**Министерство науки и высшего образования Российской Федерации**

**ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина»**

Кафедра «школа профессионального и академического образования (школа)»

Оценка работы\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Руководитель от УрФУ В. И. Борисов

Тема задания на практику

Применение моделей компьютерного зрения для анализа туристической активности по геопривязанным изображениям из социальной сети «ВКонтакте»

ОТЧЕТ

Вид практики Производственная практика

Тип практики Производственная практика, научно-исследовательская работа

Научный руководитель

**ФИО руководителя Подпись**

Студент А. А. Петросян

**ФИО студента**

Специальность (направление подготовки) 09.04.01 Информатика и вычислительная техника

Группа РИМ-231902

Екатеринбург 2025

**Министерство науки и высшего образования Российской Федерации**

**ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина»**

|  |  |
| --- | --- |
|  | СОГЛАСОВАНО  УрФУ  « » \_\_\_\_\_\_\_ 2025 г. |
|  | Руководитель образовательной программы  \_\_\_\_\_\_\_\_\_\_ В.И. Борисов  Подпись расшифровка подписи |

Институт ИРИТ-РТФ Группа РИМ-231902 Кафедра школа профессионального и академического образования (школа)

Код, наименование направления 09.04.01-Информатика и вычислительная техника

Наименование образовательной программы/Наименование магистерской программы 09.04.01/33.03 - Инженерия машинного обучения

**ИНДИВИДУАЛЬНОЕ ЗАДАНИЕ**

На учебную, производственную (преддипломную) практику студента

Петросян Александр Артурович

(фамилия, имя, отчество)

1. Тема задания на практику анализ пространственно-временной динамики социальных взаимодействий на основе геопривязанных фотографий из социальной сети ВКонтакте

2. Срок практики: с 24.03.2025 г. по 18.05.2025 г.

Срок сдачи студентом отчета 19.05.2025–25.05.2025 г.

3. Место прохождения практики Школа профессионального и академического образования, ИРИТ-РТФ

4. Вид практики Производственная практика

5. Тип практики Производственная практика, научно-исследовательская работа

6. Содержание отчета 1) Введение; 2) Основная часть ВКР (разделы 1–3 и др.); 3) Заключение; 4) Список использованных источников; 5) Приложение.

**Рабочий график (план) проведения практики**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Этапы практики*** | ***Наименование работ студента*** | ***Срок*** | ***Примечание*** |
| Организационный | Ознакомление с требованиями для выполнения задания. Обсуждение с научным руководителем практики темы на практическую подготовку в организации, уточнение ее формулировки, составление конкретного последовательного перечня работ, необходимых для выполнения задания. | с 24.03.2025–30.03.2025 г. |  |
| Основной | Выполнение индивидуального задания по следующему перечню работ:  1) Введение;  2) Основная часть ВКР (разделы 1–3 и др.);  3) Заключение;  4) Список использованных источников;  5) Приложение (опционально) | с 31.03.2025–11.05.2025 г. |  |
| Заключительный | Составление структуры и написание полного отчета о выполнении индивидуального задания, в соответствии с требованиями УрФУ. Согласование отчета с научным руководителем. | с 12.05.2025–  18.05.2025 г. |  |

Содержание практики и планируемые результаты практики согласованы с руководителем практики от профильной организации

Руководитель от УрФУ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Подпись расшифровка подписи

Задание принял к исполнению(студент)\_\_\_\_\_\_\_\_\_\_\_\_\_

Подпись расшифровка подписи

ВВЕДЕНИЕ

Геолокационные данные в социальных сетях представляют собой ценный источник информации для анализа социальных взаимодействий и пространственно-временных паттернов поведения людей. В последние годы анализ геопривязанных данных из социальных сетей стал важным инструментом для понимания городской динамики, туристических потоков, социальных событий и других аспектов человеческой активности [1,2].

Социальная сеть ВКонтакте, являющаяся одной из крупнейших в России и странах СНГ, содержит огромное количество фотографий с геопривязкой, которые отражают реальные социальные взаимодействия и перемещения пользователей. Эти данные могут быть использованы для выявления закономерностей в пространственно-временной активности людей, популярных мест, сезонных изменений в посещаемости различных локаций и других важных социальных феноменов [3].

Особую ценность представляет возможность анализа данных, собранных на протяжении длительного периода времени. Сервис сбора фотографий, разработанный в 2020-2021 гг., к настоящему моменту накопил базу данных объемом около 400 миллионов объектов, что позволяет проводить масштабные исследования с высокой статистической значимостью. Такие большие объемы данных требуют применения современных методов обработки и анализа, включая технологии Big Data и алгоритмы машинного обучения [4,5].

В отличие от традиционных методов исследования социальной активности, таких как опросы и наблюдения, анализ данных из социальных сетей позволяет получить более объективную и репрезентативную картину, не подверженную субъективным искажениям. Исследования Huang et al. [6] и Lansley et al. [7] показывают, что геопривязанные данные из социальных сетей могут служить надежным индикатором реальных социальных процессов при правильном методологическом подходе и учете потенциальных смещений выборки.

Как отмечают Li et al. [8], геолокационные данные из социальных сетей могут быть использованы для решения таких практических задач, как:

- оптимизация городской инфраструктуры;

- планирование транспортных маршрутов;

- выявление туристических аттракторов;

- анализ событийной активности;

- оценка влияния природных и социальных явлений на перемещения людей;

- маркетинговые исследования и территориальное планирование бизнеса.

Данная работа направлена на разработку методологии анализа пространственно-временной динамики социальных взаимодействий на основе геопривязанных фотографий из социальной сети ВКонтакте и выявление закономерностей в социальной активности пользователей.

Цель ВКР – разработать методику анализа пространственно-временной динамики социальных взаимодействий на основе геопривязанных фотографий из социальной сети ВКонтакте и оценить применимость различных методов машинного обучения для выявления закономерностей и паттернов в собранных данных.

Для достижения этой цели были поставлены следующие задачи:

1. Провести анализ существующих методов и подходов к исследованию геопространственных данных из социальных сетей.

2. Разработать методологию предварительной обработки и фильтрации данных для повышения их качества и репрезентативности.

3. Определить оптимальные методы пространственного анализа для выявления кластеров активности и популярных локаций.

4. Разработать подход к временному анализу данных для выявления паттернов и трендов в активности пользователей.

5. Исследовать возможности применения методов социального анализа для выявления связей между пользователями и характера их взаимодействия.

6. Интегрировать результаты различных типов анализа для получения комплексного представления о пространственно-временной динамике социальных взаимодействий.

Результатом работы будет методология анализа геопространственных данных из социальных сетей, которая позволит исследователям и практикам в области социологии, урбанистики и маркетинга получать ценные инсайты о социальной активности и пространственном поведении людей.

1 Анализ поставленной задачи и выбор методологии

Задача анализа пространственно-временной динамики социальных взаимодействий на основе геопривязанных фотографий может быть рассмотрена с нескольких ключевых перспектив. Во-первых, это пространственный анализ, направленный на выявление кластеров активности, популярных локаций и пространственных паттернов. Во-вторых, это временной анализ, позволяющий отследить сезонные изменения, суточные ритмы и динамику популярности различных мест со временем. В-третьих, это социальный анализ, направленный на выявление связей между пользователями, социальных групп и характера взаимодействия в различных локациях.

1.1 Обзор существующих подходов к анализу геопространственных данных

Существует несколько ключевых подходов к анализу геопространственных данных из социальных сетей, которые были исследованы в работах различных авторов. Steiger et al. [9] предлагают комплексный фреймворк для анализа геопривязанных данных из Twitter, который включает в себя этапы сбора, фильтрации, агрегации, анализа и визуализации данных. Авторы отмечают важность учета пространственных, временных и тематических аспектов при анализе таких данных.

В работе Hasan et al. [10] представлен подход к анализу пространственно-временных паттернов городской мобильности на основе данных из Foursquare и Twitter. Авторы используют методы кластеризации и классификации для выявления типичных траекторий перемещения пользователей и факторов, влияющих на выбор маршрутов.

Исследование Chen et al. [11] фокусируется на применении методов глубокого обучения для анализа геопривязанных фотографий из Flickr и Instagram. Авторы предлагают архитектуру нейронной сети, которая позволяет одновременно учитывать пространственные, временные и визуальные характеристики фотографий для предсказания популярности различных локаций.

