Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУ «Информатика и системы управления»

КАФЕДРА ИУ-7 «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

НА ТЕМУ:

«Метод определения аномалий в статистике посещений организации сотрудниками и посетителями на основе статистических методов»

Студент **ИУ7-86Б** _____ **А. О. Виноградов**
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель ВКР _____ **Н. В. Назаренко**
(Подпись, дата) (И.О.Фамилия)

Нормоконтролер _____ _____
(Подпись, дата) (И.О.Фамилия)

2024 г.

РЕФЕРАТ

Выпускная квалификационная работа содержит 49 с., 11 рис., 4 табл., 38 ист.

Ключевые слова: СТАТИСТИКА, ВРЕМЕННОЙ РЯД, АНОМАЛИЯ, АВТОРЕГРЕССИОННАЯ МОДЕЛЬ, ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ,

Объектом исследования являются методы выявления аномалий. Объект разработки — метод выявления аномалий. Цель работы — разработка комбинированного метода выявления аномалий на основе статистических методов.

Область применения — статистика посещений организации.

В аналитическом разделе проведен анализ предметной области, выполнен обзор существующих программных решений поставленной задачи, проведен сравнительный анализ методов выявления аномалий.

В конструкторском разделе описан разработанный метод, состоящий из обработки набора данных существующим алгоритмом выявления аномалий и последующим подтверждением корректности выявленных аномалий при помощи другого алгоритма.

В технологическом разделе выполнено обоснование программных средств разработки и описание программной реализации разработанного метода.

В исследовательском разделе проведено исследование влияния параметров разработанного метода на качество обнаружения аномалий в различных наборах данных.

Полученные результаты показывают применимость разработанного метода в сфере статистики посещений организации.

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ

В настоящей выпускной квалификационной работе используются следующие термины:

— **Аномалия** — наблюдение или набор наблюдений, которые значительно отличаются от других данных и не соответствуют общему шаблону.

— **Временной ряд** — последовательность значений, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени, обычно через равные промежутки.

— **Выброс** — результат измерения, выделяющийся из общей выборки.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящей выпускной квалификационной работе используются следующие обозначения и сокращения:

— **XB2** — метод Хольта-Винтерса с двумя сезонностями.

— **СКУД** — система контроля и управления доступом.

— **ПО** — программное обеспечение.

— **CSV** — Comma-Separated Values, текстовый формат, предназначенный для представления табличных данных.

— **AR** — Auto-Regressive model.

— **CLR** — Common Language Runtime.

— **NAB** — Numenta Anomaly Benchmark.

ВВЕДЕНИЕ

В условиях современного бизнеса анализ статистики посещений организации является неотъемлемой частью стратегического управления и принятия решений, менеджмента персонала и безопасности контрольно-пропускного режима. На основе этой статистики могут приниматься более верные и своевременные решения по управлению ресурсами, обеспечения безопасности и оптимизации производственных процессов на предприятии.

Выявление аномалий в данных статистики посещений позволяет повышать эффективность работы с учетом пиковых нагрузок посетителей, оценивать пунктуальность сотрудников и выявлять нарушения трудового график, выявлять поломки оборудования и нарушения в работе контрольно-пропускного режима.

Цель данной работы заключается в разработке и компьютерной реализации комбинированного метода выявления аномалий на основе статистических методов.

Для достижения поставленной цели необходимо выполнить следующие задачи:

1. проанализировать методы выявления аномалий в статистике посещений организации и провести их сравнительный анализ;
2. выбрать необходимые для работы метода алгоритмы;
3. разработать метод выявления аномалий в статистике посещений;
4. создать программную реализацию разработанного метода;
5. исследовать зависимость эффективности работы реализованного метода от значений параметров системы.

1 Аналитический раздел

В этом разделе представлен анализ предметной области, постановка и формализация задачи. Здесь также рассматриваются типы аномалий и существующие алгоритмы их выявления, включая их особенности, преимущества и недостатки. После определения критериев сравнения проведен сравнительный анализ алгоритмов и методов, результаты которого представлены в виде таблицы. Также здесь рассмотрены способы объединения методов и обоснован выбор одного из них.

1.1 Анализ предметной области

Временной ряд [1] — это последовательность данных, которая упорядочена по времени или другому типу последовательности. Каждая точка данных во временном ряде соответствует определенному моменту времени или периоду времени, и данные обычно собираются с равными интервалами времени. Например, временными рядами являются данные о ценах акций на бирже, погодных показателей (температура, количество осадков), пульс и уровень глюкозы в крови. Пример временного ряда приведен на рисунке 1.

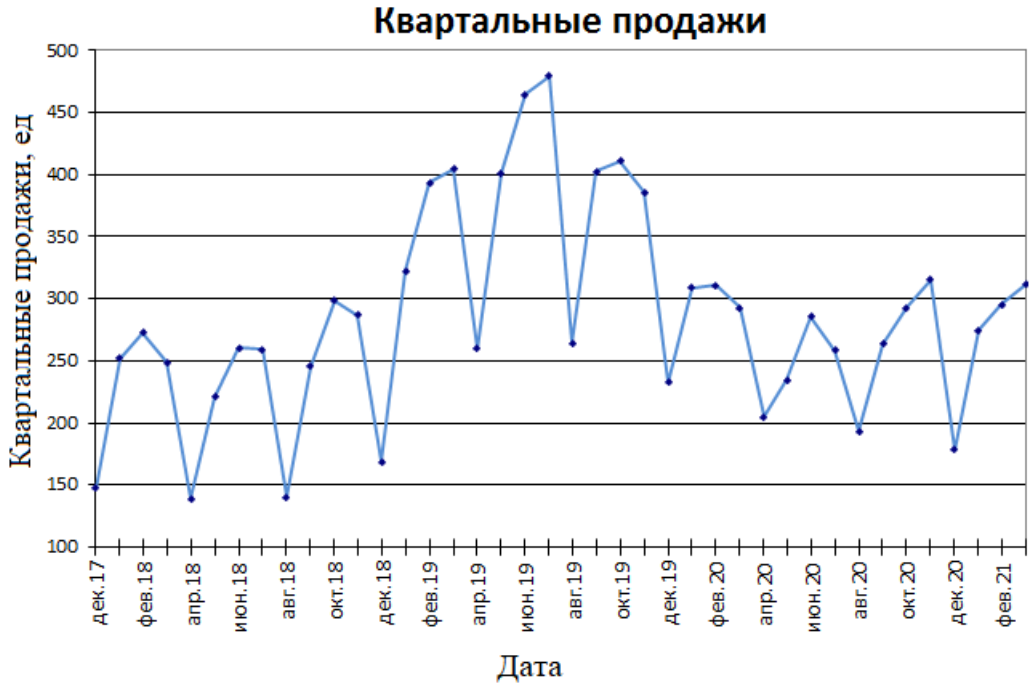

Рисунок 1 – Пример временного ряда

Особенностью временных рядов является наличие внутренней структуры, учитывающей множество компонент:

— сезонность — повторяющиеся изменения, связанные с определенными периодами, например, время года;

— циклы — изменения, происходящие с определенной регулярностью, но не обязательно связаны с календарными периодами;

— тренды — долгосрочные изменения в данных, которые могут быть восходящими или нисходящими;

— случайные колебания — непредсказуемые изменения, которые не следуют определенному шаблону.

Анализ и прогнозирование временных рядов широко применяется в различных областях, таких как экономика и медицина. В экономике временные ряды используются для анализа и предсказания цен акций, уровня безработицы и валового внутреннего продукта.

1.1.1 Статистика посещений предприятия

Статистика посещений предприятия включает в себя сбор, обработку и анализ данных о количестве визитов сотрудников и посетителей на предприятие. Данная информация является важной для управления ресурсами, обеспечения безопасности и оптимизации процессов на предприятии.

Для анализа посещений используются различные методы сбора данных, такие как:

— системы контроля и управления доступом [2] — электронные пропускные системы, фиксирующие время входа и выхода каждого сотрудника и посетителя;

— регистры и журналы посещений — ручной учет, который ведется на контрольно-пропускных пунктах;

— системы видеонаблюдения — позволяют отслеживать потоки людей в реальном времени и собирать данные для последующего анализа.

Количество посетителей и перемены в частоте посещений могут говорит о снижении и повышении популярности предприятия, изменениях в поведении и расписании людей, на которые требуется отвечать изменениями бизнес-процессов для максимизации прибыли.

Кроме того, статистика посещений работниками предприятия позволяет оценивать пунктуальность сотрудников и выявлять нарушения рабочего графика, а также корректировать графики работы и распределение ресурсов с учетом данных о пиковых нагрузках, повышая общую эффективность предприятия.

В данной работе рассматриваются данные о количестве работников или посетителей в помещении организации в определенный момент времени как временной ряд. Причем промежутком снятия показаний будет являться день, то есть анализу подлежат данные о количестве посещений предприятия в тот или иной день.

