

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего
образования
«Волгоградский государственный технический университет»

На правах рукописи



Зеленский Илья Сергеевич

**СОВЕРШЕНСТВОВАНИЕ МЕТОДОВ МОНИТОРИНГА СОСТОЯНИЯ
ГОРОДСКОЙ СРЕДЫ НА ОСНОВЕ ОТКРЫТЫХ ДАННЫХ**

2.3.1 – Системный анализ, управление и обработка информации, статистика
(технические науки)

ДИССЕРТАЦИЯ

на соискание ученой степени кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
Садовникова Наталья Петровна

Волгоград – 2026

Оглавление

Введение.....	5
ГЛАВА 1. Анализ работ в области мониторинга сложных систем	16
1.1 Структура инструментов мониторинга сложных систем	16
1.2 Ключевые особенности объекта мониторинга	21
1.3 Существующие решения и исследования в области мониторинга городской среды и их недостатки.....	24
1.4 Система мониторинга в составе национального проекта.....	31
1.5 Постановка задачи повышения эффективности мониторинга.....	35
1.6 Выводы по первой главе	41
ГЛАВА 2. Разработка метода извлечения и трансформации данных	43
2.1. Особенности первичных данных	44
2.2. Этапы метода извлечения и трансформации данных	48
2.2.1. Исследование доступных источников первичных данных	48
2.2.2. Сбор первичных данных	51
2.2.3. Трансформация текстовых первичных данных	51
2.2.4. Трансформация графических первичных данных.....	57
2.2.5. Построение структурированных описаний объектов	58
2.3. Выводы по второй главе.....	59
ГЛАВА 3. Разработка метода очистки и обогащения данных	61
3.1. Связывание данных	62
3.1.1 Обнаружение дублирования данных	63
3.1.2 Устранение дублирования данных.....	67
3.2. Прогнозирование достоверности данных	69

3.2.1	Прогнозирование принадлежности данных агентам	70
3.2.2	Выявление «сомнительных» данных	72
3.3.	Выводы по третьей главе	76
ГЛАВА 4. Разработка алгоритмов оценки состояния		78
4.1	Разработка алгоритма оценки показателя потребительских свойств....	79
4.1.1	Выбор метода оценки состояния.....	79
4.1.2	Постановка задачи оценки состояния.....	83
4.1.3	Разработка алгоритма вычисления интегральной оценки состояния....	86
4.1.4	Выбор метода определения весов частных критериев оценки состояния	88
4.1.5	Целевая функция оценки состояния	92
4.1.6	Алгоритм оценки показателя потребительских свойств жилых и коммерческих объектов.....	93
4.2	Разработка алгоритма оценки показателя обеспеченности социальной инфраструктурой.....	95
4.2.1	Оценка максимальной вместимости объектов.....	98
4.2.2	Оценка необходимой вместимости объектов	99
4.2.3	Определение перегруженных объектов.....	102
4.2.4	Оценка обеспеченности социальной инфраструктурой	102
4.2.5	Алгоритм оценки показателя обеспеченности социальной инфраструктурой.....	103
4.3	Выводы по четвертой главе	104
ГЛАВА 5. Разработка системы оценки и обоснование эффективности решений		105
5.1	Разработка архитектуры прототипа	105
5.2	Проектирование базы данных	107

5.3	Реализация модулей сбора первичных данных	107
5.4	Реализация модуля извлечения и трансформации данных	109
5.5	Реализация модуля хэширования изображений	115
5.6	Реализация модуля очистки и обогащения данных	116
5.7.1	Прогнозирование принадлежности данных агентам	116
5.7.2	Выявление «сомнительных» данных	120
5.7.3	Обнаружение дубликатов изображений	122
5.7	Разработка модуля оценки показателя потребительских свойств	124
5.8	Разработка модуля оценки показателя обеспеченности социальной инфраструктурой	126
5.9	Обоснование эффективности предложенных решений	128
5.10	Выводы по пятой главе	134
Заключение		136
Список используемых сокращений и условных обозначений		139
Список используемых источников		140
Приложение А – Схема базы данных разработанной системы		159
Приложение Б – Перечень характеристик в структурированных описаниях объектов городской среды		160
Приложение В – Конфигурации оценки потребительских		162
Приложение Г – Копии свидетельств о регистрации программ для ЭВМ		166
Приложение Д – Копия акта внедрения результатов работы		169

Введение

Актуальность темы исследования. Благополучие населения городов во многом зависит от качества и состояния городской среды, которая является сложной пространственно распределенной системой и характеризуется комплексной структурой с множеством разнородных подсистем и элементов, определяющих комфортность проживания на определенной территории. Характерными чертами современных городов являются сложное пространственное размещение людей и видов деятельности, неравномерность охвата инфраструктурой, что приводит к повышению сложности задач управления развитием городов [1].

Стремление к созданию качественной, гармонично развитой и устойчивой городской среды в РФ зафиксировано законодательством [2, 3], указами [4] и поручениями [5] Президента и Национальными проектами [6]. Перечисленные документы фиксируют на государственном уровне цели, созвучные с 11-й целью в списке целей устойчивого развития ООН [7] по созданию устойчивой городской среды.

Задача создания устойчивой и комфортной городской среды порождает задачи принятия управленческих решений по преобразованию городов. Для их решения необходимо анализировать множество факторов, влияющих на качество жизни в городах, и совершенствовать способы получения информации о них, поскольку эффективное управление сложными и динамичными пространственно распределенными системами, к числу которых принадлежат города, практически невозможно без непрерывного сбора и анализа информации об их актуальном состоянии.