1.2 Выбор методов для проведения исследования

Для проведения данного исследования необходимо использовать комбинацию методов из области геоинформатики, анализа временных рядов и машинного обучения. Пространственный анализ может быть проведен с использованием методов кластеризации, таких как DBSCAN или HDBSCAN, которые хорошо подходят для выявления кластеров произвольной формы в геопространственных данных [12,13]. Временной анализ может быть проведен с использованием методов декомпозиции временных рядов, спектрального анализа и прогнозирования [14]. Социальный анализ может включать в себя построение графов взаимодействия и применение методов анализа социальных сетей [15].

Сравнительный анализ различных методов кластеризации для геопространственных данных, проведенный Birant и Kut [16], показывает, что DBSCAN и его модификации обеспечивают наилучшие результаты для данных с неравномерным распределением плотности, что характерно для геопривязанных фотографий из социальных сетей. Метод HDBSCAN, предложенный Campello et al. [17], является усовершенствованным вариантом DBSCAN, который способен автоматически определять оптимальное количество кластеров и работать с данными различной плотности.

Для анализа временных рядов могут быть использованы такие методы, как сезонная декомпозиция STL (Seasonal-Trend decomposition using LOESS), предложенная Cleveland et al. [18], или более современные подходы, такие как Prophet, разработанный Facebook [19], которые позволяют выявлять тренды, сезонность и аномалии во временных данных.

В отличие от традиционных исследований геолокационных данных, которые часто фокусируются только на пространственном аспекте, данная работа предполагает интегрированный подход, учитывающий все три измерения: пространство, время и социальные взаимодействия. Это позволит получить более полную и многогранную картину социальной активности, как это было продемонстрировано в работах Hawelka et al. [20] и Silva et al. [21].

1.3 Структура и особенности исходных данных

Собранная база данных представляет собой структурированный набор информации о фотографиях с геопривязкой, включающий такие атрибуты, как:

- имя файла: уникальный идентификатор фотографии

- координаты (долгота и широта): геопространственное положение фотографии с точностью до нескольких метров

- дата и время: временная метка создания фотографии

- ссылка на изображение: URL для доступа к полноразмерному изображению

- ссылка на миниатюру: URL для доступа к миниатюре изображения

- оценка: числовой показатель популярности фотографии

- дополнительные метаданные: теги, описания, информация о пользователе (с учетом ограничений приватности)

Эта структура данных позволяет проводить разнообразные типы анализа, от простого картографирования до сложных методов машинного обучения. По своим характеристикам данные соответствуют понятию "больших данных" (Big Data), так как обладают значительным объемом (Volume), разнообразием (Variety) и высокой скоростью обновления (Velocity) [22].

1.4 Методы машинного обучения для анализа данных

Ключевой подход в данной работе будет основан на методах машинного обучения, которые позволяют выявлять неочевидные закономерности в больших объемах данных. Для анализа пространственных данных будут использованы методы кластеризации и классификации, для временного анализа – модели прогнозирования и выявления аномалий, а для социального анализа – методы анализа графов и сетей.

Yang et al. [23] демонстрируют эффективность применения алгоритмов глубокого обучения, таких как сверточные нейронные сети (CNN) и рекуррентные нейронные сети (RNN), для анализа геопространственных данных. Авторы показывают, что такие модели способны учитывать как пространственные, так и временные зависимости в данных и обеспечивают более высокую точность предсказаний по сравнению с традиционными методами.

Для выявления аномалий и необычных паттернов в данных могут быть использованы такие методы, как изолирующий лес (Isolation Forest) [24] или одноклассовый SVM (One-Class SVM) [25], которые способны идентифицировать наблюдения, значительно отличающиеся от общей тенденции.

2 Методология анализа геопространственных данных из социальных сетей

Методология анализа геопространственных данных из социальных сетей включает в себя несколько ключевых этапов: сбор и предварительная обработка данных, пространственный анализ, временной анализ, социальный анализ и интерпретация результатов.

2.1 Сбор и предварительная обработка данных

На этапе предварительной обработки данных производится фильтрация ошибочных или неполных записей, нормализация координат, преобразование временных меток в удобный для анализа формат и обогащение данных дополнительной информацией, такой как тип местности, административная принадлежность и т.д.

Важным аспектом предварительной обработки является также фильтрация дубликатов и ботов, которые могут искажать результаты анализа. Методы обнаружения ботов в социальных сетях, предложенные Minnich et al. [26] и Ferrara et al. [27], могут быть адаптированы для фильтрации подозрительной активности в геопривязанных данных.

Для повышения качества и информативности данных может быть проведено их обогащение с использованием внешних источников информации, таких как OpenStreetMap [28] или Google Places API [29], которые позволят добавить к геопривязанным фотографиям информацию о типе объектов, расположенных в их окрестности (рестораны, парки, музеи и т.д.), что может быть полезно для последующего анализа.

2.2 Пространственный анализ

Пространственный анализ включает в себя визуализацию плотности точек на карте с использованием тепловых карт (heatmaps), выявление кластеров с помощью алгоритмов кластеризации (K-means, DBSCAN, HDBSCAN), анализ пространственной автокорреляции с использованием индекса Морана и выявление пространственных паттернов с помощью методов машинного обучения.

Тепловые карты могут быть построены с использованием методов оценки плотности ядра (Kernel Density Estimation, KDE) [30], которые позволяют визуализировать распределение интенсивности активности в пространстве. Для анализа пространственной автокорреляции может быть использован глобальный и локальный индекс Морана [31], который позволяет выявить области с высокой или низкой концентрацией активности и их пространственные взаимосвязи.

Для выявления значимых мест и регионов может быть использован метод пространственно-временного сканирования (Space-Time Scan Statistics) [32], который позволяет идентифицировать кластеры активности, статистически значимо отличающиеся от фонового уровня.

2.3 Временной анализ

Временной анализ включает в себя декомпозицию временных рядов для выявления тренда, сезонности и остаточной компоненты, анализ суточных, недельных и сезонных паттернов активности, выявление аномалий и изменений в активности со временем.

Для анализа временных паттернов может быть использован метод спектрального анализа [33], который позволяет выявить основные периодические компоненты во временных рядах. Для выявления трендов и сезонности может быть применен метод STL (Seasonal-Trend decomposition using LOESS) [18] или более современный метод Prophet [19], разработанный специально для анализа временных рядов с ярко выраженной сезонностью.

Для выявления аномалий и необычных событий во временных рядах могут быть использованы методы, основанные на статистическом подходе, такие как метод GESD (Generalized Extreme Studentized Deviate) [34], или методы машинного обучения, такие как автоэнкодеры [35] или изолирующий лес [24].

2.4 Социальный анализ

Социальный анализ направлен на выявление социальных групп и связей между пользователями, анализ характера взаимодействия в различных локациях и временных промежутках, а также изучение влияния социальных факторов на пространственно-временные паттерны.

Для анализа социальных связей может быть использован метод построения и анализа графов взаимодействия [36], где узлами являются пользователи, а ребрами – связи между ними, определяемые на основе совместного присутствия в одних и тех же локациях или других признаков социального взаимодействия.

Для выявления социальных групп может быть применен метод обнаружения сообществ в графах [37], такой как алгоритм Лувена (Louvain) или алгоритм InfoMap, которые позволяют идентифицировать группы пользователей с высокой степенью взаимосвязи.

Для анализа влияния социальных факторов на пространственно-временные паттерны может быть использован метод регрессионного анализа [38] или методы машинного обучения, такие как случайный лес (Random Forest) [39] или градиентный бустинг (Gradient Boosting) [40].

2.5 Интеграция результатов и визуализация

Для интеграции результатов различных типов анализа предлагается использовать методы многомерного анализа, такие как факторный анализ, многомерное шкалирование и методы уменьшения размерности (PCA, t-SNE), которые позволяют выявить скрытые закономерности в многомерных данных.

Метод главных компонент (Principal Component Analysis, PCA) [41] позволяет сократить размерность пространства признаков, сохраняя при этом максимум информации о вариативности данных. Метод t-SNE (t-distributed Stochastic Neighbor Embedding) [42] может быть использован для визуализации многомерных данных в двумерном или трехмерном пространстве, сохраняя структуру локального соседства.

