простое экспоненциальное сглаживание, чтобы обеспечить прогнозирование данных с тенденцией. Этот метод включает в себя уравнение прогноза и два уравнения сглаживания (одно для уровня и одно для тренда) [8]:

$$\begin{aligned}
&\text{Уравнение прогноза} && \hat{x}_{t+h|t} = \ell_t + hb_t \\
&\text{Уравнение уровня} && \ell_t = \alpha x_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
&\text{Уравнение тренда} && b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1},
\end{aligned}$$

Здесь ℓ_t обозначает оценку уровня ряда в момент времени t, которая представляет собой средневзвешенное значение наблюдений, b_t — оценка тренда ряда во времени t, α — параметр сглаживания уровня, и β — параметр сглаживания тренда.

3.1.2. Авторегрессионное интегрированное скользящее среднее

В начале стоит рассмотреть упрощённую модель ARMA. Autoregressive Moving Average модель состоит из двух частей: авторегрессионной (AR) части и скользящего среднего(MA). Обычно используется обозначение ARMA(p,q), где p — порядок регрессионной части, а q — порядок скользящего среднего. Общая формула для этой модели:

$$\hat{x}_{t|t-min(p,q)} = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{p} \phi_i x_{t-i},$$

где $\theta_1, ..., \theta_q$ и $\phi_1, ..., \phi_p$ — параметры модели. Уравнение модели строится на основании суммирования среднего значение ряда, ошибок, и значений временного ряда, полученных на предыдущих шагах.

Autoregressive Integrated Moving Average является обобщением авторегрессионного скользящего среднего. ARMA используется для стационарных рядов (т.е. рядов, в которых нет резкой смены направления тренда), а интегрирование, представленное в модели ARIMA, является инструментом, который приводит ряд к стационарному виду. В модель добавляется третий параметр d, который называют порядком интегрирования. Она показывает, насколько элемент ряда близок по значению к d предыдущим значениям, если разность между ними минимальна. Ниже представлена формула для модели ARIMA:

$$(\triangle^d \hat{x}_{t|t-min(p,q)}) = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i (\triangle^d \varepsilon_{t-i}) + \sum_{i=1}^{p} \phi_i (\triangle^d x_{t-i})$$

Здесь $\triangle^d$ — оператор разности порядка d (последовательное взятие d раз разностей первого порядка — сначала от самого ряда, затем от полученных разностей первого порядка, затем от второго порядка и т.д.).

3.1.3. Фильтр Калмана

Одним из методом, позволяющим совершать предсказания на основе зашумлённых или неполных данных, является Фильтр Калмана [10]. Метод описывается двумя моделями: моделью движения (перехода) и моделью измерений.

Модель движения описывает, как переменные состояния эволюционируют с течением времени:

$$x_t = F_t x_{t-1} + B_t u_t + w_t,$$

где F_t — матрица перехода от состояния в момент времени $t-1$ к состоянию в момент времени t;
B_t — матрица управления;
u_t — управление;

w_t — возмущение;

Модель измерений показывает, как измерения связаны с текущим состоянием системы:

$$z_t = H_t x_t + \varepsilon_t,$$

где z_t — наблюдение в момент времени t;
H_t — матрица измерений;

Предсказания осуществляются с помощью формул:

$$\hat{x}_{t|t-1} = F_t \hat{x}_{t-1|t-1} + B_t u_t$$

$$P_{t|t-1} = F_t P_{t-1|t-1} F_t^T + Q_t$$

где $P_{t|t-1}$ — прогнозированная ковариация ошибки предсказания;
Q_t — ковариация возмущения модели движения.

Далее появляются новые измерения, они сравниваются с прогнозом, и система корректирует свою оценку, учитывая разницу между прогнозом и реальными данными. Это происходит с использованием Калмановского коэффициента, который оптимальным образом комбинирует информацию из прогноза и измерений.

$$K_t = P_{t|t-1} H_t^T (H_t P_{t|t-1} H_t^T + R_t)^{-1}$$

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t (z_t - H_t \hat{x}_{t|t-1})$$

$$P_{t|t} = (I - K_t H_t) P_{t|t-1},$$

где K_t — калмановский коэффициент;
R_t — ковариация ошибки измерений.