Сегодня наиболее важным инструментом наблюдения, анализа и прогнозирования, а также фактором в принятии обоснованных и эффективных решений в области управления сложными динамичными системами, являются методы мониторинга показателей их состояния [8]. Таким показателем

применительно к городам РФ является степень благоприятности жизни, привлекательности для населения, которая зависит от многих факторов, в том числе и от состояния городской среды [9] как показателя способности удовлетворять объективные потребности населения.

Таким образом, возникает потребность в регулярном мониторинге состояния городской среды для своевременного получения полной информации о ней и принятия необходимых управленческих решений. Задача является актуальной для РФ и имеет ряд существующих методов решения, используемых, например, при подсчете Национального рейтинга качества жизни [10] или оценке достижения целей федеральных проектов «Развитие инфраструктуры в населенных пунктах» [11] и «Формирование комфортной городской среды [12] национального проекта «Инфраструктура для жизни» [6], утвержденных на государственном уровне. Для получения всесторонней оценки состояния городской среды данные решения используют официальную статистику [10, 13], данные опросов населения [10], а также данные из открытых источников: геоинформационных систем (ГИС) [10] и др. [13]. Однако в данных решениях отсутствует строгий регламент относительно актуальности данных и механизмов их сбора [10, 13], а интерпретация первичных данных производится с участием человека. С ростом количества источников повышается вероятность использования некорректных данных, ошибочных выводов и расчетов. Также существующие решения ориентированы на оценку состояния городов в целом, не позволяя, как следствие, своевременно обнаружить негативные тенденции, складывающиеся на отдельных территориях. Кроме того, существующие решения учитывают не все факторы, влияющие на удовлетворение потребностей населения.

В связи с этим повышение эффективности мониторинга состояния городской среды на основе обработки гетерогенных данных и расширения перечня используемых показателей состояния городской среды является актуальной задачей.

Степень разработанности темы. Автоматизированному мониторингу отдельно взятых элементов ГС посвящены работы Yao B., Ma A., Feng R., Shen X., Zhang M., Yao Ya., Bautista F., Bautista-Hernández D., Pacheco A., Goguitchaichvili A., Morales Ju., Gallegos-Tavera A., Nojeng S., Jaya A., Murniati R., Финогеева А.Г., Финогеева А.А., Ляпина А.М., Парыгина Д.С., Голубева А.В., Данилиной Н.В. и др. Подход к мониторингу ГС как единой системы на основе закрытой информации из государственных отчетов предлагают Osman M.A., Tahoon K.Q., Shehata M., Lamee M.A. Мониторингу социально-экономического развития региона на основе открытых данных посвящены работы Иващенко А.В.

Munywoki M.N., Singh K., Onwuzuligbo Ch., Ejikeme J., Kenechukwu A. и др. рассматривают применение для мониторинга ГС геоинформационных систем (ГИС). Обработке пространственных данных для мониторинга и поддержки принятия решений по управлению целыми регионами посвящены труды Ямашкина С.А., Ямашкина А.А. и Бершадского А.М.

Анализу состояния городов через призму рыночных цен объектов ГС посвящены работы Buda E., Broitman D., Czamanski D, Гришина Д.

Количественной оценке состояния ГС посвящена работа Трояновского В.С., сфокусированная на оценке обеспеченности социальной инфраструктурой на основе показателей, закрепленных в нормативных документах РФ.

Целью диссертационного исследования является повышение эффективности процессов мониторинга состояния городской среды за счет использования интеллектуальных технологий и пространственного анализа данных. Эффективность определяется информативностью и полнотой получаемой информации.

Для достижения цели были сформулированы и решены следующие **задачи**:

1. Провести анализ существующих работ в области мониторинга сложных систем, в частности городской среды и её подсистем.
2. Разработать метод автоматизированного извлечения и трансформации данных о состоянии городской среды из гетерогенных источников.

3. Разработать метод очистки и обогащения геопространственных и атрибутивных данных о состоянии городской среды.

4. Разработать алгоритмы оценки состояния сложной пространственно распределенной системы, адаптированные к задачам принятия решений по преобразованию городской среды.

5. Разработать систему оценки состояния сложной пространственно распределенной системы (городской среды) и обосновать эффективность предложенных решений.

Объектом исследования является система мониторинга состояния городской среды.

Предметом исследования являются методы сбора и анализа данных для мониторинга состояния городской среды.

Методы исследования. Для решения поставленных задач были использованы методы системного анализа, моделирования; методы искусственного интеллекта (в частности, обработки естественных языков); методы статистики и обработки геопространственных данных, а также методы планирования эксперимента и обработки экспериментальных данных.

Научная новизна заключается в разработке нового подхода к получению информации о состоянии городской среды, который включает в себя:

– Новый метод автоматизированного извлечения и трансформации данных о состоянии сложной пространственно распределенной системы (городской среды) из гетерогенных источников, отличающийся сочетанием технологий обработки графических данных, естественных языков и геопространственного анализа (п. 12 паспорта специальности 2.3.1).

– Метод очистки и обогащения геопространственных и атрибутивных данных, отличающийся сочетанием методов статистики, обработки графических данных, естественных языков и машинного обучения для прогнозирования достоверности данных и решения задачи связывания данных (п. 4 паспорта специальности 2.3.1).

– Новые алгоритмы оценки состояния сложной пространственно распределенной системы, отличающиеся адаптацией к задачам принятия решений по преобразованию городской среды (п. 5 паспорта специальности 2.3.1).