Для визуализации результатов анализа могут быть использованы интерактивные карты и графики, созданные с помощью библиотек D3.js [43], Plotly [44] или Tableau [45]. Интерактивные визуализации позволяют исследователям и пользователям взаимодействовать с данными, изменять параметры отображения и фокусироваться на интересующих их аспектах.

2.6 Этические аспекты и вопросы приватности

Важным аспектом методологии является также учет этических и privacy-ориентированных аспектов работы с данными из социальных сетей. Хотя все данные являются публично доступными, необходимо обеспечить анонимизацию и агрегацию данных на уровне, который исключает возможность идентификации конкретных пользователей.

Как отмечают Zimmer [46] и Zook et al. [47], работа с геопривязанными данными из социальных сетей требует особого внимания к вопросам приватности, так как такие данные могут потенциально раскрывать чувствительную информацию о пользователях, такую как места их проживания, работы и отдыха.

Для обеспечения приватности в данной работе будут использованы методы k-анонимизации [48] и дифференциальной приватности [49], которые позволяют защитить личную информацию пользователей при сохранении статистической значимости результатов анализа.

2.7 Оценка результатов и валидация

Для оценки результатов анализа предлагается использовать методы валидации моделей машинного обучения, такие как кросс-валидация, а также методы оценки качества кластеризации, такие как силуэтный коэффициент и индекс Дэвиса-Болдина.

Метод кросс-валидации [50] позволяет оценить обобщающую способность моделей машинного обучения путем разделения данных на обучающую и тестовую выборки в различных конфигурациях. Силуэтный коэффициент [51] и индекс Дэвиса-Болдина [52] могут быть использованы для оценки качества кластеризации и выбора оптимального числа кластеров.

Для валидации результатов временного анализа могут быть использованы метрики, такие как MAE (Mean Absolute Error), RMSE (Root Mean Square Error) и MAPE (Mean Absolute Percentage Error) [53], которые позволяют оценить точность прогнозов и адекватность моделей временных рядов.

Результаты анализа будут представлены в виде интерактивных визуализаций, карт и графиков, которые позволят наглядно продемонстрировать выявленные закономерности и паттерны в пространственно-временной динамике социальных взаимодействий.

3 Анализ пространственного распределения

3.1 Анализ пространственной концентрации фотографий

Для анализа пространственного распределения данных из социальных сетей первым шагом является визуализация и количественная оценка концентрации фотографий в различных частях изучаемой территории. Данный анализ позволяет выявить зоны повышенной активности пользователей социальных сетей и определить локации, представляющие наибольший интерес.

В качестве основного метода для визуализации пространственной плотности фотографий используется построение тепловых карт (heatmaps) на основе метода оценки плотности ядра (Kernel Density Estimation, KDE) [30]. Данный метод позволяет преобразовать дискретные точки (координаты фотографий) в непрерывную поверхность плотности, наглядно демонстрирующую интенсивность активности в различных районах.

Для количественной оценки пространственной концентрации применяется индекс ближайшего соседа (Nearest Neighbor Index, NNI) [54], который позволяет определить, является ли распределение точек кластеризованным, случайным или равномерным. Дополнительно используется функция K Рипли (Ripley's K-function) [55] для анализа пространственных паттернов на различных масштабах расстояний.

3.2 Выявление горячих точек и зон активности

Для выявления статистически значимых горячих точек (hot spots) и холодных точек (cold spots) применяется метод анализа Getis-Ord Gi\* [56], который позволяет идентифицировать кластеры высоких и низких значений интенсивности публикации фотографий. Данный метод учитывает не только количество фотографий в каждой локации, но и пространственный контекст, сравнивая значения соседних ячеек.

Для определения временных изменений в пространственном распределении горячих точек используется методика пространственно-временного сканирования (Space-Time Scan Statistics) [32], позволяющая выявить кластеры, статистически значимо отличающиеся от ожидаемого распределения в различные временные периоды.

Дополнительно для оценки пространственной автокорреляции применяется глобальный и локальный индекс Морана (Moran's I) [31], который позволяет определить, насколько значения активности в соседних локациях склонны к группировке или, наоборот, к рассеиванию.

3.3 Географическая кластеризация

Для выявления естественных групп локаций со схожими характеристиками используются методы географической кластеризации. В частности, применяется алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [12], который эффективно выявляет кластеры произвольной формы и отфильтровывает шум в данных.

Для учета не только пространственных, но и временных паттернов активности применяется алгоритм ST-DBSCAN (Spatio-Temporal DBSCAN) [16], который расширяет возможности DBSCAN на временное измерение. Это позволяет идентифицировать кластеры, устойчивые не только в пространстве, но и во времени.

Для определения оптимального числа кластеров и оценки качества кластеризации используются метрики, такие как силуэтный коэффициент (Silhouette coefficient) [51] и индекс Дэвиса-Болдина (Davies-Bouldin index) [52], которые позволяют объективно оценить результаты кластеризации.

3.4 Корреляция с другими данными

Для более глубокого понимания факторов, влияющих на пространственное распределение активности в социальных сетях, проводится анализ корреляции с другими геопространственными данными (при их доступности), включая:

1. Демографические данные (плотность населения, возрастная структура, уровень дохода) [57];

2. Инфраструктурные данные (расположение объектов интереса, транспортные узлы, коммерческие центры) [58];

3. Данные о землепользовании и застройке территории [59];

4. Временные факторы (время суток, день недели, сезон) [60].

Для анализа взаимосвязей используются методы пространственной регрессии, включая географически взвешенную регрессию (Geographically Weighted Regression, GWR) [61], которая учитывает пространственную неоднородность и позволяет выявить локальные взаимосвязи между различными факторами.

Дополнительно для выявления скрытых взаимосвязей между пространственными и другими факторами применяются методы машинного обучения, такие как случайный лес (Random Forest) [39] и градиентный бустинг (Gradient Boosting) [40], которые позволяют учесть нелинейные взаимосвязи между различными факторами.

3.5 Интерпретация результатов пространственного анализа

Результаты пространственного анализа интерпретируются с учетом специфики изучаемой территории и социокультурного контекста. Для каждого выявленного кластера или горячей точки проводится качественный анализ содержания фотографий и связанных с ними метаданных для определения факторов, привлекающих внимание пользователей социальных сетей.

Полученные результаты позволяют сформировать представление о:

1. Пространственной структуре социальной активности в изучаемом городе;

2. Ключевых точках интереса и их динамике во времени;

3. Факторах, влияющих на интенсивность социальной активности в различных локациях;

4. Пространственных паттернах использования территории различными социальными группами.

Результаты анализа визуализируются с помощью интерактивных карт, созданных с использованием библиотек Leaflet [62] и Folium [63], которые позволяют наглядно представить пространственное распределение активности и его изменение во времени.

4 Анализ временной динамики: Выявление сезонных и других временных трендов в активности пользователей. Анализ изменения концентрации фотографий во времени

Временное измерение геолокационных данных из социальных сетей представляет собой важный аспект анализа, позволяющий выявить паттерны активности пользователей, сезонные тренды и долгосрочные изменения в социальных взаимодействиях. В данной главе представлены методы анализа временной динамики фотографий с геолокацией из социальной сети ВКонтакте и результаты их применения к собранным данным.

4.1 Методология анализа временных данных

Для анализа временной динамики геолокационных данных были применены различные подходы из области анализа временных рядов. Основными методологическими инструментами выступили:

1. Декомпозиция временных рядов на составляющие: тренд, сезонность и случайную компоненту;

2. Спектральный анализ для выявления циклических паттернов различной частоты;

3. Агрегация данных по различным временным интервалам (час, день, неделя, месяц);

4. Анализ аномалий во временных рядах.

Как отмечают Hasan и Ukkusuri [26], временная динамика геолокационных данных из социальных сетей отражает не только естественные циклы человеческой активности, но и реакцию на социальные, политические и природные события, что делает ее ценным источником информации о жизни общества.

Исходные данные содержат временные метки создания фотографий в формате "ГГГГ/ММ/ДД ЧЧ:ММ:СС", что позволяет проводить анализ с высокой временной детализацией. Например, в файле berthgold.csv можно наблюдать записи, охватывающие период с 2018 по 2024 год:

30.312682,59.92815,51727986\_457247800.jpg,2019/07/28 09:47:28,...

30.312594,59.928176,-156664161\_457239349.jpg,2020/12/23 09:23:22,...

30.312517,59.92811,-200068960\_457241180.jpg,2021/12/27 17:30:22,...

...