Формализованные таким образом данные, очевидно, будут обладать цикличностью — в зависимости от дня недели, месяца или времени года количество посетителей будет циклично увеличиваться и уменьшаться. Также при такой формализации данных могут наблюдаться тренды на увеличение или уменьшение числа посетителей, связанные, например, с набором или потерей популярности организации или ее сферы деятельности.

1.2 Понятие аномалии

Аномалия (выброс, отклонение, исключение) [3] — это отклонение поведения системы от стандартного (ожидаемого). Обнаружение аномалий в статистическом наборе данных связано с обнаружением отклонений или необычных паттернов, которые выходят за пределы ожидаемого поведения данных. Аномалии могут указывать на ошибки в данных, необычные события или изменения в процессе, которые могут быть важными для понимания.

Наличие аномалий может отражать различные отклонения от «нормальной» работы системы.

В качестве примеров можно описать следующие ситуации:

— попытки взлома системы или сервиса на основании показаний сетевого трафика;

— выявление нарушения рабочего графика на основании необычно ранних или поздних приходов сотрудников;

— обнаружение поломки в системе контрольно-пропускного пункта на основе анализа показаний количества проходов посетителей.

В вышеописанных ситуациях большую часть анализируемых данных, составляют некоторые значения, характеризующие стандартную, нормальную работу системы. Однако при возникновении данных ситуаций, эти показатели будут принимать некоторые отклоняющиеся от типовых, значения. Поиск отклонений в этих показателях может позволить определить, предотвратить или спрогнозировать возникновение нештатных, «аномальных» ситуаций.

1.2.1 Типы аномалий

Аномалии бывают различны по своей сути и проявляются в различных формах. Чаще всего отклонения в данных относят [4] к одному из трех основных типов.

1. Точечные аномалии — возникают в ситуации, когда отдельный экземпляр данных может рассматриваться как аномальный по отношению к остальным данным. Пример точечной аномалии приведен на рисунке 2.

2. Групповые аномалии — возникают в ситуации, когда группа точек ведет себя аномально, хотя каждая из точек по-отдельности аномальной не является. Пример групповой аномалии приведен на рисунке 3.

3. Контекстуальные аномалии — наблюдаются, когда экземпляр данных является аномальным только в определенном контексте, не связанным со значениями набора данных. В качестве примера можно привести отрицательное количество посетителей предприятия в день наблюдения.

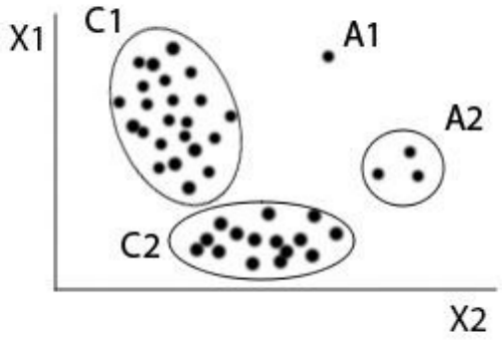

Рисунок 2 – Пример точечных аномалий. Точка A1 и группа A2 являются аномальными относительно признанных «нормальными» групп C1, C2.

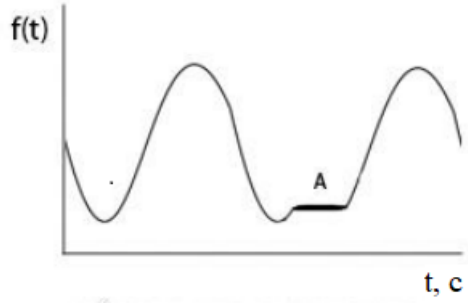

Рисунок 3 – Пример групповой аномалии. Группа точек A является аномалией относительно общего поведения временного ряда.

1.3 Существующие решения

В настоящее время существует множество программных решений для поиска аномалий в данных. Многие из них представлены в виде различных коммерческих сервисов, таких как:

— Microsoft Azure Anomaly Detector [5] — сервис на базе облачной платформы Microsoft Azure, предназначенный для обнаружения аномалий в данных временных рядов на основе подписки. Пользователи платят за использование сервиса в зависимости от объема обработанных данных и количества запросов.

— Anodot [6] — платформа для аналитики данных в реальном времени, которая использует машинное обучение для обнаружения аномалий. Данный сервис также работает на подписочной основе, а платежи зависят от объема данных и числа анализируемых метрик.

Данные иностранные сервисы предоставляют только услуги на основе подписки и не предусматривают вычислений на мощностях пользователя.

Это значит, что данные сервисы не подходят для обработки коммерческих данных, так как требуют отправки персональных данных посетителей на сторонние сервера.

Также существуют бесплатные решения с открытым кодом. В качестве примеров можно привести:

— Atlas [7] — платформа для мониторинга данных и выявления аномалий, используется для анализа метрик и оперативного обнаружения отклонений в потоках данных;

— Morgoth [8] — система для обнаружения аномалий, которая применяется для мониторинга и анализа временных рядов, поддерживающая адаптивные алгоритмы машинного обучения, что позволяет работать с различными типами данных;

— LinkedIn's ThirdEye [9] — инструмент для мониторинга данных, использующий комбинацию статистических методов и алгоритмов машинного обучения для идентификации аномалий.

Для сравнения были выделены следующие критерии:

1. открытый код — наличие открытого кода и возможности использовать продукт бесплатно;
2. локальность — возможность развернуть решение на локальной машине;
3. двойная сезонность — возможность продукта правильно оценивать аномалии в данных с двумя сезонностями;
4. комбинированные методы — применение комбинированных методов выявления аномалий для увеличения точности и уменьшение количества ложных срабатываний.

Сравнительный анализ методов по данным критериям приведен в таблице 1.

Таблица 1 – Сравнительная таблица программных решения для выявления аномалий

Решение	**Открытый код**	**Локальность**	**Двойная сезонность**	**Комбинированные методы**
Microsoft Azure	-	-	+	+
Anodot	-	-	-	+
Atlas	+	+	+	-
Morgoth	+	+	-	-
LinkedIn's ThirdEye	+	+	-	+

Ни одно из рассмотренных решений не смогло удовлетворить всем критериям, следовательно есть востребованность в реализации собственного ПО, отвечающего на все перечисленные запросы.

1.4 Постановка и формализация задачи

Задача, поставленная в данной работе может быть сформулирована следующим образом: дан массив данных о посещениях организации сотрудниками и посетителями. Необходимо разработать комбинированный алгоритм на основе статистических методов, с помощью которого определить наличие аномалий в данных и сформировать массив аномальных значений.

Формализация данной задачи может быть представлена в виде IDEF0 диаграммы (рисунок 4).

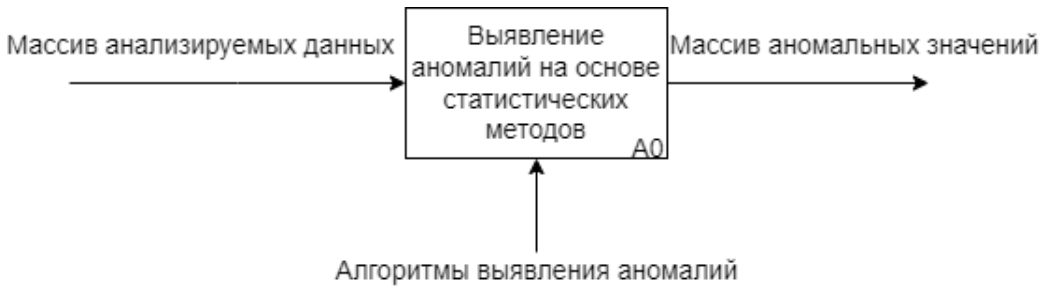

Рисунок 4 – Схема формализованной постановки задачи

В качестве ограничения, входные данные представляют собой массив числовых значений, так как алгоритмы выявления аномалий работают с количественными параметрами. Также данное ограничение позволяет обойтись без дополнительной предварительной обработки данных и сохранить возможность анализа одномерных значений, так как большинство алгоритмов работают с одномерными массивами данных.

1.5 Методы выявления аномалий

Ниже приведена классификация основных методов выявления аномалий. Далее будут рассмотрены примеры методов каждого класса.

1. Вероятностные методы — включают оценку стандартных отклонений, диапазонов значений и других статистических характеристик данных. Аномальными определяются точки, которые значительно отличаются от среднего или медианного значения.

2. Методы машинного обучения — применяют алгоритмы машинного обучения, такие как методы кластеризации, выбросов, ансамблевые методы.

3. Статистические методы — на основе статистической модели производится прогноз доверительных интервалов, в которых может находиться значение следующего экземпляра данных. Точки, реальные значения которых не попадают в доверительный интервал, признаются аномальными. Эффективность выявления аномалий при таком подходе зависит от правильного подбора статистической модели, наиболее точно описывающей анализируемые данные.