Теоретическая и практическая значимость.

Теоретическая значимость заключается в развитии методов сбора и обработки гетерогенных данных о состоянии сложной пространственно распределенной системы (городской среды).

Практическая значимость заключается в использовании следующих результатов исследования:

1. Разработанные алгоритмы оценки состояния сложной пространственно распределенной системы (городской среды) по ранее неучтенным показателям на основе данных из гетерогенных открытых источников позволяют улучшить полноту и информативность результатов мониторинга, а также могут быть использованы для управления развитием городов России с соблюдением целей устойчивого и комплексного развития городских территорий.

2. Разработанные методы автоматизированного извлечения, трансформации, очистки и обогащения данных о состоянии городской среды из гетерогенных источников позволяют получать актуальные данные для использования при анализе состояния городских территорий.

3. Внедрение разработанных методов и алгоритмов позволяет повысить эффективность и обоснованность принимаемых решений в сфере преобразования городских территорий и сократить затраты времени на анализ состояния городской среды, что подтверждается актом внедрения.

4. Результаты, полученные в диссертационном исследовании, целесообразно использовать в учебном процессе по направлению обучения «Информатика и вычислительная техника» в дисциплинах «Методы оптимизации», «Теория принятия решений» и «Математическое моделирование».

Предложенные методы и алгоритмы были использованы при разработке специализированного программного обеспечения, на которое получены свидетельства о регистрации программ для ЭВМ:

- «Программа генеративного проектирования комплексного развития городских территорий на основе знаний из онтологической модели», № 2024685595 от 30 октября 2024 г.

- «Программный комплекс для массовой оценки жилых помещений сравнительным подходом», № 2020666421 от 9 декабря 2020.

- «Программный модуль для сравнительного анализа и выявления дубликатов изображений», № 2018662442 от 8 октября 2018 г.

Результаты работы использованы при реализации 2 грантов: РФФИ № 18-37-20066-мол_a_вед, РНФ и Администрации Волгоградской области № 22-11-20024.

Достоверность и апробация результатов работы.

Достоверность изложенных в работе результатов исследования обусловлена корректным применением указанных методов исследования и успешным практическим применением результатов диссертационной работы, что отражено в акте внедрения.

Реализация результатов работы состоялась при исполнении следующих НИР:

- «Разработка фундаментальных основ для информационно-аналитической поддержки задач комплексного развития городских территорий с использованием методов онтологического инжиниринга», 2022–2024 гг., грант Российского научного фонда (РНФ) и Администрации Волгоградской области 22-11-20024;

- «Разработка научно-методического подхода к поддержке процесса управления развитием урбанизированной территории на основе интеллектуальной обработки геопространственных данных», 2018–2020 гг., грант РФФИ 18-37-20066-мол_a_вед.

Основные положения и материалы диссертационной работы представлялись и обсуждались на международных и всероссийских научных и практических мероприятиях: Основные положения и материалы диссертационной работы представлялись и обсуждались на международных и всероссийских научных и практических мероприятиях:

- XIV Всероссийское совещание по проблемам управления (Москва, 2024);
- Международная Казанская научная мультikonференция ММТТ-37 (Казань, 2024);
- 5th International Conference on Creativity in Intelligent Technologies and Data Science (CIT&DS 2023) (Волгоград, 2023);
- XXV Международная объединенная научная конференция «Интернет и современное общество» (IMS-2022) (Санкт-Петербург, 2022);
- International Conference on Information Technology in Business and Industry (ITBI 2020) (Новосибирск, 2020);
- XIII Всероссийская научно-практическая студенческая конференция «России – творческую молодёжь» (Камышин, 2020);
- Конкурс научно-исследовательских работ студентов Волгоградского государственного технического университета (Волгоград, 2020, 2021);
- Electronic Governance and Open Society: Challenges in Eurasia : 6th International Conference EGOSE 2019 (Санкт-Петербург, 2019);
- XIV Всероссийская заочная научно-практическая конференция «Инновационные технологии в обучении и производстве» (Камышин, 2019);
- VI международной конференции «Информационные технологии в науке, управлении, социальной сфере и медицине» (Томск, 2019);
- Смотр-конкурс научных, конструкторских и технологических работ студентов Волгоградского государственного технического университета (Волгоград, 2018, 2019);

– 7th International Conference on System Modeling & Advancement in Research Trends (SMART–2018, IEEE Conference ID: 44078) 2018 (Морадабад, 2018);

– XXII Региональная конференция молодых учёных Волгоградской области 2017 (Волгоград, 2017).

Реализация и внедрение результатов работы.

Результаты работы успешно внедрены в ООО «СтройБилд» с целью поддержки принятия решений по выбору места строительства и конфигурации новых жилых и нежилых зданий, что позволило повысить эффективность и обоснованность принимаемых решений и сократить затраты времени на анализ состояния городской среды.

Положения, выносимые на защиту:

1. Метод автоматизированного извлечения и трансформации данных о состоянии сложной пространственно распределенной системы (городской среды) из гетерогенных источников, обеспечивающий получение актуальных данных для анализа при мониторинге.

2. Метод очистки и обогащения геопространственных и атрибутивных данных, позволяющий повысить качество результатов мониторинга.

3. Алгоритмы оценки состояния сложной пространственно распределенной системы (городской среды), позволяющие повысить полноту и информативность результатов мониторинга.

4. Разработанная система оценки состояния сложной пространственно распределенной системы (городской среды) по ранее неучтенным показателям.