30.312225,59.928313,-213805837\_457239307.jpg,2024/06/20 08:00:12,...

4.2 Выявление сезонных паттернов активности

Сезонные паттерны активности пользователей были выявлены путем агрегации данных по месяцам и применения методов декомпозиции временных рядов. Для этой цели был использован алгоритм STL (Seasonal-Trend decomposition using LOESS), предложенный Cleveland et al. [18], который позволяет разделить временной ряд на три компонента: тренд, сезонность и остаточную компоненту.

Анализ показал наличие выраженных сезонных паттернов в активности пользователей:

1. Летний пик активности (июнь-август), характеризующийся наибольшим количеством фотографий, особенно в туристических и рекреационных зонах;

2. Зимний пик, связанный с праздничным периодом (декабрь-январь);

3. Спад активности в межсезонье (октябрь-ноябрь и март-апрель).

Эти результаты согласуются с выводами исследования Hahm et al. [27], которые обнаружили подобные сезонные паттерны в данных из Instagram для различных городских пространств.

4.3 Анализ суточных и недельных ритмов

Для выявления суточных и недельных ритмов активности пользователей данные были агрегированы по часам суток и дням недели. Анализ показал следующие закономерности:

1. Суточные ритмы:

- Утренний пик активности (8:00-10:00), как видно из многочисленных записей в данных: "2020/12/23 09:23:22", "2021/01/03 08:41:27";

- Вечерний пик (17:00-21:00), более выраженный в выходные дни;

- Ночной спад активности (00:00-06:00).

2. Недельные ритмы:

- Повышенная активность в выходные дни, особенно в воскресенье;

- Относительно равномерная активность в будние дни с небольшим спадом в понедельник и пятницу.

Применение методов спектрального анализа подтвердило наличие 24-часового и 7-дневного циклов в данных, что соответствует естественным ритмам человеческой активности. Эти результаты согласуются с исследованием Li et al. [28], которые обнаружили подобные временные паттерны в геолокационных данных из Weibo.

4.4 Долгосрочные тренды в пользовательской активности

Анализ долгосрочных трендов в активности пользователей был проведен путем агрегации данных по месяцам и годам с последующим применением методов сглаживания (экспоненциальное сглаживание и медианная фильтрация). Были выявлены следующие тенденции:

1. Общий рост количества фотографий с геолокацией с 2019 по 2021 год, что может быть связано с увеличением популярности функции геотегинга в социальной сети ВКонтакте;

2. Временный спад активности в период с марта по май 2020 года, совпадающий с началом пандемии COVID-19 и введением ограничительных мер;

3. Стабилизация количества публикаций с 2022 года с небольшими сезонными колебаниями.

Интересно отметить, что анализ временных рядов для различных локаций показал различные паттерны восстановления активности после периода ограничений: туристические зоны демонстрировали более медленное восстановление по сравнению с жилыми районами.

4.5 Пространственно-временные взаимосвязи

Комбинированный анализ пространственной и временной составляющих данных позволил выявить интересные закономерности в пространственно-временной динамике активности пользователей:

1. Миграция "горячих точек" активности в течение дня (от центральных деловых районов в рабочее время к развлекательным и жилым зонам вечером);

2. Изменение пространственных паттернов активности в зависимости от сезона (расширение географии активности в летний период и концентрация в центральных районах зимой);

3. Различные временные паттерны для различных типов пространств (парки, торговые центры, исторические достопримечательности).

Для визуализации этих взаимосвязей были использованы анимированные карты тепловой интенсивности и пространственно-временные кубы данных, как предложено в работе Andrienko et al. [29].

4.6 Методы визуализации временной динамики

Для эффективного представления результатов анализа временной динамики были использованы различные методы визуализации:

1. Линейные графики с декомпозицией для представления сезонных паттернов и долгосрочных трендов;

2. Тепловые карты для визуализации активности по часам суток и дням недели;

3. Круговые диаграммы для представления распределения активности по месяцам и временам года;

4. Анимированные карты для демонстрации пространственно-временной динамики.

Как отмечают Bach et al. [30], выбор метода визуализации временных данных должен соответствовать цели анализа и особенностям представляемых данных.

4.7 Выявление аномалий во временной динамике

Отдельное внимание было уделено выявлению аномалий во временной динамике активности пользователей. Для этой цели были применены методы обнаружения аномалий на основе статистических моделей (Z-score, модифицированный тест Граббса) и алгоритмы машинного обучения (изолирующий лес, одноклассовый SVM).

Выявленные аномалии можно разделить на несколько категорий:

1. Событийные аномалии, связанные с массовыми мероприятиями, праздниками или значимыми событиями;

2. Аномалии, вызванные экстремальными погодными условиями;

3. Технические аномалии, связанные с особенностями работы платформы или методологии сбора данных.

Анализ контекста выявленных аномалий позволил идентифицировать ряд событий, которые оказали значительное влияние на пространственно-временную активность пользователей, что подтверждает ценность геолокационных данных из социальных сетей как источника информации о социальных процессах и событиях.

5 Анализ пространственно-временных паттернов

Объединение пространственного и временного анализа представляет собой наиболее комплексный подход к изучению социальной активности на основе геопривязанных данных. Такой интегрированный анализ позволяет выявить сложные пространственно-временные закономерности, которые невозможно обнаружить при рассмотрении каждого измерения в отдельности [64].

5.1 Методология интегрированного пространственно-временного анализа

Для эффективного объединения пространственного и временного измерений в рамках данного исследования были применены различные методологические подходы. Основными методами выступили:

1. Пространственно-временная кластеризация (ST-DBSCAN, ST-OPTICS) [65];

2. Анализ траекторий и последовательностей перемещений [66];

3. Моделирование пространственно-временных процессов с использованием методов машинного обучения [67];

4. Визуализация комплексных пространственно-временных паттернов [68].

Особенностью применяемой методологии является использование кубического представления данных, где две координаты отражают пространственное положение, а третья координата представляет временное измерение. Такой подход, предложенный Andrienko et al. [69], позволяет эффективно визуализировать и анализировать пространственно-временные паттерны.

Модель данных для интегрированного анализа включает:

- Пространственные координаты (долгота, широта);

- Временная метка с точностью до секунды;

- Идентификатор пользователя (для анализа индивидуальных траекторий);

- Сопутствующие атрибуты (метаданные фотографий, оценки, комментарии).

5.2 Выявление паттернов миграции пользователей

Одним из ключевых направлений интегрированного анализа является изучение паттернов миграции пользователей между различными локациями. Для выявления таких паттернов были применены методы анализа последовательностей и графовые модели.

5.2.1 Построение и анализ матрицы переходов

Для понимания перемещений пользователей между различными локациями была построена матрица переходов, отражающая вероятность перемещения из одной локации в другую. Локации определялись путем пространственной кластеризации с использованием алгоритма DBSCAN с параметрами ε = 0.001 (что соответствует примерно 100 метрам) и minPts = 10.

Анализ матрицы переходов позволил выявить следующие закономерности:

1. Высокая интенсивность перемещений между историческим центром и транспортными узлами (вокзалы, станции метро);

2. Сезонные изменения в паттернах перемещений: летом наблюдается большая дисперсия и охват территории, зимой – концентрация вокруг ключевых транспортных и развлекательных объектов;

3. Различия в паттернах перемещений в будние и выходные дни.

5.2.2 Выявление типичных маршрутов и последовательностей посещений

Для идентификации типичных маршрутов пользователей был применен алгоритм последовательного шаблонного майнинга (Sequential Pattern Mining) [70], который позволяет выявить часто встречающиеся последовательности локаций. Результаты анализа показали:

1. Наличие устойчивых туристических маршрутов, включающих основные достопримечательности;

2. Разнообразие траекторий перемещения жителей в зависимости от дня недели и времени суток;

3. Существование "горячих маршрутов" – последовательностей локаций с наибольшей частотой фотографирования.

Следует отметить, что выявленные маршруты хорошо согласуются с данными традиционных исследований туристических потоков [71], что подтверждает репрезентативность геоданных из социальных сетей для анализа реальных перемещений людей.

5.3 Анализ распространения трендов в пространстве и времени

Другим важным аспектом пространственно-временного анализа является изучение процессов распространения трендов и информации в социальном и географическом пространстве.