В контексте решения задачи определения тренда матрица управления является нулевой, а матрица перехода размером 2x2, поскольку движение моделируется как изменение текущего значения тренда и скорости изменения.

3.1.4. Линейная регрессия

В случае линейной регрессии линия тренда описывается следующим уравнением:

$$x_t = kt + b + \varepsilon_t,$$

где k, b — коэффициенты, которые показывают, как изменяется x_t при изменении t.

Выбор регрессионной линии, описывающей взаимосвязь данных наилучшим образом, заключается в минимизации функции потерь, представленной в виде среднеквадратичной ошибки. Линия должна проходить через данные таким образом, чтобы в среднем разница квадратов ожидаемых и реальных значений была минимальна. Данный метод называется методом наименьших квадратов.

3.1.5. SPS

Метод знаковозмущённых сумм (Sign-Perturbed Sums, SPS) представляет собой метод, используемый для построения доверительных интервалов и оценки неопределенности параметров в системах уравнений. В качестве модели для линии тренда предлагается также использовать уравнение

$$x_t = kt + b + \varepsilon_t,$$

но в данной модели минимизируется уже не среднеквадратичная ошибка, а строятся доверительные оценки параметров k, b.

Метод состоит из двух этапов: инициализация параметров и построение доверительного множества. На первом этапе задаётся необходимая доверительная вероятность, на втором — проверяется, попадает ли прямая с выбранными k, b в доверительное множество или нет.

Основные шаги SPS:

1. Выбор начальных параметров: количество знаковозмущённых сумм M и уровень значимости q, с помощью которых задаётся доверительная вероятность $p = 1 - q/M, M > q > 0$.

2. Генерируются $T(M-1)$ случайных одинаково распределённых независимых бернуллиевских величин. T в данном случае является количеством измерений.

$$\beta_{i,j} = \pm 1 : P(\beta_{i,j} = 1) = P(\beta_{i,j} = -1) = 0.5$$

для $j = 1, ..., T$ $i = 1, ..., M$.

3. Считается ссылочная сумма

$$H_0(k, b) = \sum_{j=1}^{T} (x_j - kt_j - b).$$

4. Генерируются знако-возмущённые суммы

$$H_i(k, b) = \sum_{j=1}^{T} \beta_{i,j}(x_j - kt_j - b)$$

для $i = 1...M - 1$.

5. Полученные на предыдущих двух шагах суммы возводятся в квадрат и сортируются по возрастанию.

6. Позиция $H_0^2(k, b)$ в упорядоченном наборе называется рангом r. В случае если ранг $r >= q$, то считается, что прямая с данными коэффициентами лежит в доверительном множестве.

3.2. Сравнение существующих решений

На сегодняшний день существует большое число систем, позволяющих предсказывать тренд во временных рядах [1, 2, 12, 15, 16, 17]. Некоторые из них представляют собой среды для визуального программирования и анализа данных, другие же — библоитечные пакеты.

В таблице 1 показаны особенности бесплатных и открытых систем Orange [13], forecast [20], KNIME [9], Prophet [14], которые больше всего подходят для решения задачи определения тренда во временных рядах.

Orange предоставляет широкий спектр инструментов для визуального анализа данных и машинного обучения. Библиотека forecast в R предлагает множество методов и моделей для прогнозирования временных рядов, включая ARIMA и многие другие. KNIME является платформой с открытым исходным кодом, предназначенной для анализа данных и создания рабочих процессов. Prophet, решение от Facebook, специально предназначен для анализа и прогнозирования временных рядов.

Однако ни одна из этих систем не предлагает полного решения для всех требований.

	Orange	forecast	KNIME	Prophet
Настройки алгоритмов	+	+	+	+
Генерация данных	-	-	+	-
Прогнозирование готовых датасетов	+	+	+	+
Чтение данных в режиме реального времени	-	-	+	-
Пользовательские алгоритмы	+	-	+	-
Модель шума	-	-	-	-
Экспорт результатов	+	+	+	+
GUI	+	-	+	-

Таблица 1: Сравнительная характеристика систем

3.3. Выводы

Исследованные алгоритмы решают разные задачи и годятся как для предсказания будущих значений, так и для определения тренда. Например, метод Хольта позволяет найти локальный тренд, то есть тренд в окрестности одного измерения, в то время как линейная регрессия находит тренд на всех данных, а метод знаковозмущённых сумм — доверительные интервалы для предсказанных значений тренда. Было решено

реализовать все эти алгоритмы в системе, для того, чтобы пользователь мог максимально широко исследовать временной ряд.