Публикации. Материалы диссертации опубликованы в 28 печатных работах, из них 10 статей – в рецензируемых журналах, рекомендуемых ВАК, 5 публикаций – в изданиях, включенных в международные базы данных Scopus и Web of Science.

Личный вклад автора. В настоящей диссертации представлены результаты исследований, выполненных самим автором или под его

непосредственным руководством. Личный вклад автора состоит в постановке задач исследования, разработке теоретических и прикладных методов их решения, составляющих научную новизну, в обработке, анализе, обобщении полученных результатов и формулировке выводов. В публикациях с соавторами авторский вклад распределяется пропорционально. Программная реализация методов и алгоритмов выполнена лично автором.

Структура и объем диссертации. Диссертационная работа состоит из введения, пяти глав, заключения, списка литературы (144 наименования на 19 страницах) и пяти приложений. Основной текст работы (без приложений и списка литературы) занимает 139 страниц машинописного текста, в том числе 47 рисунков, 21 таблица. Общий объем работы – 169 страниц.

Во **введении** обоснована актуальность темы, сформулирована цель и задачи исследования, выбраны методы исследования, отмечены научная новизна и практическая значимость работы, приведены основные положения, выносимые на защиту.

В **первой главе** проведен анализ существующих работ в области мониторинга состояния сложных систем, в том числе городской среды как целого и отдельно взятых её подсистем, а также используемых в этих исследованиях источников данных и методов их обработки и анализа. Определены недостатки существующих методов и систем мониторинга состояния городской среды с точки зрения качества данных и полноты покрытия объекта мониторинга. Выделена система мониторинга на основе показателей федерального проекта «Развитие инфраструктуры в населенных пунктах» как наиболее актуальное в настоящее время для РФ решение задачи. Отмечены значимые показатели состояния городской среды, недостаточно покрытые существующими решениями в области мониторинга: потребительские свойства жилых и коммерческих объектов городской среды и обеспеченность социальной инфраструктурой (учреждениями образования и здравоохранения). Составлена обобщенная схема архитектуры системы мониторинга комплексных объектов и рассмотрены принципы оценки эффективности таких систем. Предложена модификация

системы мониторинга на основе показателей федерального проекта, дополненная оценкой неучтенных ранее значимых показателей состояния городской среды. Целью предлагаемых изменений является повышение эффективности мониторинга состояния городской среды посредством учета дополнительных значимых показателей состояния и применения автоматизированного интеллектуального анализа гетерогенных открытых данных, ранее в процессе мониторинга не использованных. Определены критерии эффективности мониторинга для повышения. Сформулированы задачи исследования, решение которых необходимо для реализации предложенной концепции системы оценки состояния городской среды.

Во **второй главе** предложен метод извлечения и трансформации данных о состоянии городской среды из гетерогенных открытых источников, отличающийся сочетанием технологий обработки графических данных, естественных языков и геопространственного анализа. Исследованы особенности данных об объектах городской среды, публикуемых в рассматриваемых открытых источниках: ГИС, социальных сетях, тематических площадках для размещения объявлений об операциях с недвижимым имуществом.

В **третьей главе** описан метод очистки и обогащения данных о состоянии городской среды, осуществляющий связывание данных, а также прогнозирование их достоверности с использованием методов машинного обучения, векторного представления слов, методов работы с графической информацией и статистики.

В **четвертой главе** описаны алгоритмы оценки состояния сложной пространственно распределенной системы (городской среды) по выделенным ранее неучтенным показателям: потребительским свойствам жилых и коммерческих объектов городской среды и обеспеченности социальной инфраструктурой (образованием и здравоохранением).

В **пятой главе** описана программная реализация предложенных методов и алгоритмов в виде прототипа системы оценки состояния сложной пространственно распределенной системы (городской среды) по ранее неучтенным показателям, включающая реализацию методов, описанных в

предыдущих главах. Приведено обоснование эффективности предложенных решений.

В **заключении** приводятся основные научные и прикладные результаты, полученные в процессе выполнения диссертационной работы, и выделяются возможные направления дальнейших исследований.

В **приложениях** приведены: схема базы данных разработанной системы, перечень характеристик в структурированных описаниях жилых и коммерческих объектов городской среды, конфигурации оценки состояния городской среды по критерию потребительских свойств для разработанного прототипа системы, копии свидетельств о государственной регистрации программ для ЭВМ и акта внедрения результатов научной работы.

Благодарности. Автор выражает благодарность заведующему кафедрой «Цифровые технологии в урбанистике, архитектуре и строительстве» ВолгГТУ к.т.н. Д.С. Парыгину за помощь при написании и оформлении диссертации, а также участие в обсуждении полученных результатов.

ГЛАВА 1. Анализ работ в области мониторинга сложных систем

1.1 Структура инструментов мониторинга сложных систем

Системный подход предполагает рассмотрение мониторинга как целостного комплекса связанных элементов, реализующих общие функции. На текущий момент не существует единой структуры инструментов мониторинга сложных систем, конкретный перечень элементов определяется прикладной задачей мониторинга и системой, для мониторинга которой инструмент предназначен.

Реализации систем мониторинга и оценке их эффективности посвящены работы П. В. Дмитриенко и В. В. Коренькова [14-15], рассматривающих данный вопрос применительно к наблюдению за функционированием комплексных систем и контролю их ресурсов. Авторами описана общая архитектура систем мониторинга, приведено формальное описание модели взаимодействия системы мониторинга с подконтрольной системой [14] и предложены метрики эффективности систем мониторинга [15]. Перечисленные труды позволяют сформировать представление об общих принципах устройства систем мониторинга, их архитектуре и ключевых подсистемах.