1.6 Вероятностные методы выявления аномалий

К вероятностным методам относят оценку стандартных отклонений, диапазонов значений и других статистических характеристик данных. Аномальными определяются точки, которые значительно отличаются от среднего или медианного значения. Эти методы позволяют обнаруживать выбросы и отклонения в данных, основываясь на их вероятностных характеристиках. Одним из основных преимуществ вероятностных методов является их способность учитывать естественные колебания данных и устанавливать гибкие пороги для выявления аномалий. В

1.6.1 Метод Z-скорректированных отклонений

Z-скорректированные отклонения [10] — один из методов выявления аномалий во временных рядах, основанный на стандартном отклонении и среднем значении ряда.

В начале работы вычисляется среднее значение μ и стандартное отклонение σ для временного ряда.

Затем для каждой точки временного ряда по формуле (1.1) вычисляется Z-скорректированное значение, то есть количество стандартных отклонений, на которое данная точка отклоняется от среднего значения временного ряда.

$$Z_t = \frac{y_t - \mu}{\sigma}, \tag{1.1}$$

где y_t и Z_t — значение временного ряда в точке t и Z-скорректированное значение для этой точки соответственно.

Затем устанавливается пороговое значение отклонения Z. Как правило [11], устанавливают пороговое значение, равное 3, то есть аномальными признаются точки, значение которых отклонилось от среднего более чем на 3 стандартных отклонения.

К минусам данного метода относится неправильное определение аномалий при наличии выбросов и требовательность к данным: метод хорошо работает только на нормально распределенных данных. Кроме того, метод может недооценивать аномалии в данных, обладающих ярко выраженной сезонностью.

1.6.2 Метод кумулятивных сумм

Метод кумулятивных (накопленных) сумм [12] использует накопленные суммы ошибок между фактическими и прогнозируемыми значениями ряда.

Для каждого момента времени t вычисляются накопленные суммы S_t^+ и S_t^-, которые представляют собой суммы положительных и отрицательных отклонений между фактическими и прогнозируемыми значениями ряда соответственно. Начальные значения накопленных сумм обычно равны нулю.

Затем анализируются значения накопленных сумм. Если значение S_t^+ и S_t^- превышает некоторый порог h, который определяется пользователем, это может свидетельствовать о наличии аномалии во временном ряду в момент времени t. После обнаружения аномалии, накопленные суммы обычно обнуляются или устанавливаются в ноль, чтобы продолжить мониторинг следующих значений ряда.

Недостатками данного метода является требование к настройке параметра h, значение которого находится экспериментально. Также данный метод может быть малоэффективен при обнаружении краткосрочных и редких аномалий.

1.7 Методы машинного обучения

Ниже перечислены наиболее популярные методы, включаемые в данную категорию [13].

1. Изолирующий лес [13] — данные случайным образом разделяются, строя деревья решений, пока каждое наблюдение не будет выделено в своем собственном листе. Аномальными признаются наблюдения, заканчивающиеся на более коротком пути от корня дерева, так как аномалии требуют меньшего числа разделений до изоляции.

2. Гауссовская смесь распределений [13] — модель вероятностного распределения, которая представляет собой комбинацию нескольких гауссовских распределений. Гауссовская смесь моделирует это смешанное распределение путем совмещения (смешивания) нескольких гауссовских компонент, аномальными признаются точки данных, имеющие низкую вероятность по сравнению с другими точками.

3. Метод опорных векторов [13] — основан на поиске оптимальной гиперплоскости, которая лучше всего разделяет два класса данных в пространстве признаков. Основная идея метода заключается в том, чтобы максимизировать расстояние между гиперплоскостью и ближайшими точками обучающего набора данных, которые называются опорными векторами. Аномальными признаются наблюдения, находящиеся слишком далеко от построенной гиперплоскости.

4. Ансамблевые алгоритмы [13] — такие методы, как бэггинг, бустинг и стекинг, позволяющие объединять нескольких базовых моделей для получения более точных и устойчивых прогнозов.

Общими недостатками методов машинного обучения являются:

1. требование больших объемов [14] данных для обучения;

2. потребность в эмпирической настройке гиперпараметров;

3. сложность интерпретации — объяснение принимаемых решений и выявление причин ошибок работы является затруднительным.

1.8 Статистические методы

В данной группе методов значение признается аномальным, если оно выбивается из доверительного интервала прогноза, построенного на основе статистической модели. Чаще всего применяются методы на основе экспоненциального сглаживания или авторегрессионные модели.

Статистические методы позволяют учитывать сезонные и трендовые компоненты данных, что делает их полезными в различных областях, включая экономику, инженерные науки и управление запасами. Основное преимущество этих методов заключается в их способности адаптироваться к изменениям в данных, что повышает точность прогнозов и выявления аномалий.

1.8.1 Методы экспоненциального сглаживания

Далее рассматриваются наиболее популярные методы экспоненциального сглаживания [15]:

1. метод простого экспоненциального сглаживания (метод Брауна);
2. метод тройного экспоненциального сглаживания (метод Хольта-Винтерса);
3. метод Хольта-Винтерса с двумя сезонностями (сокращенно ХВ2);

Особенность метода простого экспоненциального сглаживания заключается в том, что для предсказания используются только значения предыдущих уровней временного ряда, взятые с определенным весом.

1.8.1.1 Метод Брауна

В методе Брауна [15] прогноз на момент времени $t+1$ определяется по следующей формуле:

$$\hat{y}_{t+1} = \alpha \cdot y_t + (1 - \alpha) \cdot \hat{y}_t, \tag{1.2}$$

где:

— y_t — значение временного ряда в момент t;

— $\alpha \in (0, 1)$ — коэффициент сглаживания;

— $\hat{y}_t$ — предыдущий прогноз.

Параметр α определяет вес текущего наблюдения по сравнению с предыдущими значениями при вычислении прогноза. Если значение параметра

близко к 1, то больший вес придается последним наблюдениям, и модель быстрее адаптируется к последним изменениям в данных. Если α близко к 0, то все наблюдения имеют примерно одинаковый вес, и модель более устойчива к колебаниям в данных.

В свою очередь предыдущий прогноз $\hat{y}_t$ аналогичным образом зависит от прогноза $\hat{y}_{t-1}$.

Таким образом, обозначив $\beta = 1 - \alpha$, можно представить формулу (1.2) в следующем виде:

$$\hat{y}_{t+1} = \alpha \cdot y_t + \alpha \cdot \beta \cdot y_{t-1} + \ldots + \alpha \cdot \beta^k \cdot y_{t-k} + \ldots + \beta^t \cdot y_0, \qquad (1.3)$$

где y_i при $i = \overline{0, t}$ — значение временного ряда в момент i.

Главными недостатками данного метода являются:

1. неучет таких особенностей временного ряда, как цикличность и наличие трендов;
2. чувствительность к точности начальных значений и параметра α;
3. неучет внешних факторов, потенциально влияющих на поведение прогнозируемой величины.

1.8.1.2 Метод Хольта-Винтерса

Метод Хольта-Винтерса [16] является модификацией метода Брауна, учитывающей сезонность и трендовость в данных.

Прогноз на момент времени t определяется по следующей формуле:

$$\hat{y}_t = L_{t-1} + P_{t-1} + S_{t-m}, \qquad (1.4)$$

где:

— $L_t = \alpha \cdot (y_t - S_{t-m}) + (1-\alpha)(L_{t-1} + P_{t-1})$ — компонент уровня за последний период;

— $P_t = \beta \cdot (L_t - L_{t-1}) + (1 - \beta)P_{t-1}$ — компонент тренда за последний период;

— $S_t = \gamma \cdot (y_t - L_t) + (1 - \gamma)S_{t-m}$ — сезонный компонент за последний период;

— m — длина сезонного периода;

— $\alpha \in (0; 1)$ — коэффициент сглаживания уровня — вес текущего наблюдения по сравнению с предыдущими;

— $\beta \in (0; 1)$ — коэффициент сглаживания тренда;

— $\gamma \in (0; 1)$ — коэффициент сглаживания сезонности.

Параметр β определяет, какая доля изменения уровня временного ряда должна быть включена в оценку текущего тренда. Чем ближе значение этого параметра к 1, тем больший вес имеют последние изменения уровня. Когда β ближе к 0, тренд более устойчив к шуму, но медленнее адаптируется к изменениям.

Параметр γ определяет скорость изменения сезонных паттернов. Когда значение ближе к 1, модель быстрее реагирует на изменения в сезонности, но может быть менее устойчивой к шуму. Когда γ ближе к 0, модель более устойчива к шуму, но медленнее адаптируется к изменениям в сезонности.

Оптимальные параметры α, β, γ предлагается определять экспериментальным путем и исходя из вида данных.