По результатам сравнительной характеристики существующих систем видно, что ни одна из них не позволяет задавать модели шума. Чаще всего системы позволяют лишь генерировать случайные шумы. Также не все системы обладают пользовательским интерфейсом, необходимым для быстрого тестирования алгоритмов.

4. Архитектура системы

В этой главе описываются основные решения об организации системы: из каких компонентов она состоит, как эти компоненты взаимодействуют между собой, как они взаимодействуют с окружением, мотивация принятых решений.

На рис. 1 представлена контекстная диаграмма, которая кратко описывает работу системы. Пользователь загружает или генерирует данные, выбирает алгоритм, вводит параметры. Далее, сформированная модель подается на вход выбранного алгоритма, и он запускается. После выполнения вычислений пользователь получает доступ к результатам работы в виде графиков или текстового файла. При желании пользователь может добавить свой алгоритм в систему.

Рис. 1: Контекстная диаграмма системы

Языком реализации серверной части системы был выбран Python, так как он обладает обширной экосистемой библиотек для математического моделирования и веб-разработки.

Flask [6] и Django [5] — это два известных фреймворка для создания веб-приложений на языке программирования Python, каждый из которых обладает своими уникальными особенностями. Для реализации серверной части системы был выбран Flask вместо Django, поскольку он является более легковесным и предоставляет лишь необходимый минимум инструментов для разработки. Кроме того, Flask считается более простым в изучении. Также фреймворк Flask обладает обширной экосистемой расширений, что позволяет интегрировать различный функционал в приложение. Он легко интегрируется с другими технологиями и фреймворками, такими как, например, React, с помощью которой решено реализовать клиентскую часть приложения. Данная библиотека считается более популярной по сравнению с аналогами Vue [18] и Angular [3] по данным Stack Overflow Trends[2] и npm trends[3].

Архитектура системы представлена на рис. 2.

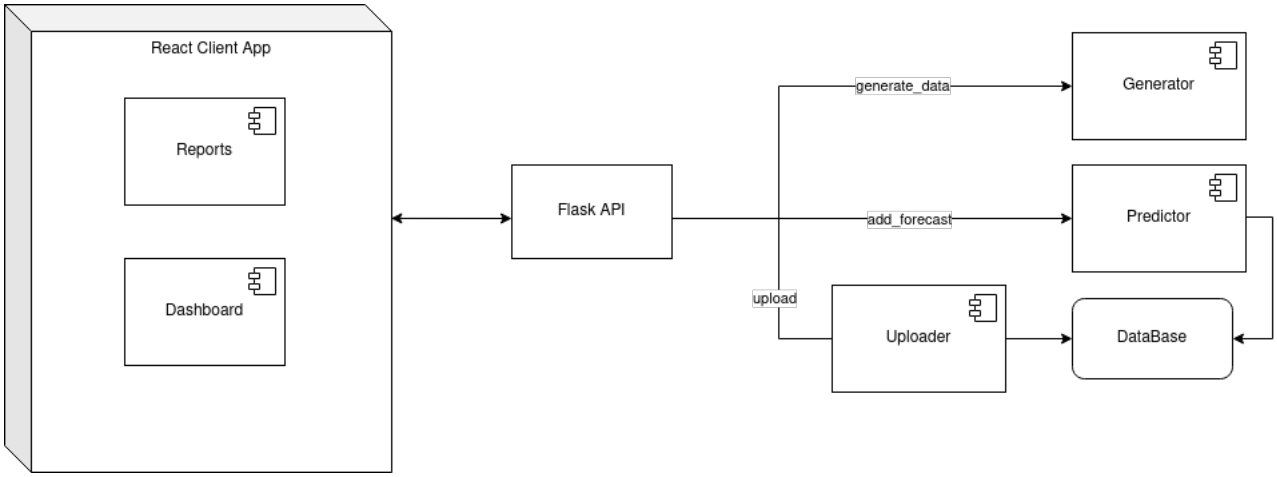

Рис. 2: Архитектура системы

На сегодняшний момент доступно большое количество UI-библиотек компонентов для создания пользовательских интерфейсов. Они облегчают работу программисту, так как предоставляют готовые стилевые решения. В качестве основной библиотеки было решено использовать Material UI [11]. К сожалению, она не предоставляет широкого функ-

[2]https://insights.stackoverflow.com/trends?tags=angular%2Creactjs%2Cvue.js — StackOverflowTrends for React, Angular, Vue (дата обращения 26.10.2023).