С точки зрения системного анализа, в задаче мониторинга выделяют две взаимосвязанные системы: подконтрольную, для эффективного управления которой применяются средства мониторинга – объект мониторинга, и комплекс данных средств – систему мониторинга [14-15].

Состояние объекта мониторинга описывается набором показателей, за которыми осуществляется регулярное наблюдение для анализа протекающих процессов и своевременного выявления нештатных состояний элементов [15]. Нештатным или кризисным состоянием объекта мониторинга или его элемента называют состояние, которое препятствует его корректной работе. Под нештатной ситуацией следует понимать нахождение одного или более элементов объекта мониторинга в нештатном состоянии.

В обобщенном виде система мониторинга описывается следующим кортежем [8]:

$$MS = \langle S, M, T, G, O, P \rangle, \quad (1.1)$$

где S – предмет мониторинга как совокупность процессов и явлений, а также состояние объекта, его функционирование и развитие в течение определенного периода времени;

M – методы мониторинга: опрос, анализ документов, наблюдение, эксперимент, анкетирование, интервьюирование и др.;

T – задачи мониторинга: сбор, обработка и исследование информации;

G – цель мониторинга: организация эффективного функционирования и развития управляемого объекта и его элементов;

O – объект мониторинга: подконтрольная система или её элемент, на который направлены конкретные мониторинговые процедуры;

P – субъекты мониторинга: носители мониторинговых функций объекта.

Помимо получения сведений о текущем состоянии и функционировании объекта мониторинга, а также обеспечения своевременной реакции на нештатные ситуации, результаты мониторинга используются для изучения закономерностей в повторяющихся нештатных ситуациях, связей между элементами объекта с целью повышения эффективности их работы, а также для прогнозирования поведения объекта при различных изменениях и внешних воздействиях [14-15].

Общая концептуальная схема системы мониторинга приведена на Рисунке 1.1 [8, 14-15]. Сплошные линии со стрелками обозначают передачу информации между элементами системы мониторинга. Пунктирные линии со стрелками отмечают потоки информации, направленные извне системы (от объекта мониторинга) и за её рамки к лицу, принимающему решения (ЛПР) об оказании управляющих воздействий на подконтрольную систему.

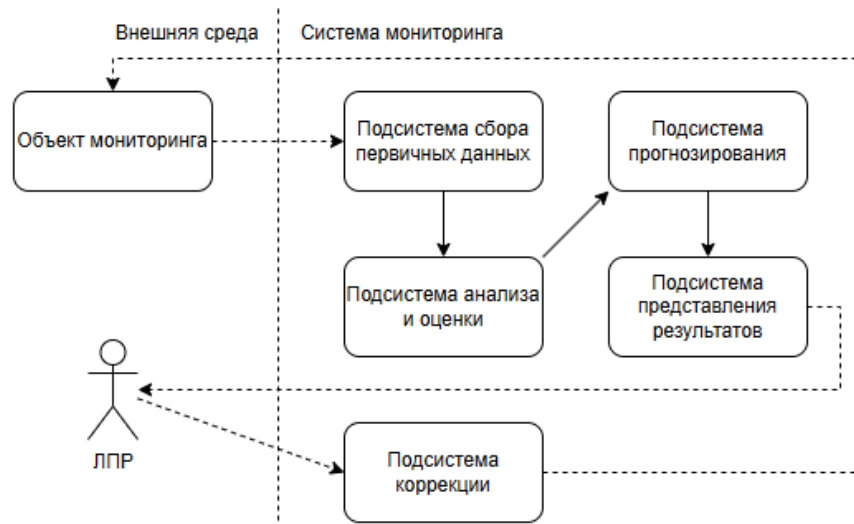


Рисунок 1.1 – Общая концептуальная схема системы мониторинга

На Рисунке 1.1:

- подсистема сбора первичных данных – совокупность датчиков, получающих информацию о первичных показателях состояния объекта мониторинга, и подсистем первичной обработки поступающей информации;
- подсистема анализа и оценки преобразовывает первичные показатели в оценки целевых показателей состояния системы и анализирует их с точки зрения факта наступления нештатных ситуаций [14];
- подсистема прогнозирования отвечает за прогноз дальнейшего поведения объекта мониторинга и потенциального наступления нештатного состояния на основе ранее выявленных закономерностей;
- подсистема представления результатов осуществляет передачу полученной информации ЛПР и представление её в наглядном виде [15];
- подсистема коррекции позволяет ЛПР осуществлять управляющее воздействие на объект мониторинга с целью исключения наступления нештатных состояний или вывода из нештатного состояния [15].

Теоретико-множественное описание системы мониторинга имеет следующий вид [15]:

$$MS = \langle O, P, f_p, D_o, D_p, S, f_u \rangle, \quad (1.2)$$

где O – множество элементов объекта мониторинга;

P – множество свойств, которыми обладают элементы объекта мониторинга;

f_p – функция принадлежности свойства некоторому элементу объекта, имеющая следующий вид $f_p = O \otimes P \rightarrow \{0,1\}$ [15];

D_o – множество упорядоченных пар элементов, обозначающих наличие зависимости стабильного состояния второго элемента пары от стабильного состояния первого;

D_p – множество упорядоченных четверок $\langle o_1, p_1, o_2, p_2 \rangle$, обозначающих наличие зависимости стабильного состояния свойства p_2 объекта o_2 от стабильного состояния свойства p_1 объекта o_1 ;

S – множество датчиков;

f_u – функция применимости конкретного датчика для контроля состояния конкретного свойства, имеющая вид $f_u = S \otimes P \rightarrow \{0,1\}$ [15];

Таким образом, мера определенности состояния подконтрольной системы описывается следующей формулой [15]:

$$D = \frac{\sum_{i=1}^c w_{p_i} n_{p_i}}{\sum_{j=1}^d w_{p_j} n_{p_j}}, \quad (1.3)$$

где p_i – свойства объекта мониторинга, охваченные системой мониторинга;

p_j – все свойства элементов объекта мониторинга;

w_{p_i} – вклад свойства p_i в функционирование объекта мониторинга;

n_{p_i} – число элементов объекта мониторинга, обладающих свойством p_i .