В данном методе аномальными определяются точки временного ряда, чьи значения не попали в соответствующие им предсказанные доверительные интервалы отклонения. Для определения указанных интервалов используется алгоритм Брутлага [17]. Согласно ему доверительным признается интервал $[\hat{y}_t - p \cdot d_{t-m}; \hat{y}_t + p \cdot d_{t-m}]$, где:

— $\hat{y}_t$ — спрогнозированное значение;

— $d_t = \gamma \cdot |y_t - \hat{y}_t| + (1 - \gamma) d_{t-m}$ — прогнозируемое отклонение значения;

— m — длина сезонного периода;

— p > 0 — коэффициент доверительного интервала алгоритма Брутлага.

$$d_t = \gamma \cdot |y_t - \hat{y}_t| + (1 - \gamma) d_{t-m1}$$
$$w_t = \theta \cdot |y_t - \hat{y}_t| + (1 - \theta) w_{t-m2}$$

1.8.1.3 Метод Хольта-Винтерса с двумя сезонностями

Метод Хольта-Винтерса с двумя сезонностями [18] — модификация метода Хольта-Винтерса, которая применяется для данных, обладающий двумя сезонностями. Примером таких данных можно назвать и статистику посещений организации — количество посещений в день будет варьироваться как от дня недели, так и от времени года. Очевидно, что количество посещений во время выходных или ежегодных праздников значительно меньше, чем в рабочие дни.

Прогноз на момент времени t определяется по следующей формуле:

$$\hat{y}_t = L_{t-1} + P_{t-1} + S_{t-m1} + W_{t-m2}, \tag{1.5}$$

где:

- $L_t = \alpha \cdot (y_t - S_{t-m1} - W_{t-m2}) + (1-\alpha)(L_{t-1} + P_{t-1})$ — компонент уровня за последний период;
- $P_t = \beta \cdot (L_t - L_{t-1}) + (1-\beta)P_{t-1}$ — компонент тренда за последний период;
- $S_t = \gamma \cdot (y_t - L_t - W_{t-m2}) + (1-\gamma)S_{t-m1}$ — сезонный компонент первой сезонности за последний период;
- $W_t = \theta \cdot (y_t - L_t - S_{t-m1}) + (1-\theta)W_{t-m2}$ — сезонный компонент второй сезонности за последний период;
- m1 — длина более короткого сезонного периода;
- m2 — длина более длинного сезонного периода;
- $\alpha \in (0;1)$ — коэффициент сглаживания уровня — вес текущего наблюдения по сравнению с предыдущими;
- $\beta \in (0;1)$ — коэффициент сглаживания тренда;
- $\gamma \in (0;1)$ — коэффициент сглаживания более короткой сезонности;
- $\theta \in (0;1)$ — коэффициент сглаживания более длинной сезонности.

1.8.2 Авторегрессионные методы

Далее перечислены наиболее популярные [19] методы, относящиеся к данному классу:

1. авторегрессионная модель (AR);
2. модель авторегрессии — скользящего среднего (ARMA);
3. модель Бокса–Дженкинса (ARIMA);
4. модель ARIMAX.

1.8.2.1 AR-метод

AR-метод как правило используется для моделирования сезонности временных рядов. Предполагается, что каждый член временного ряда образуется при помощи p предыдущих.

Прогноз на момент времени t вычисляется по следующей формуле:

$$\hat{y}_t = c + \sum_{i=1}^{p} a_i y_{t-i} + \varepsilon_t, \tag{1.6}$$

где:

— $c = const$ — постоянный уровень временного ряда — среднее значение временного ряда;

— y_{t-i} для $i = \overline{1,p}$ — предыдущие значения временного ряда;

— a_i для $i = \overline{1,p}$ — коэффициенты авторегрессии;

— ε_t — случайная ошибка с нулевым средним и постоянной дисперсией;

— p — порядок авторегрессии.

Математический смысл параметра ε_t заключается в том, что он представляет случайную составляющую, которая не может быть объяснена моделью и вносит случайную ошибку в прогнозы. Параметр ε_t может быть оценен с помощью метода наименьших квадратов [20] или метода максимального правдоподобия [21], и его оценка может быть использована для проверки гипотез о значимости коэффициентов модели.

1.8.2.2 ARMA-метод

ARMA-метод [22] комбинирует метод AR и скользящую среднюю [23]. Данную модель можно представить следующим образом:

$$\hat{y}_t = c + \sum_{i=1}^{p} a_i y_{t-i} + \varepsilon_t + \sum_{j=0}^{q} b_j \varepsilon_{t-j}, \tag{1.7}$$

где:

— q — порядок скользящего среднего;

— b_j для $j = \overline{1, q}$ — коэффициенты скользящего среднего, определяющие влияние предыдущих ошибок на текущее значение.

Остальные обозначения соответствуют себе из формулы (1.6).

Эта модель хорошо подходит для прогнозирования данных, стабильно сохраняющих свою динамику на всем рассматриваемом временном периоде. Если данные не обладают, таким свойством, прогнозы получаются отличными от реальности.

1.8.2.3 ARIMA-метод

ARIMA-метод или метод Бокса-Дженкинса [14] — модификация ARMA, предполагающая, что данные имеют авторегрессию, случайные колебания и интеграцию:

$$\Delta^d y_t = c + \sum_{i=1}^{p} \Delta^d y_{t-1} + \sum_{j=0}^{q} b_j \varepsilon_{t-j} + \varepsilon_t. \tag{1.8}$$

Интеграция данных подразумевает наличие стабильной разности некоторого порядка. То есть $\Delta^d y_t = y_{t-d} - y_t$ сохраняет свой вид и поведение при любом t.

Метод считается точным [14] для краткосрочных и долгосрочных прогнозов, но требует [14] большого набора данных и трудоемкого анализа.

1.8.2.4 ARIMAX-метод

ARIMAX-метод [24] — модификация ARIMA, которая также позволяет учитывать зависимость прогнозируемых данных от внешних факторов.

$$\Delta^d y_t = c + \sum_{i=1}^{p} \Delta^d y_{t-1} + \sum_{j=0}^{q} b_j \varepsilon_{t-j} + \sum_{k=0}^{r} \beta_k x_t^k + \varepsilon_t, \tag{1.9}$$

где:

— r — количество внешних факторов, влияющих на прогнозируемую величину;

— β_k для $k = \overline{1, r}$ — коэффициенты внешних факторов, определяющие влияние предыдущих ошибок на текущее значение;

— x_t^k для $k = \overline{1, r}$ — значение k-го коэффициента в момент времени t.

Из результатов исследований [25] и [26] можно сделать вывод, что при переходе с метод ARIMA на метод ARIMAX общее качество прогнозов может как ухудшиться, так и улучшиться.

Таким образом, выбор между ARIMA и ARIMAX во многом зависит от предметной области, формализации данных, количества и качества формализации внешних факторов. Также важен правильный подбор коэффициентов, участвующих в вычислениях, так как от них во многом зависит итоговые характеристики полученной модели.

1.8.3 Двухэтапный подход к объединению

Помимо перечисленных методов в статье [27] был предложен двухэтапный подход, предполагающий оценку точек временного ряда при помощи двух различных статистических моделей.

Для реализации алгоритма предлагается сначала использовать сравнительно простую модель, способную выделить аномалии сравнительно быстро и с меньшей точностью. Затем точки, помеченные простой моделью как потенциально аномальные, анализируются при помощи второй модели — позволяющей оценить аномалии с большей точностью, но затрачивая на каждую аномалию больше ресурсов, чем первая модель.

Так как вторая модель обрабатывает только точки, отмеченные первой моделью как потенциально аномальные, максимальное количество реальных аномалий, выявленное комбинированным методов, ограниченно количеством точек, отмеченных первой моделью. Таким образом основная задача более простой модели — выявление максимального числа истинно положительных аномалий. Перед второй моделью ставится задача из точек, отмеченных как потенциально аномальные, отсеять наибольшее число ложно положительных результатов.

Лучшим подходом к подбору используемых моделей является применение моделей одного семейства. К примеру, формулы, используемые в ARIMA и ARIMAX моделях, для прогнозирования значения временного ряда имеют общие члены. Таким образом, при применении ARIMA в качестве более простой модели, а ARIMAX — в качестве более точной, удается сократить вычисления, необходимые для прогноза с помощью ARIMAX за счет использования значений членов уравнения, уже подсчитанных для данного шага в ходе прогноза моделью ARIMA.

В отличие от других популярных подходов к объединению алгоритмов (голосование [28], взвешенное голосование [29]) двухступенчатый подход позволяет обрабатывать больше одного раза только точки, признанные на первом этапе аномальными, что сокращает общее количество вычислений.

В связи с названными преимуществами для дальнейшей реализации комбинированного метода выбран двухэтапный подход к объединению методов выявления аномалий.

1.9 Сравнительный анализ

В ходе исследования данных, подлежащих анализу, было выявлено присутствие частых выбросов и наличие как минимум двух значимых сезонностей.