[3]https://npmtrends.com/angular-vs-react-vs-vue — npm trends for React, Angular, Vue (дата обращения 26.10.2023).

ционала по работе с графиками. Этот функционал необходим для возможности сравнивать работу разных алгоритмов. Помимо реализации стандартного интерфейса необходимо, чтобы была возможность динамически обновлять графики при получении новых данных, строить интервалы предсказаний. Поэтому было решено дополнительно использовать библиотеку d3.js[19], так как она позволяет, например, использовать масштабирование, даёт возможность отрисовывать большое количество компонентов на графике. Также библиотека содержит большое количество примеров, что облегчает ее освоение и использование.

5. Реализация

В практических задачах особенностью временных рядов является изменчивость направления тренда с течением времени. Тренд не является статичным и может меняться под влиянием различных факторов, таких как, например, случайные колебания или изменения в поведении пользователей. Для эффективного анализа временных рядов и своевременного выявления изменений в тренде необходимо было включить применение методов, которые способны адаптироваться к измененияющимся условиям. В этой главе представлено описание таких подходов в реализации системы, а также описана реализация модификации алгоритма SPS для системы.

5.1. Режим скользящего окна

Одним из методов также является использование режима скользящего окна. Подход позволяет непрерывно обновлять анализируемую информацию, включая только последние данные, тем самым обеспечивая быструю реакцию на изменения в динамике ряда. Это делает анализ более локальным. При фиксировании изменения направления тренда появляется возможность корректировать предсказания.

Метод заключается в том, что выделяется некоторое число данных, окно, на основе которого выполняется основная работа всех алгоритмов. Окно перемещается, «скользит» по всему исходному набору данных, и для каждого окна вычисляются свои характеристики, такие как, например, наклон линии или коэффициенты регрессионной модели.

Сравнение работы алгоритма с режимом скользящего окна и без него представлено на рисунках 3 и 4.

5.2. Интервал прогнозирования

Хотя метод скользящего окна обладает значительными преимуществами, точность определения тренда не является абсолютной, и оценки тренда могут быть подвержены неопределенности. Для учета этой

Рис. 3: Работа алгоритма с режимом скользящего окна

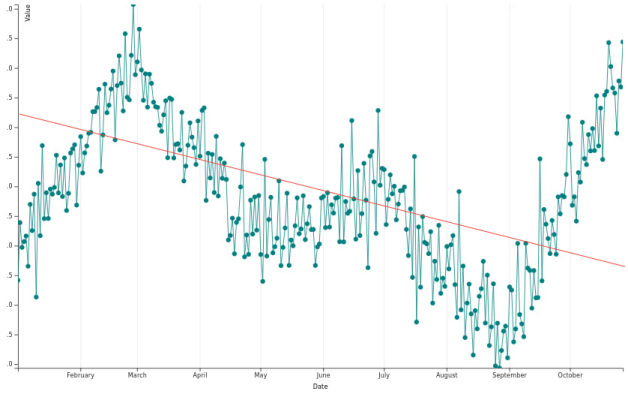

Рис. 4: Работа алгоритма без использования режима скользящего окна

неопределенности рекомендуется использовать доверительные интервалы. Доверительный интервал представляет собой диапазон значений, в котором с заданной вероятностью (уровнем доверия) находится истинное значение параметра тренда. Он даёт возможность оценивать степень уверенности в полученных результатах и учитывать возможные отклонения. При этом следует помнить, что размер окна влияет на количество наблюдений, используемых для оценки тренда, что, в свою очередь, влияет на ширину доверительного интервала.