D не может быть больше 1 и принимает это значение только в том случае, если для каждого из свойств, определяющих состояние подконтрольной системы, в системе мониторинга существует хотя бы один метод мониторинга, применимый для контроля этого свойства [15]. На практике же системы мониторинга, осуществляющие полный охват свойств объекта, различают по полноте предоставляемой ими информации [15]. В реальных системах мониторинга различные диагностические сообщения обладают разной информативностью, позволяя с большей или меньшей точностью локализовать

конкретный элемент и/или его свойство, служащие источником нештатной ситуации. Сообщения, а значит, и порождающие их датчики могут предоставлять разное количество дополнительных данных о состоянии подконтрольной системы, важных для принятия решения об управляющем воздействии.

Таким образом, эффективность мониторинга состояния подконтрольной системы с точки зрения точности результатов, их достоверности и возможности их использования среди прочих определяются следующими факторами [15]:

- полнотой информации о состоянии объекта мониторинга или полнотой охвата процессом мониторинга её составляющих: в идеальном случае в перечне первичных показателей должно быть учтено состояние всех подсистем в полном объёме;
- актуальностью и достоверностью используемой первичной информации: источники данных должны быть надежны, данные должны подвергаться строгому контролю качества и обновляться не реже, чем вступают в силу изменения в объекте мониторинга;
- объективностью первичной информации: для оценки состояния объекта необходимо использовать исходные данные напрямую от датчиков, а получения данных через посредников – избегать, чтобы не получать промежуточную интерпретацию данных третьей стороной;
- степенью информативности результатов: обобщение, усреднение оценок состояния по большому количеству подсистем может приводить к искажению результатов и потере локальных девиаций, видимых по первичным показателям, из которых можно выделить конкретный элемент и/или его свойство, приводящие к нештатным ситуациям;
- возможностью прогнозирования последствий изменений, что позволяет избежать решений, приводящих к дестабилизации системы.

1.2 Ключевые особенности объекта мониторинга

Городская среда характеризуется совокупностью природных, архитектурно-планировочных, экологических и других факторов, формирующих среду жизнедеятельности на определенной территории города и определяющих комфортность проживания на ней [16]. На основе существующих трактовки понятия городской среды [16-19] было составлено теоретико-множественное описание городской среды как системы:

$$BE = \langle C, L, T, G, F, P(C, F) \rangle, \quad (1.4)$$

где C – множество элементов (подсистем) городской среды;

L – множество связей между элементами, отражающих их взаимное влияние друг на друга;

T – территория, на которой расположена городская среда;

G – цель городской среды как системы;

F – функции городской среды как системы;

$P(C, F) \in [0; 1]$ – показатель участия подсистемы из множества C в реализации функций системы из множества F .

Целью G является создание комфортной и безопасной среды для пребывания и удовлетворения потребностей различных групп населения [20].

Система реализует функции:

1. Социальную: создание условий для восстановления и развития человеческого капитала, здравоохранения и образования [21-22].
2. Транспортно-промышленную: обеспечение транспортной доступности для людей и грузов, движения в городские районы и из них [22].
3. Экономическую: сосредоточение различных предприятий на ограниченной территории с целью сократить издержки производства [22].
4. Обеспечения жилья, что в комбинации с дополнительными элементами комфорта или дискомфорта формирует качество жилой среды [23].

5. Предоставления пространства для коммерческой деятельности: наличие развитой, комфортной городской среды приводит к появлению новых рабочих мест и увеличению налоговых поступлений [24].

Хотя единой общепринятой структурной схемы городской среды как сложной пространственно распределенной системы не существует, работа [25] предлагает схему элементов, представленную на Рисунке 1.2 в расширенном виде с раскрытием элементов более низкого уровня (ключевые для исследования элементов выделены цветом).