В связи с этим для сравнения методов выявления аномалий были выделены следующие критерии:

1. устойчивость к выбросам — способность метода качественно выявлять аномалии в наборе данных, имеющих резкие выбросы;

2. сезонность — возможность метода правильно оценивать аномалии в данных с явной сезонностью.

3. двойная сезонность — возможность метода правильно оценивать аномалии в данных с двумя сезонностями сезонностью.

Сравнительный анализ методов по данным критериям приведен в таблице 2.

Таблица 2 – Сравнительная таблица методов выявления аномалий временных рядов

Метод	**Устойчивость к выбросам**	**Сезонность**	**Двойная сезонность**
Z-скорректированные отклонения	-	-	-
Кумулятивные суммы	-	-	-
Брауна	-	-	-
Хольта-Винтерса	+	+	**-**
Хольта-Винтерса с двумя сезонностям	+	+	+
AR	-	+	-
ARMA	+	+	-
ARIMA	+	-	-
ARIMAX	+	-	-

Как видно из таблицы, наилучшими качествами по данным параметрам обладают методы Хольта-Винтерса и Хольта-Винтерса с двумя сезонностями, поэтому они выбраны для дальнейшей разработки комбинированного метода. Кроме того, данные методы принадлежат одному семейству, что позволяет сократить количество вычислений при реализации выбранного подхода к объединению алгоритмов.

1.10 Вывод

В этом разделе был проведён анализ предметной области. Была проведена постановка и формализация задачи. Также были рассмотрены существующие подходы к выявлению аномалий, были рассмотрены сами алгоритмы выявления аномалий. Для дальнейшей реализации был выбраны методы Хольта-Винтерса и Хольта-Винтерса с двумя сезонностями, объединяемые с использованием двухэтапного подхода.

2 Конструкторский раздел

Для поиска аномалий с использованием комбинации алгоритмов, выбранной в аналитическом разделе, необходимо получить массив потенциально аномальных точек при помощи модели Хольта-Винтерса, затем отсеять ложно-положительные аномалии с использованием модели Хольта-Винтерса с двумя сезонностями. Данный метод выявления аномалий можно представить на диаграмме IDEF0 в виде функциональных блоков (рисунок 5).

Рисунок 5 – Схема метода выявления аномалий

В данном разделе будут подробно описаны основные этапы работы метода выявления аномалий и алгоритмы, используемые при его реализации.

2.1 Алгоритм поиска аномалий

После получения анализируемого набора данных, необходимо их обработать выбранными алгоритмами. Обработка происходит в два этапа — сначала более быстрой, но менее точной моделью Хольта-Винтерса. Точки, отмеченные в ходе анализа данным алгоритмом как аномальные затем анализируются более точной, но более требовательной к ресурсам моделью Хольта-Винтерса с двумя сезонностями.

Схема комбинированного алгоритма поиска аномалий представлена на рисунке 6.

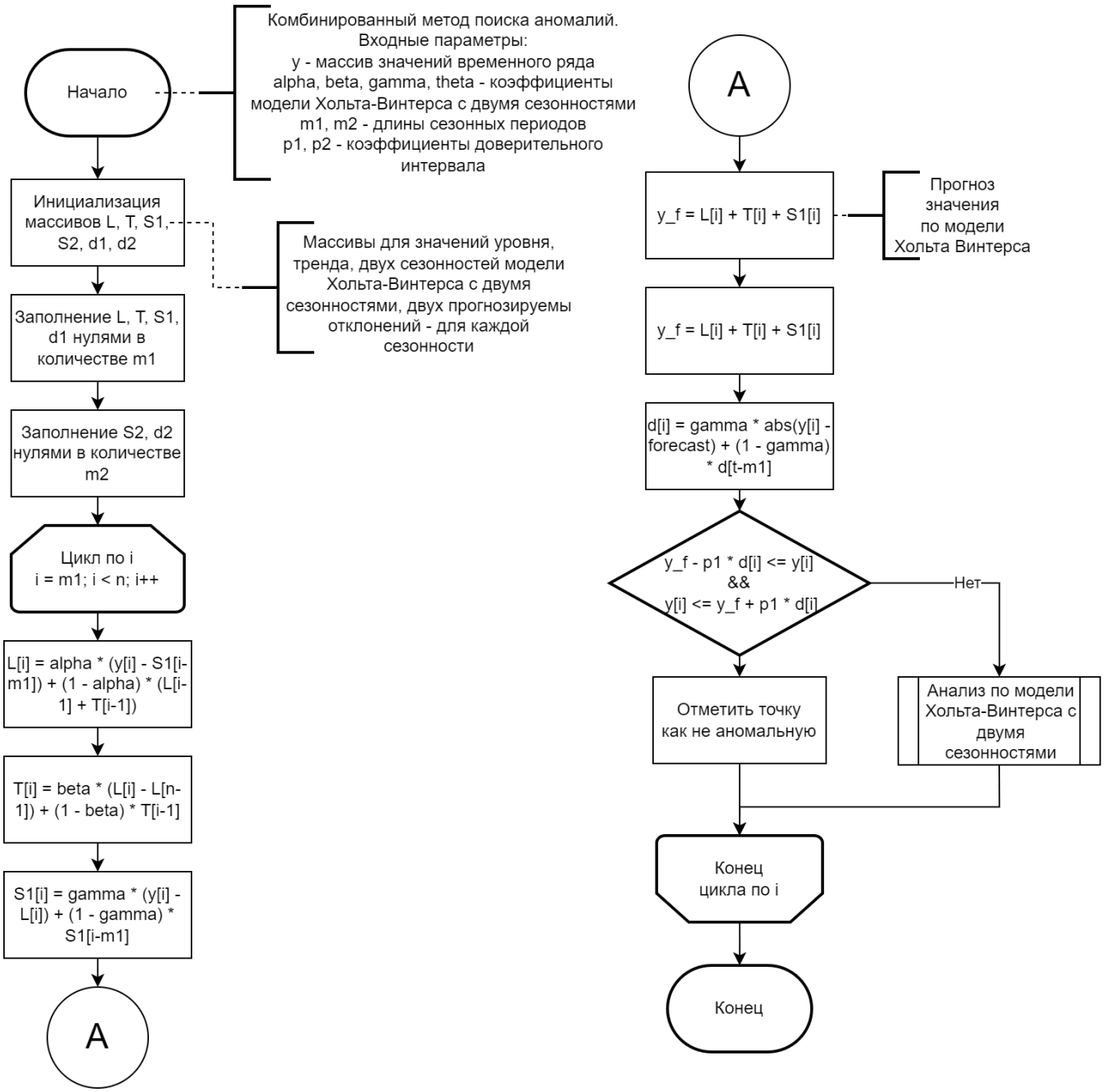

Рисунок 6 – Схема комбинированного алгоритма поиска аномалий

В случае, если точка отмечена как аномальная, ее необходимо дополнительно проверить на истинно положительную аномалию. Для этого значение прогноза, полученное по модели Хольта-Винтерса, корректируется по модели Хольта-Винтерса с двумя сезонностями. Данная модель частично использует значения, уже посчитанные на первом этапе алгоритма, что сокращает используемое процессорное время.

Схема второго этапа алгоритма — проверки по модели Хольта-Винтерса с двумя сезонностями, представлена на рисунке 7.

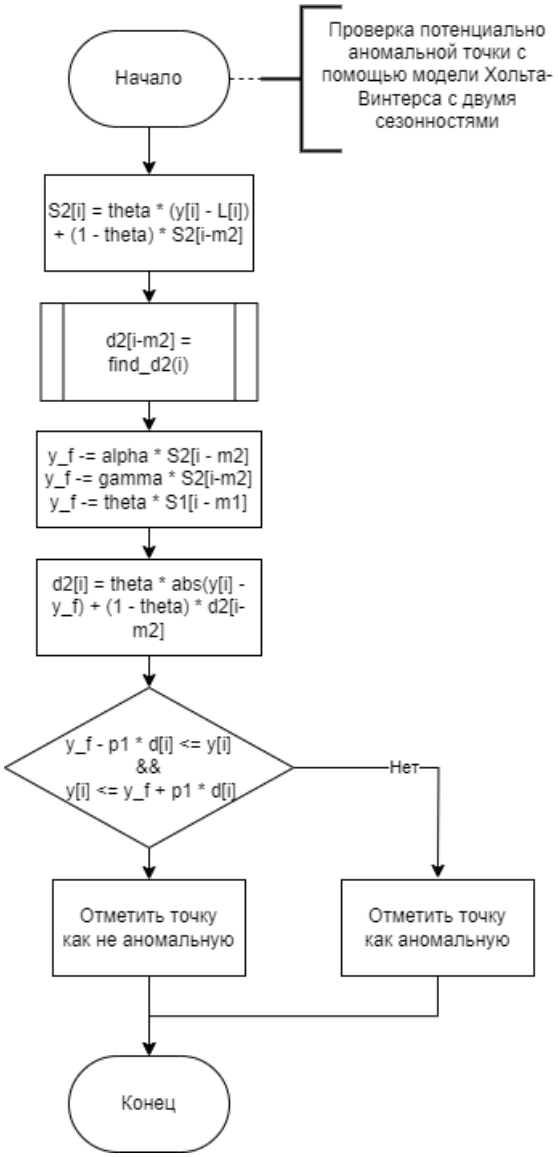

Рисунок 7 – Схема второго этапа алгоритма — проверки по модели Хольта-Винтерса с двумя сезонностями

Так как значения прогнозируемых отклонений по второй сезонности d2 вычисляются не для каждой точки, а только для аномальных, может сложиться ситуация, в которой значение $d2[i - m2]$ не было посчитано. Для разрешения этой проблемы используется рекурсивная функция вычисления $d2[i - m2]$. Ее схема представлена на рисунке 8.