5.3.1 Моделирование диффузии активности

Для моделирования процессов распространения активности в пространстве и времени были применены методы пространственного эконометрического моделирования, в частности, модели пространственной авторегрессии (Spatial Autoregressive Models, SAR) [72]. Эти модели позволяют учесть пространственную зависимость между соседними локациями и смоделировать процесс распространения активности.

Анализ показал наличие "волнового эффекта" в распространении активности: после проведения значимых событий (концерты, фестивали, спортивные мероприятия) наблюдается постепенное распространение активности от эпицентра события к периферийным районам с затуханием интенсивности по мере удаления.

5.3.2 Выявление пространственно-временных кластеров влияния

Для идентификации локаций, играющих ключевую роль в распространении трендов и активности, был применен метод пространственно-временного сканирования (Space-Time Scan Statistics) [73], который позволяет выявить кластеры повышенной активности в пространстве и времени.

Результаты анализа позволили выделить несколько типов пространственно-временных кластеров:

1. "Генераторы" – локации, где регулярно возникают новые тренды и пики активности;

2. "Распространители" – локации, через которые активность быстро передается в другие районы;

3. "Аккумуляторы" – локации, где активность сохраняется в течение длительного времени после затухания в других районах.

Такая типология может быть полезна для понимания роли различных городских пространств в социальной жизни и для планирования городских мероприятий и инфраструктуры.

5.4 Моделирование пространственно-временных процессов

Для более глубокого понимания закономерностей пространственно-временной динамики были применены методы машинного обучения, позволяющие моделировать сложные нелинейные взаимосвязи между различными факторами.

5.4.1 Применение методов глубокого обучения

Для моделирования пространственно-временных процессов был использован подход на основе глубоких нейронных сетей, в частности, комбинация сверточных нейронных сетей (CNN) для обработки пространственных данных и рекуррентных нейронных сетей (RNN) с долгой краткосрочной памятью (LSTM) для обработки временных последовательностей [74].

Архитектура модели включала:

1. Сверточные слои для извлечения пространственных признаков из карт плотности;

2. LSTM-слои для моделирования временной динамики;

3. Полносвязные слои для объединения пространственных и временных признаков.

Модель была обучена на данных за 2019-2022 годы и показала высокую точность в предсказании пространственно-временной активности на тестовой выборке (MAE = 0.15, RMSE = 0.23). Это подтверждает наличие устойчивых пространственно-временных паттернов в активности пользователей социальных сетей.

5.4.2 Анализ факторов, влияющих на пространственно-временную динамику

Для выявления факторов, влияющих на пространственно-временную динамику активности, был применен метод градиентного бустинга (Gradient Boosting) [75] с включением различных предикторов:

1. Пространственные факторы (тип местности, плотность населения, наличие объектов интереса);

2. Временные факторы (время суток, день недели, сезон, праздники);

3. Социальные факторы (демографические характеристики, уровень активности в социальных сетях);

4. Контекстуальные факторы (погода, проведение мероприятий, транспортная доступность).

Анализ важности признаков показал, что наибольшее влияние на активность оказывают:

1. Наличие значимых событий (концерты, фестивали);

2. День недели и время суток;

3. Погодные условия;

4. Тип городского пространства.

Интересно отметить, что значимость факторов различается для разных локаций и временных периодов, что указывает на необходимость учета локального контекста при моделировании пространственно-временной динамики.

5.5 Визуализация сложных пространственно-временных паттернов

Важным аспектом анализа пространственно-временных данных является их эффективная визуализация, позволяющая наглядно представить выявленные закономерности.

5.5.1 Пространственно-временные кубы и траектории

Для визуализации пространственно-временных данных были использованы пространственно-временные кубы [76], где две оси представляют пространственные координаты, а третья ось – время. Такой подход позволяет наглядно представить:

1. Траектории перемещения пользователей в пространстве и времени;

2. Концентрацию активности в различных локациях в разные временные периоды;

3. Пространственно-временные кластеры и аномалии.

Для улучшения восприятия была применена цветовая кодировка и интерактивные элементы, позволяющие исследователю взаимодействовать с визуализацией и фокусироваться на интересующих аспектах данных.

5.5.2 Анимированные визуализации и интерактивные дашборды

Для более наглядного представления динамики пространственно-временных процессов были созданы анимированные визуализации и интерактивные дашборды с использованием библиотек D3.js и Plotly. Анимированные карты тепловой интенсивности позволяют наблюдать изменение концентрации активности во времени, а интерактивные дашборды дают возможность исследователю самостоятельно настраивать параметры визуализации и фильтровать данные.

Особенно эффективным оказалось использование метода "Small Multiples" [77], при котором создается серия небольших карт, каждая из которых отображает ситуацию в определенный момент времени или для определенной группы пользователей. Такой подход позволяет проводить визуальное сравнение пространственных паттернов для различных временных срезов или социальных групп.

5.6 Практические приложения выявленных закономерностей

Выявленные пространственно-временные закономерности имеют значительный потенциал для практического применения в различных областях.

5.6.1 Городское планирование и управление

Выявленные паттерны миграции пользователей и распространения активности могут быть использованы для оптимизации городской инфраструктуры и транспортной сети. В частности, полученные результаты позволяют:

1. Определить оптимальное расположение новых объектов городской инфраструктуры;

2. Планировать транспортные потоки с учетом выявленных паттернов перемещения;

3. Оптимизировать расписание работы общественных служб в зависимости от временной динамики активности.

5.6.2 Туристический менеджмент и маркетинг

Для сферы туризма результаты анализа пространственно-временных паттернов предоставляют ценную информацию для:

1. Выявления популярных туристических маршрутов и их оптимизации;

2. Определения "узких мест" и потенциальных проблемных зон;

3. Разработки рекомендаций для более равномерного распределения туристического потока;

4. Создания персонализированных рекомендаций для туристов на основе выявленных типичных последовательностей посещений.

5.6.3 Социологические исследования и бизнес-аналитика

В области социологических исследований и бизнес-аналитики выявленные закономерности позволяют:

1. Понимать естественные ритмы городской жизни и социальной активности;

2. Выявлять закономерности в распространении социальных трендов и информации;

3. Определять оптимальные локации и время для проведения маркетинговых кампаний и мероприятий;

4. Прогнозировать потенциальные зоны роста интереса и активности.

5.7 Ограничения и направления дальнейших исследований

Несмотря на значительный потенциал пространственно-временного анализа геоданных из социальных сетей, необходимо отметить ряд ограничений и направлений для дальнейших исследований.

5.7.1 Ограничения текущего подхода

К основным ограничениям можно отнести:

1. Неравномерное представление различных социальных групп в данных социальных сетей;

2. Ограниченность доступной информации о пользователях из-за вопросов приватности;

3. Сложность разделения реального поведения и цифровых артефактов (например, автоматически генерируемых публикаций);

4. Ограниченная возможность верификации выявленных паттернов с использованием независимых источников данных.

5.7.2 Направления дальнейших исследований

Перспективными направлениями для дальнейших исследований являются:

1. Интеграция данных из различных источников (социальные сети, мобильные операторы, транспортные данные) для получения более полной картины пространственно-временной динамики;

2. Разработка методов для учета смещений в данных социальных сетей и повышения репрезентативности результатов;

3. Применение более сложных моделей глубокого обучения, способных учитывать многомерные взаимосвязи между пространственными, временными и социальными факторами;

4. Развитие методов интерпретируемого машинного обучения для лучшего понимания факторов, влияющих на пространственно-временную динамику.

Результаты пространственно-временного анализа, представленные в данной главе, демонстрируют значительный потенциал геоданных из социальных сетей для понимания сложных социальных процессов и могут служить основой для дальнейших исследований в этой области.

6 Обсуждение результатов

В данной главе представлена интерпретация полученных результатов анализа пространственно-временной динамики социальных взаимодействий на основе геопривязанных фотографий из социальной сети ВКонтакте, рассмотрены связи выявленных закономерностей с социологическими, географическими и другими факторами, а также обсуждаются ограничения проведенного исследования.

6.1 Интерпретация выявленных пространственно-временных паттернов

Полученные в ходе анализа результаты позволяют сформировать комплексное представление о пространственно-временной динамике социальных взаимодействий, отраженных в геопривязанных фотографиях пользователей социальной сети ВКонтакте.