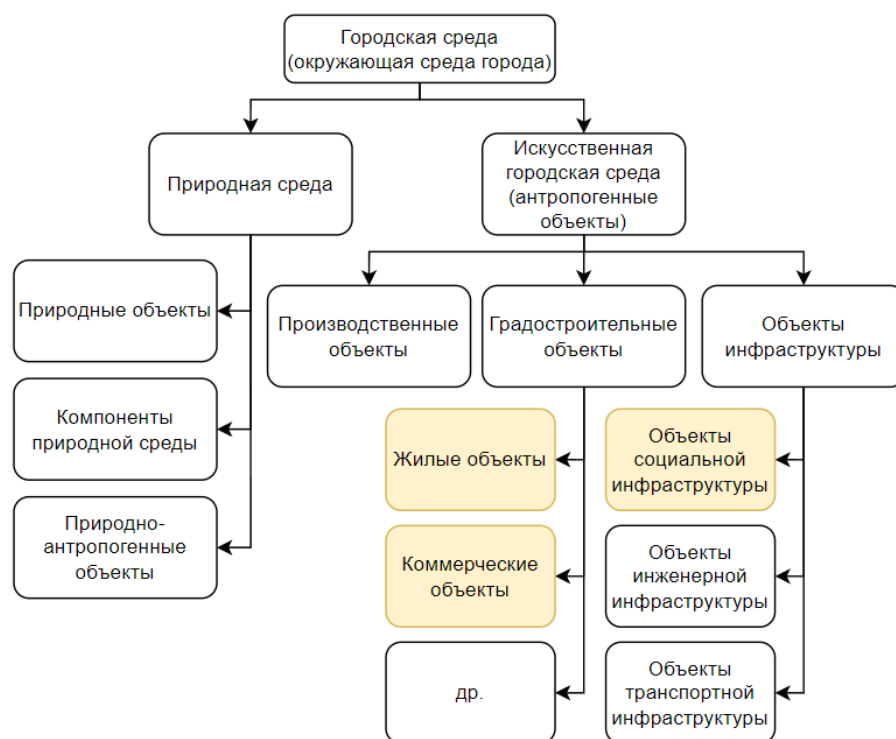


Рисунок 1.2 – Элементы городской среды как сложной пространственно распределенной системы

Опираясь на описания состава элементов городской среды [13, 16, 17, 19, 20, 25-26], множество элементов (подсистем) C было детализировано следующим образом:

$$C = \langle Arch, Soc, Nat \rangle, \quad (1.5)$$

где $Arch$ – архитектурно-планировочные (антропогенные) элементы;

Soc – социальные элементы;

Nat – природные элементы.

Множество архитектурно-планировочных элементов *Arch*, в свою очередь, описывается как [16, 20]:

$$Arch = \langle LRe, CRe, S, I, Com, Tr \rangle, \quad (1.6)$$

где *LRe* – жилье (жилые объекты) и прилегающие пространства;

CRe – коммерческие объекты;

S – общегородские пространства;

I – объекты инфраструктуры, обеспечивающие потребности населения (социально-досуговой, общественно-деловой и пр.);

Com – коммуникации (улично-дорожная сеть, канализация, линии электропередач, линии подключения к сети Интернет и пр.);

Tr – транспорт.

Были выделены следующие особенности городской среды как объекта мониторинга [13, 16, 20, 27-30]:

- Многообразие и взаимосвязанность элементов, что затрудняет выбор ключевых показателей состояния и может требовать применения принципиально разных подходов для оценки различных показателей.
- Отсутствие единых стандартов оценки, что может приводить к различным интерпретациям одних и тех же первичных данных.
- Ограниченная доступность данных: не все данные имеют открытый доступ, часть невозможно получить без физического посещения городских территорий.
- Динамичность: изменения в городской среде могут опережать процесс сбора и анализа данных.
- Территориальная неоднородность: характеристики городской среды могут значительно различаться на разных городских территориях.

– Разнообразие источников и большие объемы данных: города порождают массивы данных, которые сложно обрабатывать без использования современных технологий.

1.3 Существующие решения и исследования в области мониторинга городской среды и их недостатки

Теме автоматизированного мониторинга состояния различных элементов городской среды как сложной пространственно распределенной системы посвящены работы Yao B., Ma A., Feng R., Shen X., Zhang M., Yao Ya. [31], Bautista F., Bautista-Hernández D., Pacheco A., Goguitchaichvili A., Morales Ju., Gallegos-Tavera A. [32], Roopa Lakshmi R. [33], Nojeng S., Jaya A., Murniati R. [34], Финогеева А.Г., Финогеева А.А., Ляпина А.М., Парыгина Д.С., Голубева А.В. [35] и др. Большинство трудов сосредоточено на мониторинге отдельно взятых подсистем городской среды без их связи с остальными. Например, мониторинг загрязнения воздуха [31-32] или мониторинг дорожного трафика и распознавание объектов и явлений (заторов, аварий) [33] в нём, обнаружение перегрузок городской электросети [34] и пр. Мониторинг городской среды как целого в рамках перечисленных исследований не рассматривается. Сбор первичных данных реализуется с помощью внедрения технических устройств-датчиков: автоматических измерителей качества воздуха, камер видеонаблюдения и т.п. Оценка, анализ и прогнозирование основаны на методах машинного обучения и искусственного интеллекта, выбранных в соответствии со спецификой конкретной задачи мониторинга и объекта. Оценке отдельных аспектов развития городской среды посвящены также научные труды Беляевой Е.Л. в части социальной оценки проектов благоустройства городских территорий [36], Акопова А.С., Бекларян А.Л., Сагателян А.К., Саакян Л.В., Беляева О.А., Тепаносян Г.О. в контексте поддержки принятия решения по рациональному

озеленению городов [37], Ложкиной О.В., Рогозинского Г.Г., Ложкина В.Н., Малыгина И.Г., Комашинского В.И. в части применения интеллектуальных технологий при принятии решений в сфере экологической безопасности городов [38], Sztubecka M., Skiba M, Mrówczyńska M., Bazan-Krzywoszanska A. в сфере многокритериальной оценки энергоэффективности городских зданий с применением ГИС [39], Н.В. Данилиной в части оценки состояния улиц [40], доступности общественного транспорта [41] и др. Упомянутые труды рассматривают один либо несколько выделенных аспектов жизнедеятельности городов и не позволяют получить комплексную оценку состояния системы в целом, учитывающую влияние смежных подсистем друг на друга.