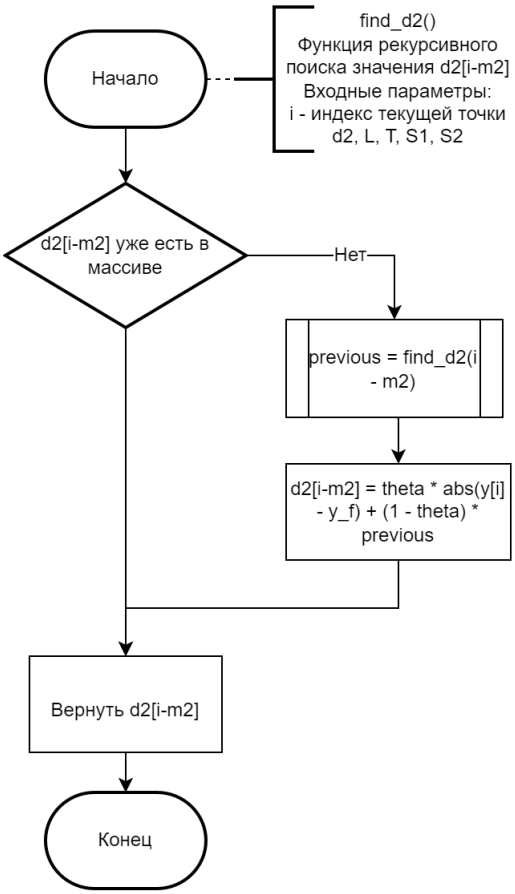

Рисунок 8 – Схема функции рекурсивного вычисления значения прогнозируемого отклонения по второй сезонности d2

2.2 Вывод

В данном разделе была представлена диаграмма IDEF0 в виде функциональных блоков разрабатываемого метода выявления аномалий в массивах данных о посещениях организации сотрудниками и посетителями, определены его основные этапы. Также были подробно описаны комбинированный алгоритм выявления аномалий и используемые им функции — проверки потенциально аномальной точки при помощи метода Хольта-Винтерса с двумя сезонностями и рекурсивного вычисления прогнозируемого отклонения для данного метода.

3 Технологический раздел

Данный раздел содержит обоснование выбора средств программной реализации разработанного метода и вспомогательных библиотек, а также описание структуры разработанного ПО. Целью этого раздела является разработка программного обеспечения, реализующего описанный метод выявления аномалий в массивах данных о посещении организации сотрудниками и посетителями.

3.1 Выбор языка программирования

Для реализации ПО был выбран объектно-ориентированный подход, так как он позволяет легко модифицировать программу и дает возможности для дальнейшего расширения функционала программы.

В качестве язык программирования был выбран объектно-ориентированный язык C# [30]. Данный язык программирования используется для разработки приложений на платформе .NET Framework [31]. Сам язык C#. изначально разрабатывался для платформы .NET Framework и работы с CLR под влиянием таких языков, как C++ и Java.

Платформа .NET Framework — выпущена компанией Microsoft в 2002 году. Основой платформы является общеязыковая среда исполнения CLR (Common Language Runtime), которая подходит для разных языков программирования. Функциональные возможности CLR доступны в любых языках программирования, использующих эту среду.

В выбранном языке есть все требующиеся инструменты для реализации программы, а именно библиотеки для работы с файлами, создания пользовательского интерфейса и визуализации данных.

3.2 Выбор вспомогательных библиотек

Также следует упомянуть сторонние библиотеки, которые использовались при реализации программы:

— CSVHelper [32] — библиотека для работы с CSV файлами для платформы .NET, позволяет создавать и читать файлы в формате CSV;

— OxyPlot [33] — кроссплатформенная библиотека .NET для построения графиков и визуализации;

— WPF [34] — это технология, разработанная Microsoft для создания гра-

фических пользовательских интерфейсов в приложениях Windows.

3.3 Структура разработанного ПО

На рисунке 9 изображена диаграмма компонентов разработанного ПО.

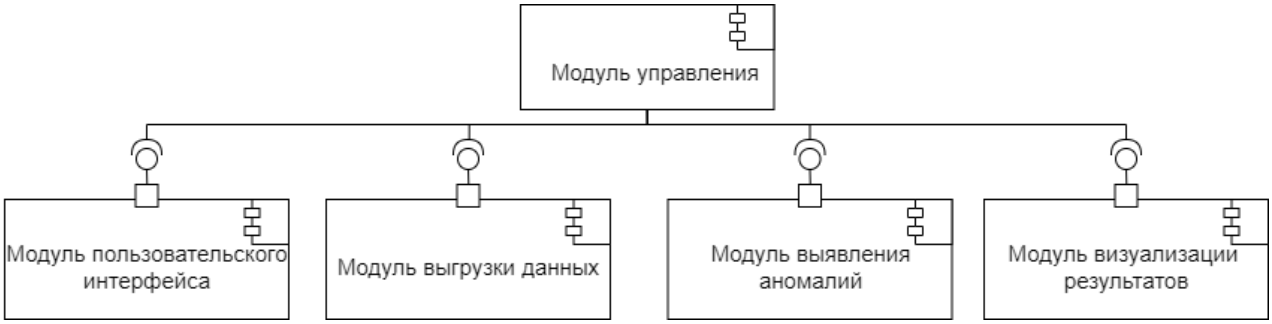

Рисунок 9 – Диаграмма компонентов ПО

Каждый модуль, показанный на диаграмме, представляет собой группу классов, объединенных по их функциональному назначению. Модуль управления отвечает за управление работой модулей выгрузки, интерфейса, выявления аномалий и визуализации результатов. Модуль выгрузки данных отвечает за загрузку исходных данных и их преобразование к формату, используемом в методе выявления аномалий. Модуль визуализации результатов реализует графическое представление временного ряда исходных данных и выявленных аномалий.

3.4 Развертывание ПО

Технические характеристики и программное обеспечение машины, на которой производились разработка, отладка и демонстрация ПО:

— операционная система W indows 10 Домашняя Версия 21H1 x86_64 [35];

— оперативная память 8 Гбайт 2133 МГц;

— процессор Intel Core i5-8300H с тактовой частотой 2.30 ГГц, 4 физических ядра, 8 логических ядер [36].

3.5 Пример результатов работы программы

В процессе анализа результатов важное значение имеет визуализация работы алгоритма обнаружения аномалий. Так как рассматриваемый ряд одномерный, его визуализацию можно наглядно изобразить в виде графика.

Пример отображения результатов, полученных в процессе работы разработанного ПО, изображен на рисунке 10.

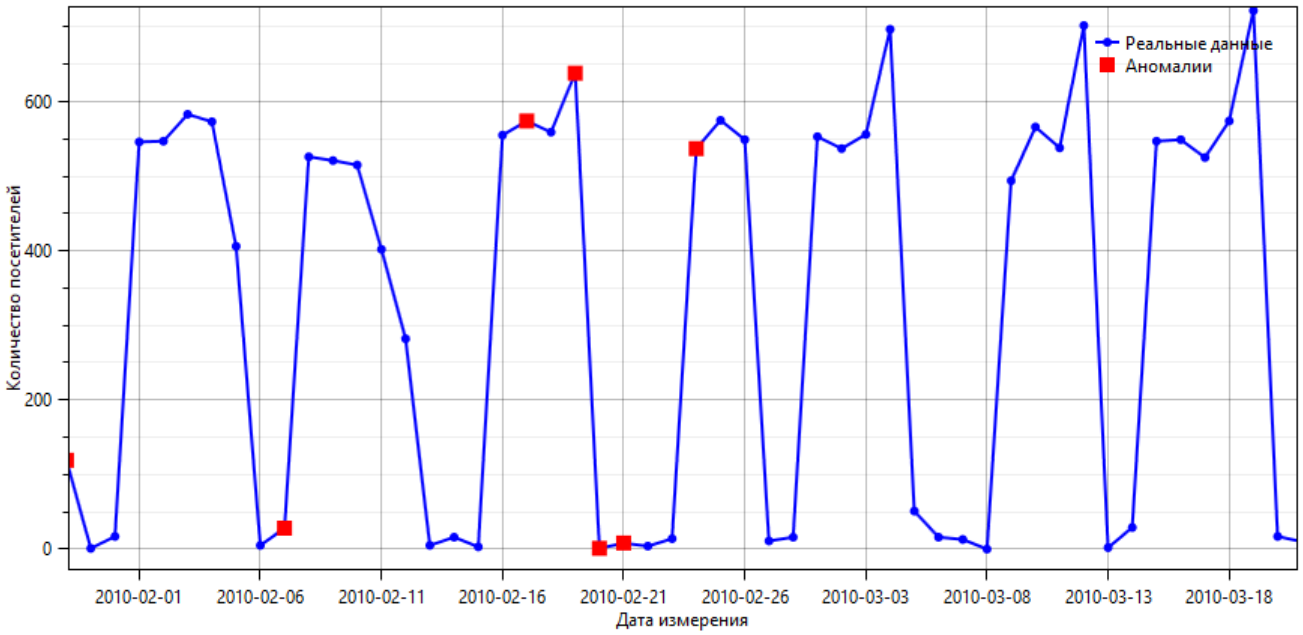

Рисунок 10 – Пример визуализации результатов. Красные точки обозначают значения, признанные аномальными