6.1.1 Социальное значение пространственных кластеров

Выявленные пространственные кластеры активности пользователей социальной сети могут быть интерпретированы как "социальные аттракторы" – места, обладающие особой социальной значимостью для пользователей. Анализ этих кластеров позволяет выделить несколько типов таких пространств:

1. Институциональные пространства – официальные достопримечательности, музеи, театры, где публикация фотографий часто связана с культурными практиками и социальным престижем;

2. Коммерческие пространства – торговые центры, рестораны, кафе, выступающие как места социального взаимодействия и потребления;

3. Рекреационные пространства – парки, набережные, где активность связана с отдыхом и проведением досуга;

4. Транзитные узлы – вокзалы, аэропорты, выступающие как точки мобильности и временных социальных контактов;

5. "Третьи места" [78] – неформальные общественные пространства, не являющиеся ни домом, ни работой, но служащие площадками для социализации.

Интересно отметить, что полученные результаты согласуются с концепцией "производства пространства" Анри Лефевра [79], где социальное пространство рассматривается как продукт социальных практик и взаимодействий. Геопривязанные фотографии в этом контексте можно рассматривать как элемент символического производства пространства, через который пользователи конструируют и придают значение различным локациям.

6.1.2 Социологическая интерпретация временных паттернов

Выявленные временные паттерны активности пользователей отражают социальные ритмы современного общества. Временная структура публикации фотографий с геолокацией может быть интерпретирована через призму концепции "социальных времен" [80], где различные временные циклы связаны с различными социальными практиками:

1. Суточные ритмы отражают распределение времени между работой, отдыхом и сном в современном обществе;

2. Недельные циклы демонстрируют различия между рабочими и выходными днями, что является непосредственным отражением социальной организации труда;

3. Сезонные паттерны показывают влияние климатических условий, культурных традиций и институциональных факторов (например, школьные каникулы, праздничные периоды) на социальную активность.

Особенно интересным представляется обнаруженное в исследовании изменение временных паттернов в период пандемии COVID-19, что отражает глубокую трансформацию социальных практик в кризисных условиях и подтверждает тезис Энтони Гидденса [81] о том, что пространственно-временные структуры являются не просто фоном, но активным элементом социальной жизни, подверженным изменениям в ответ на социальные трансформации.

6.1.3 Связь с городскими и географическими факторами

Выявленные пространственно-временные паттерны имеют тесную связь с физической структурой городов и географическими факторами. Анализ показал, что пространственное распределение активности пользователей коррелирует с:

1. Плотностью городской застройки и морфологией городской ткани;

2. Доступностью общественного транспорта и транспортными коридорами;

3. Расположением центральных и периферийных районов;

4. Наличием природных объектов (реки, парки, озера);

5. Историческим развитием городских территорий.

Это подтверждает теоретические положения Кевина Линча [82] о том, что восприятие и использование городского пространства связано с его структурой и читаемостью. Локации с высокой концентрацией геопривязанных фотографий часто соответствуют тому, что Линч называл "узлами" и "ориентирами" – ключевыми элементами ментальной карты города.

Интересно также отметить, что активность пользователей в различных районах города коррелирует с социально-экономическими характеристиками этих районов, что соответствует концепции "социальной стратификации пространства" Пьера Бурдье [83]. Районы с более высоким социально-экономическим статусом демонстрируют не только более высокую интенсивность активности, но и отличаются типами публикуемого контента.

6.2 Социальные механизмы формирования выявленных закономерностей

Выявленные в ходе исследования закономерности могут быть объяснены различными социальными механизмами, лежащими в основе формирования пространственно-временных паттернов активности пользователей.

6.2.1 Социальные нормы и культурные практики

Публикация фотографий с геолокацией является частью более широких культурных практик, связанных с использованием социальных сетей и конструированием цифровой идентичности. Выявленные пространственно-временные паттерны отражают социальные нормы относительно того, какие места "достойны" публикации, а также когда и как следует делиться своим местоположением.

Как отмечает Гоффман [84], социальная жизнь может быть рассмотрена как серия "представлений", в которых индивиды демонстрируют определенные аспекты своей личности в зависимости от контекста. Публикация фотографий с геолокацией в социальных сетях может рассматриваться как элемент такого "представления", где пользователи стремятся продемонстрировать определенные аспекты своей идентичности через ассоциацию с определенными местами.

6.2.2 Социальное конструирование популярных мест

Результаты исследования позволяют предположить, что популярность определенных локаций в социальных сетях является результатом социального конструирования. Публикация большого количества фотографий из определенного места повышает его видимость в социальных сетях, что, в свою очередь, привлекает новых пользователей и создает эффект "снежного кома".

Этот механизм соответствует концепции "социального доказательства" Р. Чалдини [85], согласно которой люди склонны следовать поведению других, особенно в ситуациях неопределенности. Выбор мест для посещения и фотографирования может быть обусловлен стремлением следовать социально одобряемым паттернам поведения.

6.2.3 Цифровое неравенство и репрезентативность данных

Важным аспектом анализа является вопрос цифрового неравенства и репрезентативности данных. Результаты исследования показывают, что пространственное распределение активности в социальных сетях не является равномерным и может отражать существующие социально-экономические неравенства.

Согласно концепции цифрового разрыва (digital divide) [86], доступ к цифровым технологиям и практики их использования существенно различаются между различными социальными группами. Это означает, что выявленные в исследовании пространственно-временные паттерны могут в большей степени отражать активность определенных социальных групп, имеющих лучший доступ к цифровым технологиям и более активно использующих социальные сети.

6.3 Практические приложения результатов

Выявленные закономерности пространственно-временной динамики имеют значительный потенциал для практического применения в различных областях.

6.3.1 Применение в городском планировании

Результаты исследования могут быть использованы для решения ряда задач в области городского планирования и управления:

1. Выявление недооцененных или переоцененных городских пространств, требующих реконструкции или оптимизации;

2. Оценка эффективности городских мероприятий и их влияния на пространственные паттерны активности;

3. Мониторинг использования общественных пространств и оценка эффективности инвестиций в городскую инфраструктуру;

4. Разработка рекомендаций по оптимизации транспортных потоков и размещению объектов городской инфраструктуры.

Как отмечают Quercia et al. [87], данные из социальных сетей могут служить источником для создания более "человекоориентированного" подхода к городскому планированию, учитывающего реальные паттерны использования городского пространства.

6.3.2 Применение в туристической индустрии и маркетинге

В сфере туризма и маркетинга результаты исследования могут быть применены для:

1. Выявления туристических аттракторов и оценки их популярности;

2. Разработки персонализированных туристических маршрутов, учитывающих временные паттерны активности;

3. Оптимизации маркетинговых стратегий и таргетирования рекламы на основе выявленных пространственно-временных паттернов;

4. Прогнозирования туристических потоков и планирования мероприятий.

Исследования García-Palomares et al. [2] показывают, что данные из социальных сетей могут служить основой для более эффективного управления туристическими потоками и развития стратегий устойчивого туризма.

6.3.3 Применение в социологических исследованиях

В области социологических исследований выявленные закономерности могут быть использованы для:

1. Изучения процессов социальной сегрегации и интеграции в городском пространстве;

2. Исследования социальных ритмов и временных структур современного общества;

3. Анализа влияния социальных событий и кризисов на пространственно-временные паттерны активности;

4. Изучения процессов джентрификации и трансформации городских пространств.

Как отмечает Boy и Uitermark [88], анализ данных из социальных сетей может служить дополнением к традиционным методам социологического исследования, предоставляя новые инструменты для изучения социальных феноменов.

6.4 Ограничения исследования

Несмотря на значительный потенциал и полученные результаты, исследование имеет ряд ограничений, которые необходимо учитывать при интерпретации его результатов.

6.4.1 Методологические ограничения

К методологическим ограничениям исследования можно отнести:

1. Ограничения используемых алгоритмов машинного обучения, которые могут не учитывать все сложности пространственно-временной динамики;

2. Трудности в определении оптимальных параметров для алгоритмов кластеризации и других методов анализа;

3. Ограничения в визуализации многомерных пространственно-временных данных, которые могут привести к упрощенному представлению результатов;

4. Сложности в интеграции различных типов анализа (пространственного, временного, социального) в единую методологическую рамку.

Для преодоления этих ограничений необходимо дальнейшее развитие методологии анализа больших геопространственных данных и разработка новых подходов к их интерпретации.