5.3. Вычисление момента изменения тренда

Ключевой задачей в анализе временных рядов является определение момента времени, в который тренд меняет своё направление. Этот процесс позволяет идентифицировать ключевые точки, в которых поведение данных изменяется. Алгоритм, реализованный в системе, основан

на вычислении угла наклона линии тренда в каждом окне. После вычисления углы сравниваются. Если разница превышает заданный порог (например, 30 градусов), это указывает на возможное изменение тренда. В этом случае последняя точка текущего окна отмечается как точка изменения тренда.

На рисунке 5 представлен пример построения графика алгоритма системы. Моменты изменения тренда отмечены вертикальными синими линиями.

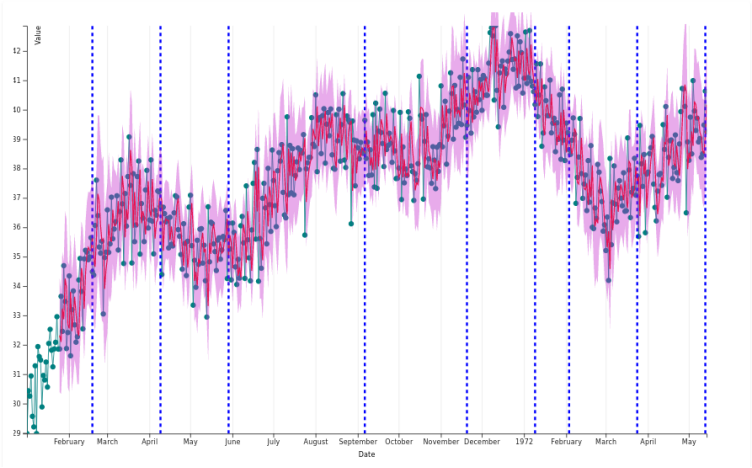

Рис. 5: Пример построения графика алгоритма системы

5.4. Модификация SPS

В случае определения доверительных интервалов для тренда модель наблюдений перепишется следующим образом:

$$Y = X^T\theta + E,$$

где $Y = \begin{pmatrix} x_1 \\ \vdots \\ x_T \end{pmatrix}$ — вектор значений тренда;

$X = \begin{pmatrix} t_1 & t_2 & \dots & t_T \\ 1 & 1 & \dots & 1 \end{pmatrix}$ — матрица признаков;

$\theta = \begin{pmatrix} k \\ b \end{pmatrix}$ — вектор параметров k, b, которые необходимо оценить;

$E = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{pmatrix}$ — вектор случайных ошибок.

Далее выполняются следующие шаги инициализации.

1. Выбор M, q.

2. Заполняется матрица знаков

$$SIGNS = \{\beta_{t,j}\} = \begin{pmatrix} 1 & \pm 1 & \cdots & \pm 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \pm 1 & \cdots & \pm 1 \end{pmatrix}_{T \times M}$$

   где каждый из знаков ± 1 выбирается случайным образом с вероятностью $\frac{1}{2}$.

3. Строится доверительное множество $\Theta = \{\theta \in \mathbb{R}^2 | SPS_{\text{indicator}}(\theta) = \text{True}\}$.

Процедура $SPS_{\text{indicator}}$:

1. Заполняется матрица для подсчёта знаковозмущённых сумм

$$\Delta = \{\beta_{i,j}(x_j - kt_j - b)\} = \begin{pmatrix} \varepsilon_1(\theta) & \pm\varepsilon_1(\theta) & \cdots & \pm\varepsilon_1(\theta) \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_T(\theta) & \pm\varepsilon_T(\theta) & \cdots & \pm\varepsilon_T(\theta) \end{pmatrix}$$

2. Генерируются знако-возмущённые суммы (элементы матрицы Δ суммируются по столбцам)

$$H_0(x) = \sum_{j=1}^{T}(x_j - kt_j - b) = \sum_{j=1}^{T}\varepsilon_j(\theta)$$

$$H_i(x) = \sum_{j=1}^{T}\beta_{i,j}(x_j - kt_j - b) = \sum_{j=1}^{T}\beta_{i,j}\varepsilon_j(\theta)$$

   для $i = 1...M-1$.

3. В случае если ранг H_0 $r \geq q$, то прямая с коэффициентами k, b лежит в доверительном множестве.