3.6 Вывод

В этом разделе в качестве языка программирования для реализации ПО был выбран язык C# с использованием платформы .NET. Для визуализации результатов применялась библиотека OxyPlot. Для реализации графического интерфейса пользователя была использована технология WPF, а для считывания данных в формате CSV — библиотека CSVHelper. Также в данном разделе представлены структура разработанной программы и описание результатов ее работы. Была создана программная реализация метода выявления аномалий.

4 Исследовательский раздел

В данном разделе анализируются результаты, полученные в ходе работы реализованного метода с реальными данными статистики посещений, а также проводится исследование зависимости эффективности работы разработанного алгоритма от различных значений параметров и наборов данных.

Для исследования использованы два набора данных из коллекции промаркированных временных рядов, а также набор данных, основанный на реальных данных статистики посещения организации.

4.1 Анализ результатов применения метода к реальным данным

Для исследования применимости разработанного метода к реальным данным, были использованы данные с контрольно-пропускного пункта бизнес-центра за период 15.01.2007–31.03.2019. Всего 4459 дня со значением количества посетителей в каждый из этих дней.

Исследование проводилось на следующих значениях параметров:

- $\alpha = 0,7$ — коэффициент компоненты уровня;
- $\beta = 0,03$ — коэффициент компоненты тренда;
- $\gamma = 0,99$ — коэффициент сезонной компоненты по m1;
- $\theta = 0.9$; — коэффициент сезонной компоненты по m2;
- $m1 = 7$ — длина более короткого сезонного цикла (еженедельный);
- $m2 = 365$ — длина более длинного сезонного цикла (ежегодный);
- $p1 = 1,01$ — коэффициент доверительного интервала более короткого сезонного цикла;
- $p2 = 1,05$ — коэффициент доверительного интервала более длинного сезонного цикла;

Результаты, полученные с использованием данных параметров показали наименьшее среднеквадратичное отклонение [37] от данных, что говорит о наиболее точном прогнозировании значений по сравнению с результатами при других исследованных значениях параметров.

Таким образом данные параметры обеспечивают наиболее качественное выявление аномалий среди исследованных значений.

На рисунке 11 приведен результат обработки реального массива данных разработанным алгоритмом.

Рисунок 11 – Результат обработки реального массива данных разработанным алгоритмом. Красными квадратами обозначены точки, признанные аномальными

Всего в ходе первого этапа (обработки менее точной моделью) было выявлено 17 потенциально аномальных точек. После обработки этих данных в ходе второго этапа (более точной моделью) 8 точек были признаны аномалиями, а 9 были отсеяны как ложно положительные.

4.2 Сравнение результатов работы с разными наборами данных

Разработанное ПО было происследовано на трех разных наборах данных. Два из них были взяты из Numenta Anomaly Benchmark (NAB) [38] — коллекции промаркированных временных рядов реальных данных. Третий набор данных — это искусственно созданный ряд на основе данных из пункта 4.1.

Так как разработанный алгоритм направлен на обработку данных с двумя сезонностями, из NAB также были выбраны два временных ряда с двумя сезонностями. Набор данных «Такси» содержит информацию о количестве пассажиров такси Нью-Йорка за период 01.07.2014–31.01.2015. Набор

данных «Температура» содержит информацию о температуре работы индустриальной машины за период 02.12.2013–19.02.2014. Набор данных «Посещения» — временной ряд на основе данных из пункта 4.1. После сглаживания и экстраполяции данного временного ряда в него вносились аномалии путем добавления случайной компоненты к значению в точках через случайные промежутки на оси времени.

Характеристики временных рядов представлены в таблице 3.

Таблица 3 – Характеристики наборов данных

Набор данных	Интервал	Сезон. №1	Сезон. №2	Аномалии	Всего
«Такси»	30 мин	Ежедневно	Еженедельно	5	10320
«Температура»	5 мин	Ежедневно	Еженедельно	5	22695
«Посещения»	1 день	Еженедельно	Ежегодно	9	4459

Для подбора значений был использован метод градиентного спуска [39] — метод нахождения минимального значения функции потерь [40]. Согласно данному методу параметры модели инициализируются случайными значениями из области определения, а затем вычисляется рекуррентно по следующей формуле:

$$\overrightarrow{\vartheta}_k = \overrightarrow{\vartheta}_{k-1} - \eta \cdot \nabla_\vartheta J(\overrightarrow{\vartheta}_{k-1}), \tag{4.1}$$

где:

- $\overrightarrow{\vartheta}_k$ — значение вектора параметров $\overrightarrow{\vartheta}$ на текущем шаге;
- $\overrightarrow{\vartheta}_{k-1}$ — значение вектора параметров $\overrightarrow{\vartheta}$ на предыдущем шаге;
- η — шаг обучения;
- $J(\overrightarrow{\vartheta}_{k-1})$ — функция потерь. В ходе исследования в качестве функции потерь использовалось среднеквадратичное отклонение [37];
- $\nabla_\vartheta J(\overrightarrow{\vartheta}_{k-1})$ — градиент [41] функции потерь.

Преобразования по формуле (4.1) выполняются до тех пор, пока модель не достигнет приемлемого уровня точности.

Для оценки качества использовались метрики, применяемые в бинарной классификации данных, такие как точность и полнота. Это связано с тем,

что при выявлении аномалий данные разделяются на два класса: аномальные и нормальные [42]. При вычислении этих метрик обработанные данные классифицируются на четыре категории:

— True Positive (TP) — правильно идентифицированные аномалии;

— False Positive (FP) — ложные срабатывания;

— True Negative (TN) — правильно классифицированные нормальные данные;

— False Negative (FN) — аномалии, не распознанные как таковые. Метрики рассчитываются по формулам (4.2) и (4.3).

$$Prec = \frac{TP}{TP + FP}, \tag{4.2}$$

где

— $Prec$ — значение метрики точности,

— TP — количество элементов в выборке, отмеченных как True Positive,

— FP — количество элементов в выборке, отмеченных как False Positive.

$$Rec = \frac{TP}{TP + FN}, \tag{4.3}$$

где

— Rec — значение метрики полноты,

— TP — количество элементов в выборке, отмеченных как True Positive,

— FN — количество элементов в выборке, отмеченных как False Negative.

Результаты исследования приведены в таблице 4.

В таблицу 4 занесены только те значения параметров, на которых разработанное ПО показало лучший результат для одного из наборов данных. Таким образом при параметрах со строки 1 был получен лучший результат для временного ряда «Посещения», со строки 2 — для ряда «Такси», со строки 3 — для ряда «Температура».

Таблица 4 – Результаты исследования

					Посещения		**Такси**		**Температура**	
№	α	β	γ	θ	**Prec**	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**
1	**0,70**	**0,03**	**0,99**	**0,90**	0,75	1	0,20	1	0,33	0,60
2	**0,80**	**0,10**	**0,70**	**0,80**	0,69	0,81	0,80	1	0,63	0,18
3	**0,4**	**0,07**	**0,5**	**0,83**	0,625	0,83	0,63	0,56	1	0,71

Из таблицы видно, что значения параметров необходимо подбирать индивидуально — для каждого ряда наиболее точный результат обеспечивается разным набором значений параметров, определяющимся спецификой анализируемых данных. Определить однозначно лучшие и наиболее универсальные значения параметров не представляется возможным.

Использование для обработки ряда значений параметров, подобранных для другого ряда, в общем случае не приведет к удовлетворительному результату выявления аномалий.