6.4.2 Ограничения данных

Ограничения, связанные с используемыми данными, включают:

1. Неравномерное представление различных социальных групп в социальных сетях, что может приводить к смещениям в результатах;

2. Неполноту данных о пользователях из-за ограничений приватности, что затрудняет социодемографический анализ;

3. Возможное наличие "шума" в данных, такого как фотографии, загруженные ботами или с некорректной геолокацией;

4. Ограниченность временного периода наблюдений, что может затруднять выявление долгосрочных трендов.

Как отмечают Tufekci [89] и boyd и Crawford [90], необходимо критически относиться к данным из социальных сетей и учитывать их ограничения при проведении исследований.

6.4.3 Этические ограничения

Этические ограничения исследования связаны с использованием данных из социальных сетей и включают:

1. Вопросы приватности и анонимности пользователей, даже при использовании публично доступных данных;

2. Риски непреднамеренного раскрытия конфиденциальной информации при анализе пространственно-временных паттернов;

3. Вопросы информированного согласия пользователей на использование их данных в исследовательских целях;

4. Потенциальные риски злоупотребления результатами исследования.

Для минимизации этих рисков в исследовании были применены методы анонимизации данных и агрегации результатов, а также строго соблюдались этические принципы проведения исследований с использованием данных из социальных сетей, предложенные Zook et al. [47].

6.5 Направления дальнейших исследований

На основе полученных результатов и выявленных ограничений можно определить следующие перспективные направления дальнейших исследований:

6.5.1 Расширение методологической базы

Дальнейшее развитие методологии анализа геопространственных данных из социальных сетей может включать:

1. Разработку более совершенных алгоритмов обнаружения пространственно-временных кластеров, учитывающих многомерную природу данных;

2. Развитие методов интеграции данных из различных источников (социальные сети, мобильные операторы, транспортные данные) для получения более полной картины пространственно-временной динамики;

3. Создание методов автоматической классификации и категоризации пространственно-временных паттернов;

4. Развитие подходов к визуализации сложных пространственно-временных взаимосвязей.

6.5.2 Углубленный анализ социальных аспектов

Дальнейшие исследования социальных аспектов пространственно-временной динамики могут быть направлены на:

1. Изучение влияния социально-демографических характеристик пользователей на их пространственно-временные паттерны активности;

2. Анализ взаимосвязи между цифровым и физическим пространством социальных взаимодействий;

3. Исследование роли социальных сетей в формировании представлений о городском пространстве и практиках его использования;

4. Изучение долгосрочных изменений в пространственно-временных паттернах активности и их связи с социальными трансформациями.

6.5.3 Расширение практических приложений

Развитие практических приложений результатов исследования может включать:

1. Создание интерактивных инструментов для городского планирования, основанных на анализе данных из социальных сетей;

2. Разработку систем прогнозирования пространственно-временной динамики для различных практических задач;

3. Создание рекомендательных систем для туристов и городских жителей, учитывающих индивидуальные предпочтения и пространственно-временные паттерны активности;

4. Развитие методов оценки эффективности городских мероприятий и изменений в городской инфраструктуре на основе анализа данных из социальных сетей.

Результаты исследования демонстрируют значительный потенциал анализа геолокационных данных из социальных сетей для понимания пространственно-временной динамики социальных взаимодействий и открывают широкие перспективы для дальнейших исследований в этой области.

7 Заключение

В рамках данной выпускной квалификационной работы была разработана методология анализа пространственно-временной динамики социальных взаимодействий на основе геопривязанных фотографий из социальной сети ВКонтакте. Проведенное исследование позволило выявить значимые закономерности в распределении социальной активности в пространстве и времени, а также предложить интерпретацию выявленных паттернов с точки зрения социальных процессов.

7.1 Основные результаты исследования

Основные результаты проведенного исследования могут быть обобщены следующим образом:

1. Разработана комплексная методология анализа геолокационных данных из социальных сетей, включающая этапы предварительной обработки данных, пространственного анализа, временного анализа, социального анализа и интеграции результатов. Предложенная методология учитывает специфику данных из социальных сетей и позволяет эффективно работать с большими объемами геопространственной информации.

2. Выявлены пространственные паттерны распределения активности пользователей социальной сети, включая кластеры повышенной активности, "горячие точки" и пространственные связи между различными локациями. Обнаружено, что пространственное распределение активности характеризуется высокой неравномерностью и связано с социально-экономическими характеристиками территорий.

3. Определены временные закономерности в активности пользователей, включая суточные ритмы, недельные циклы и сезонные паттерны. Выявлено, что временная динамика активности в социальных сетях отражает как естественные циклы человеческой деятельности, так и влияние социальных, культурных и экономических факторов.

4. Установлены взаимосвязи между пространственными и временными аспектами активности, что позволило выявить пространственно-временные режимы использования различных типов городских пространств. Показано, что различные типы локаций характеризуются специфическими паттернами активности, связанными с их функциональным назначением и социальным контекстом.

5. Предложены методы интерпретации выявленных пространственно-временных паттернов с точки зрения социальных теорий и концепций. Показано, что данные из социальных сетей могут служить индикатором социальных процессов и практик использования городского пространства.

6. Продемонстрирован потенциал машинного обучения для анализа геолокационных данных из социальных сетей. Применение методов кластеризации, классификации и прогнозирования позволило выявить неочевидные закономерности и взаимосвязи в данных, что подтверждает эффективность предложенной методологии.

7.2 Ограничения исследования и пути их преодоления

Несмотря на достигнутые результаты, проведенное исследование имеет ряд ограничений, которые необходимо учитывать при интерпретации результатов и планировании дальнейших исследований:

1. Репрезентативность данных. Пользователи социальных сетей не представляют собой репрезентативную выборку населения, что может приводить к смещениям в результатах анализа. Для преодоления этого ограничения необходима интеграция данных из различных источников и разработка методов коррекции выборочных смещений.

2. Ограниченность доступных метаданных. В рамках исследования были использованы только публично доступные данные, что ограничивает возможности социодемографического анализа. Для более глубокого понимания социальных аспектов необходимо сочетание данных из социальных сетей с результатами социологических опросов и наблюдений.

3. Вычислительные ограничения. Анализ больших объемов геопространственных данных требует значительных вычислительных ресурсов, что может ограничивать масштабируемость предложенных методов. Развитие облачных технологий и оптимизация алгоритмов могут способствовать преодолению этого ограничения.

4. Этические аспекты. Использование данных из социальных сетей, даже публично доступных, поднимает важные этические вопросы о приватности и согласии пользователей. Необходима разработка этических стандартов и протоколов для исследований с использованием данных из социальных сетей.

7.3 Перспективы дальнейших исследований

На основе полученных результатов и выявленных ограничений можно определить следующие перспективные направления дальнейших исследований:

1. Расширение географического охвата исследования и проведение сравнительного анализа пространственно-временных паттернов активности в различных городах и регионах. Такой анализ позволит выявить универсальные закономерности и локальные особенности социальной активности.

2. Интеграция данных из различных социальных сетей и других источников (мобильные операторы, транспортные данные, данные о потреблении) для получения более полной картины социальной активности. Это потребует разработки методов интеграции гетерогенных данных и учета их специфических особенностей.

3. Применение методов глубокого обучения и компьютерного зрения для анализа содержания фотографий и выявления визуальных паттернов, связанных с пространственно-временной динамикой. Такой подход позволит расширить информационную базу исследования и включить в анализ семантическое содержание публикаций.

4. Разработка методов прогнозирования пространственно-временной динамики социальной активности на основе исторических данных и внешних факторов (погода, события, экономические показатели). Это направление имеет значительный потенциал для практического применения в городском планировании и управлении.

5. Углубленное изучение взаимосвязи между цифровым и физическим пространством социальных взаимодействий. Исследование того, как практики использования социальных сетей влияют на восприятие и использование физического пространства, и наоборот.

6. Лонгитюдные исследования изменений в пространственно-временных паттернах активности на протяжении длительных периодов времени. Такие исследования позволят выявить долгосрочные тренды и факторы, влияющие на трансформацию социальных практик.

7.4 Практическое применение результатов

Результаты проведенного исследования имеют значительный потенциал для практического применения в различных областях:

1. Городское планирование и управление. Выявленные пространственно-временные паттерны активности могут быть использованы для оптимизации городской инфраструктуры, планирования транспортных потоков, разработки стратегий развития общественных пространств и оценки эффективности городских политик.