Финальным доверительным интервалом для точки является расстояние между самой самой верхней прямой с коэфициентами из множества Θ и самой нижней (рис. 6).

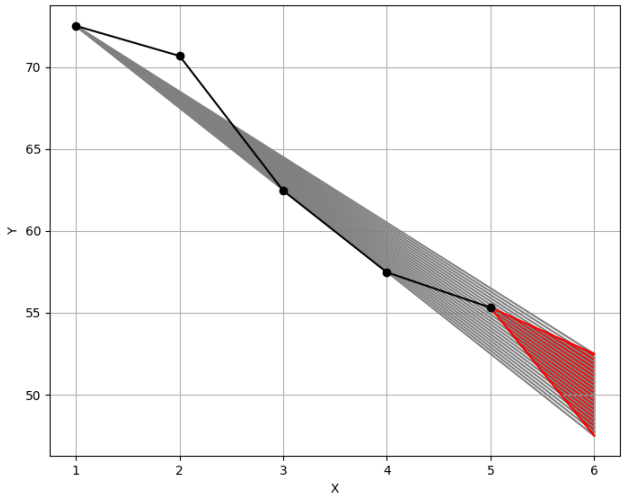

Рис. 6: Пример построения доверительного интервала с помощью модификации SPS

6. Эксперименты

В главе представлено сравнение результатов работы алгоритмов, реализованных в системе.

6.1. Данные и параметры алгоритмов

Было решено использовать сгенерированные данные, так как для них точно известны моменты изменения тренда. Тестирование проводится на двух наборах данных. Первый набор содержит 500 зашумлённых измерений и шесть изменений тренда, второй построен на основе первого с помощью операции прореживания и содержит 50 измерений. В ходе тестирования необходимо проверить, как алгоритмы справляются с определением тренда на малом числе входных данных и как быстро обнаруживают изменение тренда.

Первый набор будет тестироваться в режиме скользящего окна размером 50, второй с соответствующим ему по количеству данных окном размера 10. Параметры алгоритмов выбираются оптимальным образом. Так для определения параметров экпоненциального сглаживания параметры подбираются путём максимизации логарифмического правдоподобия, параметры для модели ARIMA подбираются с помощью проведения различных статистических тестов (например, расширенного теста Дикки-Фуллера), оптимизации информационных критериев. Более подробно о подборе параметров всех алгоритмов можно почитать в документации соответствующих библиотек.

6.2. Анализ определения тренда и построения доверительных интервалов

Для оценки точности предсказания считаются среднее абсолютное отклонение предсказанных значений от фактических (MAE), корень из среднеквадратичного отклонения (RMSE), средняя абсолютная ошибка в процентном выражении (MAPE). Для всех алгоритмов строятся

95%-ые доверительные интервалы. Для оценки доверительных интервалов вычисляются доля фактических значений, которые попадают в предсказанный доверительный интервал (CR) и средняя ширина доверительных интервалов (CW). Результаты тестирования алгоритмов на временных рядах представлены в таблице 2.

Ряд	Метрика	Holt	Регрессия	Kalman	ARIMA	SPS
1	MAE	0.9962	0.6469	0.5751	0.5707	0.8534
	RMSE	1.2310	0.8053	0.7088	0.7310	1.0651
	MAPE	2.6625	1.7290	1.5363	1.5272	2.2649
	CR	0.8105	0.9103	0.9124	0.9674	0.9888
	CW	3.3396	2.6357	4.1720	3.2580	5.3217
2	MAE	1.4829	0.9759	0.7609	1.2022	1.3789
	RMSE	1.7676	1.2240	0.9139	1.4312	1.6950
	MAPE	3.8779	2.5466	1.9809	3.1367	3.5990
	CR	0.9024	0.8536	0.9512	0.9756	1.0000
	CW	6.4080	3.3932	8.4067	5.8717	13.313

Таблица 2: Результаты анализа метрик для различных алгоритмов

6.3. Анализ моментов изменения тренда

При анализе стоит учитывать, что моменты изменения тренда зависят не только выставленных параметров алгоритмов, но и от размера скользящего окна. Так, если необходимы изменение глобальных тенденций, то размер скользящего окна должен быть как можно больше. Если же нужно принимать решения сиюминутно, то размер окна должен быть как можно меньше. Выбор оптимального окна зависит от специфики данных и конкретной задачи.