4.3 Вывод

Данные исследования показали, что разработанный метод позволяет выделять аномальные значения во временном ряде, содержащем данные о количестве посещений организации в день, а также в других наборах данных. Тем не менее для получения наиболее точного результата выявления аномалий разработанное ПО требует индивидуального подбора значений параметров, обусловленных спецификой анализируемых данных. Определить однозначно лучшие и наиболее универсальные значения параметров не представляется возможным. Возможным решением проблемы подбора значений параметров могут быть применение градиентного спуска или машинного обучения.

ЗАКЛЮЧЕНИЕ

В рамках данной квалификационной работы был разработан и программно реализован комбинированный метод выявления аномалий на основе статистических методов. Кроме того были выполнены все поставленные задачи.

— Были проанализированы методы выявления аномалий в статистике посещений организации и проведен их сравнительный анализ. Для реализации были выбраны методы

— Были спроектированы необходимые для работы метода алгоритмы.

— Был разработан метод выявления аномалий в статистике посещений.

— Была создана программная реализация разработанного метода.

— Была исследована зависимость эффективности работы реализованного метода от значений параметров системы. Исследование показало, что разработанный метод применим к реальным наборам данных, однако требует индивидуального подбора параметров системы для каждого отдельного временного ряда.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Кильдишев Г. С., Френкель А. А. Анализ временных рядов и прогнозирование — 2-е издание. — М. : Ленанд, 2021. — 104 с.

2. Козлов А. Е. Система контроля и управления доступом на предприятие: понятие, характеристика и основные требования // Вестник ВГТУ. — 2019. — №1.

3. Антипов С. Г., Фомина М. В. Проблема обнаружения аномалий в наборах временных рядов — Программные продукты и системы — 2012 — с. 78–82.

4. Шкодырев В.П. , Ягафаров К. И., Баштовенко В. А. Обзор методов обнаружения аномалий в потоках данных. // Ильина Е. Э. // Second Conference on Software Engineering and Information Management. — 2017.

5. Microsoft Azure AI Anomaly Detector. [Электронный ресурс]. — Режим доступа — URL: https://azure.microsoft.com/ru-ru/products/ai-services/ai-anomaly-detector — (дата обращения: 02.05.2024).

6. Anodot. [Электронный ресурс]. — Режим доступа — URL: https://www.anodot.com/about/ — (дата обращения: 02.05.2024).

7. Atlas Docs. [Электронный ресурс]. — Режим доступа — URL: https://netflix.github.io/atlas-docs/ — (дата обращения: 02.05.2024).

8. Morgoth. [Электронный ресурс]. — Режим доступа — URL: https://github.com/nathanielc/morgoth — (дата обращения: 02.05.2024).

9. LinkedIn's ThirdEye. [Электронный ресурс]. — Режим доступа — URL: https://github.com/project-thirdeye/thirdeye — (дата обращения: 02.05.2024).

10. Kappal S. Data normalization using median absolute deviation MMAD based Z-score for robust predictions vs. min–max normalization — London Journal of Research in Science: Natural and Formal — 2019 — Т. 19 — 10 с.

11. Z-Scores: Understanding Standardization in Statistics. [Электронный ресурс]. — Режим доступа — URL:

https://www.alooba.com/skills/concepts/statistics/z-scores/ — (дата обращения: 02.05.2024).

12. Lazarevic A., Kumar V. Feature bagging for outlier detection — InProceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining — 2005 — С. 157–166.

13. Omar S., Ngadi A., Jebur H. H. Machine learning techniques for anomaly detection: an overview — International Journal of Computer Applications — 2013 — № 79.

14. Трегуб А. В., Трегуб И. В. Методика построения модели ARIMA для прогнозирования динамики временных рядов. Лесной вестник/Forestry bulletin — 2011 — №5 — С. 179–183.

15. Косовцева Т. Р., Беляев В. В. Технологии обработки экономической информации. Адаптивные методы прогнозирования. Учебное пособие. — СПб: Университет ИТМО, 2016, — 31 с.

16. Поздняков А. С. Применение метода Хольта-Винтерса при анализе и прогнозировании динамики временных рядов — Проблемы организации и управления на транспорте 2017 — с. 57–64.

17. Single Seasonal Time Series Anomaly Detection with Brutlag's Algorithm and Holt-Winter ETS. [Электронный ресурс]. — Режим доступа — URL: https://medium.com/@tle3006/single-seasonal-time-series-anomaly-detection-with-brutlags-algorithm-and-holt-winter-ets-d8aea1fd1bfc — (дата обращения: 02.05.2024).

18. Szmit M., Szmit A. Usage of Modified Holt-Winters Method in the Anomaly Detection of Network Traffic: Case Studies. // Journal of Computer Networks and Communications. — 2012. — №34.

19. Васильев К. К., Дементьев В. Е. Авторегрессионные модели многомерных изображений — Наукоемкие технологии. — 2013. — Т. 14, №5. — с. 12–15.

20. Least Squares Method. [Электронный ресурс]. — Режим доступа — URL: https://www.investopedia.com/terms/l/least-squares-method.asp — (дата обращения: 02.05.2024).

21. Maximum Likelihood Estimation. [Электронный ресурс]. — Режим доступа — URL: https://online.stat.psu.edu/stat415/lesson/1/1.2 — (дата обращения: 02.05.2024).

22. Makridakis S., Hibon M. ARMA models and the Box–Jenkins methodology — Journal of forecasting — 1997 — №16 — 63 с.

23. Su Y., Cui C., Qu H. Self-Attentive Moving Average for Time Series Prediction. // Applied Sciences. — 2022.

24. Fan J., Shan R., Cao X. The analysis to tertiary-industry with ARIMAX model — Journal of Mathematics Research — 2009 — 1(2) — 156 с.

25. Peter D., Silvia P. ARIMA vs. ARIMAX — which approach is better to analyze and forecast macroeconomic time series. — InProceedings of 30th international conference mathematical methods in economics — 2012 — Т. 2 — с. 136–140.

26. Anggraeni W., Vinarti R. A, Kurniawati Y. D. Performance comparisons between arima and arimax method in moslem kids clothes demand forecasting: Case study. // Procedia Computer Science — 2015 — № 72 — 7 с.

27. Sperl R., Chung S. Two-Step Anomaly Detection for Time Series Data. // 2019 International Conference on Data and Software Engineering. — 2019.

28. Saha S., Ekbal A. Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. // Data & Knowledge Engineering. — V. 85. — 2013. — C. 15-39.

29. Rosas R., Garreta J. Combining Multiple Classifiers with Dynamic Weighted Voting. // Hybrid Artificial Intelligence Systems. — 2009.

30. C# language documentation. [Электронный ресурс]. — Режим доступа — URL: https://learn.microsoft.com/en-us/dotnet/csharp/ — (дата обращения: 20.05.2024).

31. .NET Framework. [Электронный ресурс]. — Режим доступа — URL: https://dotnet.microsoft.com/en-us/ — (дата обращения: 20.05.2024).

32. CsvHelper. [Электронный ресурс]. — Режим доступа — URL: https://joshclose.github.io/CsvHelper/ — (дата обращения: 20.05.2024).

33. OxyPlot [Электронный ресурс]. — Режим доступа — URL: https://oxyplot.github.io — (дата обращения: 20.05.2024).

34. Windows Presentation Foundation [Электронный ресурс]. — Режим доступа — URL: https://learn.microsoft.com/ru-ru/dotnet/desktop/wpf/overview/ — (дата обращения: 20.05.2024).

35. Windows [Электронный ресурс]. — Режим доступа — URL: https://www.microsoft.com/en-us/windows — (дата обращения: 02.05.2024).

36. Процессор Intel Core i5 [Электронный ресурс]. Режим доступа — URL: https://www.intel.com/processors/core/i5/docs — (дата обращения: 02.05.2024).

37. Трофимова Е. А., Кисляк Н. В., Гилев Д. В. Теория вероятностей и математическая статистика. — Екб.: Издательство Уральского университета, 2018. — 164 с.

38. Numenta Anomaly Benchmark. [Электронный ресурс]. — Режим доступа — URL: https://www.numenta.com/resources/htm/numenta-anomoly-benchmark/ — (дата обращения: 20.05.2024).

39. Ruders S. An overview of gradient descent optimization algorithms. [Электронный ресурс]. — Режим доступа — URL: https://arxiv.org/abs/1609.04747 — (дата обращения: 20.05.2024).

40. Loss Function. [Электронный ресурс]. — Режим доступа — URL: https://deepai.org/machine-learning-glossary-and-terms/loss-function — (дата обращения: 20.05.2024).

41. Gradient Definition [Электронный ресурс]. — Режим доступа — URL: https://www.cuemath.com/geometry/gradient-definition/ — (дата обращения: 20.05.2024).

42. Функционалы качества бинарной классификации. [Электронный ресурс]. — Режим доступа — URL: https://dyakonov.org/2019/05/31/функционалы-качества-в-задаче-бинарн/ — (дата обращения: 20.05.2024).

ПРИЛОЖЕНИЕ А