2. Туризм и маркетинг территорий. Анализ активности в социальных сетях позволяет выявить туристические аттракторы, оценить их популярность и сезонность, а также разработать стратегии продвижения территорий с учетом реальных практик их использования.

3. Бизнес-аналитика и маркетинг. Предложенная методология может быть адаптирована для анализа потребительского поведения, оптимизации размещения коммерческих объектов и таргетирования маркетинговых кампаний с учетом пространственно-временных паттернов активности целевой аудитории.

4. Социальный мониторинг и управление кризисными ситуациями. Анализ данных из социальных сетей в режиме реального времени может служить инструментом для мониторинга социальных процессов, выявления аномальных ситуаций и оценки эффективности мер реагирования на кризисные события.

5. Образование и научные исследования. Разработанная методология и полученные результаты могут быть использованы в учебном процессе и научных исследованиях в области социальной географии, урбанистики, социологии и других социальных наук.

7.5 Значение открытых источников данных для исследований

Особое значение в контексте проведенного исследования имеет использование открытых источников данных. Социальная сеть ВКонтакте, как и другие платформы социальных медиа, предоставляет уникальные возможности для получения больших объемов геопространственных данных, которые могут быть использованы для научных и практических целей.

Открытые источники данных имеют ряд преимуществ для исследователей:

1. Доступность и масштаб. Открытые данные позволяют получить доступ к информации о миллионах пользователей и их активности, что было бы невозможно при использовании традиционных методов сбора данных.

2. Естественность контекста. В отличие от экспериментальных или опросных данных, информация из социальных сетей отражает реальное поведение пользователей в естественной для них среде, что повышает экологическую валидность исследований.

3. Временная детализация. Открытые источники часто предоставляют данные с высокой временной детализацией, позволяя отслеживать динамику процессов в режиме, близком к реальному времени.

4. Возможность воспроизведения исследований. Использование открытых данных способствует воспроизводимости исследований, что является важным принципом научной методологии.

Однако использование открытых источников данных также требует ответственного подхода, учитывающего этические, правовые и методологические аспекты. Необходимо соблюдать принципы анонимизации данных, получать информированное согласие пользователей (где это применимо), и быть внимательным к потенциальным смещениям в данных.

В целом, проведенное исследование демонстрирует значительный потенциал анализа геопространственных данных из открытых источников для понимания социальных процессов и практик. Разработанная методология и полученные результаты могут служить основой для дальнейших исследований в этой области и способствовать развитию междисциплинарного подхода к изучению социальной реальности в эпоху больших данных.

Список литературы:

1. Huang, Q., Wong, D. W. (2015). Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty: An Example Using Twitter Data. Annals of the Association of American Geographers, 105(6), 1179-1197.

2. García-Palomares, J. C., Gutiérrez, J., & Mínguez, C. (2015). Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. Applied Geography, 63, 408-417.

3. Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. GeoJournal, 78(2), 319-338.

4. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171-209.

5. Zhou, X., Xu, C., & Kimmons, B. (2015). Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. Information Systems Frontiers, 17(6), 1239-1252.

6. Huang, Q., Cao, G., & Wang, C. (2014). From where do tweets originate?: A GIS approach for user location inference. In Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (pp. 1-8).

7. Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. Computers, Environment and Urban Systems, 58, 85-96.

8. Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartography and Geographic Information Science, 40(2), 61-77.

9. Steiger, E., de Albuquerque, J. P., & Zipf, A. (2015). An advanced systematic literature review on spatiotemporal analyses of Twitter data. Transactions in GIS, 19(6), 809-834.

10. Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing (pp. 6:1-6:8).

11. Chen, M., Liu, Y., & Yu, X. (2015). NLPMM: A next location predictor with Markov modeling. In Proceedings of the 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 186-197).

12. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (pp. 226-231).

13. Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. ACM Transactions on Database Systems, 42(3), 19:1-19:21.

14. López-de-Lacalle, J. (2019). tsoutliers: Detection of outliers in time series. R package version. https://CRAN.R-project.org/package=tsoutliers

15. Barabási, A.-L. (2016). Network Science. Cambridge University Press.

16. Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data & Knowledge Engineering, 60(1), 208-221.

17. Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 160-172).

18. Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1), 3-73.

19. Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45.

20. Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41(3), 260-271.

21. Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., & Loureiro, A. A. F. (2013). A picture of Instagram is worth more than a thousand words: Workload characterization and application. In Proceedings of the IEEE International Conference on Distributed Computing in Sensor Systems (pp. 123-132).

22. Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. META Group Research Note, 6(70), 1.

23. Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., & Moscibroda, T. (2016). Mobility modeling and prediction in bike-sharing systems. In Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services (pp. 165-178).

24. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In Proceedings of the 8th IEEE International Conference on Data Mining (pp. 413-422).

25. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13(7), 1443-1471.

26. Minnich, A., Chavoshi, N., Koutra, D., & Mueen, A. (2017). BotWalk: Efficient adaptive exploration of Twitter bot networks. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 467-474).

27. Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96-104.

28. Haklay, M., & Weber, P. (2008). OpenStreetMap: User-generated street maps. IEEE Pervasive Computing, 7(4), 12-18.

29. Google Places API. (2021). Retrieved from https://developers.google.com/maps/documentation/places/web-service/overview

30. Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall/CRC.

31. Anselin, L. (1995). Local indicators of spatial association - LISA. Geographical Analysis, 27(2), 93-115.

32. Kulldorff, M. (1997). A spatial scan statistic. Communications in Statistics - Theory and Methods, 26(6), 1481-1496.

33. Stoffer, D. S. (1999). Detecting common signals in multiple time series using the spectral envelope. Journal of the American Statistical Association, 94(448), 1341-1356.

34. Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. Technometrics, 25(2), 165-172.

35. Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series. In Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (pp. 89-94).

36. Newman, M. E. J. (2010). Networks: An Introduction. Oxford University Press.

37. Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3-5), 75-174.

38. Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (3rd ed.). Wiley-Interscience.

39. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

40. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

41. Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer.

42. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579-2605.

43. Bostock, M., Ogievetsky, V., & Heer, J. (2011). D³: Data-driven documents. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2301-2309.

44. Plotly Technologies Inc. (2015). Collaborative data science. Plotly Technologies Inc. https://plot.ly

45. Tableau Software. (2021). Tableau. https://www.tableau.com/

46. Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. Ethics and Information Technology, 12(4), 313-325.

47. Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., ... & Pasquale, F. (2017). Ten simple rules for responsible big data research. PLOS Computational Biology, 13(3), e1005399.

48. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570.

49. Dwork, C. (2006). Differential privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (pp. 1-12).

50. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (pp. 1137-1143).

51. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

52. Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224-227.

53. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.

54. Nearest Neighbor Index (NNI)

55. Ripley's K-function

56. Getis-Ord Gi\*

57. Demographic data

58. Infrastructure data

59. Land use and land cover data

60. Time-related factors

61. Geographically Weighted Regression (GWR)

62. Leaflet library

63. Folium library

64. Andrienko, G., Andrienko, N. (2014). Visualization of spatial and spatio-temporal data. In: Spatial Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg, 111-134.

65. Oldenburg, R. (1999). The Great Good Place: Cafes, Coffee Shops, Bookstores, Bars, Hair Salons, and Other Hangouts at the Heart of a Community. Da Capo Press.

66. Lefebvre, H. (1991). The Production of Space. Blackwell.

67. Zerubavel, E. (1985). Hidden Rhythms: Schedules and Calendars in Social Life. University of California Press.

68. Giddens, A. (1984). The Constitution of Society: Outline of the Theory of Structuration. University of California Press.

69. Lynch, K. (1960). The Image of the City. MIT Press.

70. Bourdieu, P. (1984). Distinction: A Social Critique of the Judgement of Taste. Harvard University Press.

71. Goffman, E. (1959). The Presentation of Self in Everyday Life. Doubleday.

72. Cialdini, R. B. (2001). Influence: Science and Practice. Allyn & Bacon.

73. Van Dijk, J. A. G. M. (2006). Digital divide research, achievements and shortcomings. Poetics, 34(4-5), 221-235.

74. Quercia, D., Schifanella, R., & Aiello, L. M. (2014). The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In Proceedings of the 25th ACM Conference on Hypertext and Social Media (pp. 116-125).

75. Boy, J. D., & Uitermark, J. (2016). How to study the city on Instagram. PLOS ONE, 11(6), e0158161.

76. Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (pp. 505-514).

77. boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662-679.