Для оценки моментов изменения трендов считаются точность (Precision) и полнота (Recall) и среднее время задержки (Mean Delay Time, MDT). Алгоритмы определения трендов работают с некоторым запаздыванием или опережением, поэтому первые две метрики адаптированы, т.е. при их подсчёте учитывается интервал времени, в течение которого обнаружение изменения тренда считается допустимым. В случае для первого временного ряда это 30 измерений, для второго —

три измерения. Также важно отметить, что при прореживинии данных некоторые из трендов удаляются. Если при наборе в 500 измерений было 6 изменений, то после прореживания осталось лишь 4. Учитывая всё это, были получены следующие результаты (табл. 3).

Ряд	Алгоритм	MDT	Precision	Recall
1	Holt	15.0892	0.8	0.5714
	Регрессия	15.1734	0.8	0.7142
	Калман	-22.0143	0.75	0.4285
	ARIMA	22.3456	1.0	0.7142
	SPS	-5.7896	1.0	0.8571
2	Holt	-0.25	0.50	0.50
	Регрессия	1.25	0.75	0.75
	Калман	-1.25	0.75	0.75
	ARIMA	1.25	0.75	0.75
	SPS	-2.75	1.00	1.00

Таблица 3: Результаты анализа метрик оценки изменений тренда

6.4. Выводы

Проведённые эксперименты продемонстрировали различную эффективность алгоритмов при анализе временных рядов, содержащих изменения тренда, в зависимости от объёма данных и настроек параметров. Ниже приведены ключевые выводы.

Алгоритм ARIMA показал наивысшую точность предсказаний по метрикам MAE, RMSE и MAPE на первом наборе данных, что указывает на его способность хорошо моделировать сложные временные ряды с множеством изменений тренда. Лучшие результаты по доле фактических значений, попадающих в предсказанный доверительный интервал, показали модели ARIMA и SPS.

Оценки параметров, полученные при выполнении некоторых естественных условий с помощью метода наименьших квадратов, используемого в линейной регрессии, стремятся к истинным значениям параметров системы с ростом объема данных. При увеличении объема данных ошибка оценки параметров приближается к нормальному распреде-

лению. Это свойство позволяет использовать результаты центральной предельной теоремы для построения приближенных доверительных областей вокруг оценок параметров.

Однако эти доверительные области имеют ограничения: они гарантированы только асимптотически, когда количество данных стремится к бесконечности. В случае конечного количества данных доверительные области, построенные на основе метода наименьших квадратов, становятся лишь эвристическими — они не имеют строгих теоретических гарантий, что может означать повышенный риск недооценки вариативности данных, в то время как метод SPS их имеет, хоть и даёт наибольшие значения ширины доверительных интервалов при разном количестве данных. Также этот метод продемонстрировал наивысшую точность и полноту при обнаружении изменений тренда на обоих наборах данных, что делает его предпочтительным выбором для задач, требующих точного определения изменений тренда.

7. Заключение

В ходе данной работы были получены следующие результаты.

1. Сформулированы требования к системе: определены режимы работы с данными (обработка данных в режиме реального времени, генерация данных, обработка готовых наборов данных); сформулированы требования по настройке алгоритмов и некоторые другие.

2. Перед реализацией системы были проведены:

   - исследование следующих алгоритмов для предсказания временных рядов и определения тренда в них: метод Хольта, линейная регрессия, авторегрессионное интегрированное скользящее среднее, фильтр Калмана, метод знаковозмущенных сумм;
   - обзор существующих решений: открытые системы Orange и KNIME, библиотечные пакеты forecast и Prophet;
   - сделаны выводы о том, что существующие решения не удовлетворяют сформулированным требованиям; было решено реализовать представленные в обзоре алгоритмы.

3. Разработана архитектура системы: определены основные компоненты и механизмы их взаимодействия, выбраны технологии – решено разрабатывать веб-приложение (Python, Javascript) с использованием библиотек Flask (серверная часть), React (клиентская часть), Material UI (компоненты клиентской части), d3.js (отрисовка графиков).

4. Создан прототип системы, использующий для предсказания и определения тренда скользящее окно, построение доверительного интервала и вычисление моментов изменения тренда.

5. Проведены эксперименты, которые показали, что для различных типов задач и условий могут быть предпочтительны разные ал-

горитмы, и важно тщательно подбирать параметры и методы в зависимости от поставленной задачи:

- лучшую точность при большом числе данных в задаче определения тренда показала модель ARIMA;
- в задаче определения моментов изменения тренда наилучшим образом продемонстрировал себя метод SPS при разном числе входных данных;
- наилучшие показатели по доле фактических значений, попадающих в предсказанный доверительный интервал, продемонстрировали ARIMA и SPS (при этом метод SPS даёт максимальную ширину доверительных интервалов);
- ширина доверительных интервалов оказалась наименьшей у регрессии, что указывает на более узкие предсказанные диапазоны, но это также может означать повышенный риск недооценки вариативности данных.

Исходный код доступен по ссылке на репозиторий[4] Github.

Результаты работы были представлены на конференции «Современные технологии в теории и практике программирования» и опубликованы в сборнике её материалов [21].

[4]https://github.com/Liana2707/TimeSeriesForecastingAlgorithms — репозиторий проекта на Github (дата обращения: 27.05.2024).

Список литературы

[1] Alteryx. — https://www.alteryx.com/. — Accessed: 14.12.2023.

[2] Amibroker. — https://www.amibroker.com/. — Accessed: 14.12.2023.

[3] Angular. — https://angular.io/. — Accessed: 24.04.2024.

[4] Csáji Balázs, Campi Marco, Weyer Erik. Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non-Asymptotic Confidence Regions in Linear Regression Models // Signal Processing, IEEE Transactions on. — 2015. — 01. — Vol. 63. — P. 169–181.

[5] Django Framework. — https://www.djangoproject.com/. — Accessed: 24.04.2024.

[6] Flask Framework. — https://flask.palletsprojects.com/en/3.0.x/. — Accessed: 24.04.2024.

[7] Hamilton James D. Time Series Analysis. — PRINCETON, NEW JERSEY : PRINCETON UNIVERSITY PRESS, 1994.

[8] Hyndman Rob J., Athanasopoulos George. Forecasting: principles and practice. — OTexts, 2014.

[9] KNIME. — https://www.knime.com/. — Accessed: 14.12.2023.

[10] Kalman Rudolph Emil. A New Approach to Linear Filtering and Prediction Problems // Transactions of the ASME–Journal of Basic Engineering. — 1960. — Vol. 82, no. Series D. — P. 35–45.

[11] Material UI. — https://mui.com/material-ui/. — Accessed: 24.04.2024.

[12] MetaTrader. — https://www.metatrader5.com/ru. — Accessed: 14.12.2023.

[13] Orange. — https://orangedatamining.com/. — Accessed: 14.12.2023.

[14] Prophet. — https://facebook.github.io/prophet/. — Accessed: 14.12.2023.

[15] SAS (Statistical Analysis System). — https://www.sas.com/en_us/software/stat.html. — Accessed: 14.12.2023.

[16] TIBCO Spotfire. — https://www.spotfire.com/. — Accessed: 14.12.2023.

[17] TradingView. — https://www.tradingview.com/. — Accessed: 14.12.2023.

[18] Vue.js. — https://vuejs.org/. — Accessed: 24.04.2024.

[19] d3,js. — https://d3js.org/. — Accessed: 24.04.2024.

[20] forecast. — https://github.com/robjhyndman/forecast. — Accessed: 14.12.2023.

[21] Л. И. Нафикова О. Н. Граничин. Прототип системы определения тренда во временных рядах // Современные технологии в теории и практике программирования: Сборник материалов научно-практической конференции студентов, аспирантов и молодых ученых. — Санкт-Петербург : Санкт-Петербургский политехнический университет Петра Великого, 2024. — P. 76–77. — EDN VHXLBC.

[22] Марина Волкова. Рандомизированные алгоритмы оценивания параметров инкубационных процессов в условиях неопределённостей и конечного числа наблюдений : Диссертация / Волкова Марина ; СПбГУ. — 2018.