Таким образом, для мониторинга различных подсистем городской среды применяются принципиально разные методы, и для мониторинга состояния городской среды как целого необходимо применение ансамбля методов в соответствии со спецификой включаемых в мониторинг подсистем.

Более комплексным подходом к мониторингу городов как единых систем характеризуются исследования Osman M.A., Tahoona K.Q., Shehata M., Lamee M.A., в частности, работа [42], посвященная пространственному мониторингу и классификации угасания городов Египта с использованием ГИС. Важным аспектом данного исследования является привязка к специфичным для конкретной страны нормативным документам, а также использование для большей части оценок закрытой информации из правительственных отчетов, сформированных за прошедшие периоды [42]. В частности, состояние конкретных объектов городской застройки оценивается на основе данных, собранных за год до проведения исследования, что не позволяет гарантировать актуальность результатов на момент его завершения. Следует также отметить использование в исследовании [42] метода парных сравнений Саати [43] для определения значимости каждого показателя (частного критерия оценки) состояния городской среды.

Применение ГИС в контексте мониторинга городских систем рассматривают и другие исследования. Так, мониторингу развития

инфраструктуры городов на базе ГИС с применением сравнительного анализа для оценки целевых показателей посвящена работа [44] ученых Munywoki M.N. и Singh K. Исследование [45] за авторством Onwuzuligbo Ch., Ejikeme J., Kenechukwu A. описывает применение ГИС для мониторинга загрязнения воздуха на основе корреляции между показателями атмосферного отражения и качества воздуха. В работе [46] исследователей Olorunfemi I.E., Fasinmirin J.T., Olufayo A.A., Komolafe A.A. ГИС применены для анализа изменения землепользования в контексте устойчивости городской среды. В основе исследования лежат методы классификации картографических изображений для измерения их подобия, а также изучения динамики озеленения и температуры для оценки влияния антропогенных факторов на устойчивость среды. Onwuzuligbo Ch., Okeke U.Ch., Udochukwu O.E., Effiom I.U., Kikroye J.U. в своем исследовании [47] применили ГИС и удаленные сенсоры для мониторинга и моделирования разрастания городов. Роль ГИС и их применение в градостроительной деятельности рассматривают в своих исследованиях также Аль Савафи М.Х. [48] и Каргашина М.А. [49]. Однако упомянутые работы сосредоточены в основном на теоретических изысканиях и концептуальных схемах систем поддержки принятия и не касаются практической реализации и апробации предлагаемых методов.

Из вышеперечисленных трудов [42, 44-49] можно сделать вывод, что ГИС являются источником актуальных пространственных данных о городской среде, взаимном расположении входящих в неё объектов, транспортной доступности, что подтверждается их использованием в составе существующих актуальных решений [10], но ГИС не несут актуальных данных о детальном состоянии объектов.

Теме консолидации, обработки и интерпретации разнородных пространственных данных для поддержки принятия решений по управлению территориально распределенными организационными системами, ориентированных на обеспечение условий устойчивого развития на уровне целых регионов, т.е. минимизацию рисков природного, техногенного и гуманитарного характера и максимизацию эффективности использования пространственно-

распределенных ресурсов, посвящены исследования Ямашкина С.А. [50-52], Ямашкина А.А. [51] и Бершадского А.М. [52].

Вопросам мониторинга социально-экономического развития региона на основе открытых данных с помощью цифровой платформы интегрального мониторинга посвящены работы Иващенко А.В., Додоновой Е.А и Дубининой И.Н. [53, 54].

Также следует отметить исследования, рассматривающие анализ состояния и привлекательности городов через призму рыночных цен объектов городской среды как недвижимого имущества. Следует выделить работы Buda E., Broitman D., Czamanski D. [1] и Гришина Д. [55], в которых авторы оценивают стоимость земли на основе цен на жилье, применяя специальные гедонистические модели к большим данным о сделках с жилыми объектами на детальном пространственном уровне. Полученные оценки служат маркером привлекательности территории, а также мерой разрыва между центральными и периферийными территориями городов и динамики такого разрыва.

Наконец, нельзя не упомянуть труды Лейфера Л.А. [56], а также работы Астафьева С.А. и Guo Yi [57], также связанные с рыночной оценкой объектов городской среды и принятием связанных решений. В данных трудах выявлены значимые закономерности влияния отдельных характеристик объектов на их потребительские свойства, привлекательность и, как следствие, цены.

Применяется также мониторинг качества жизни населения в городах, основанный на методах опросов жителей, геоаналитики и обработки статистических данных. Примером такого решения является Национальный рейтинг качества жизни [10], однако оно имеет ряд проблем с точки зрения актуальности, достоверности информации, а также эффективности её получения и обработки, что является направлением для совершенствования.

Актуальным зарубежным решением задачи мониторинга городов с точки зрения качества жизни является ежегодный публичный рейтинг EIU Global Liveability Index (GLI) [58]. Каждому городу GLI присваивает оценку по более чем 30 показателей, объединенным в пять обширных групп: стабильность,

здравоохранение, культура и окружающая среда, образование и инфраструктура [59]. Для категориальных показателей оценка присуждается на основе суждений экспертов-аналитиков, проживающих на территории соответствующей страны, и полевых корреспондентов в городе. Для количественных показателей рейтинг рассчитывается с использованием внешних источников данных [59], в качестве которых выступают, среди прочих, Всемирный банк. Детально методика расчета рейтинга не раскрывается, однако по опубликованной информации можно сделать вывод о сильной зависимости результатов мониторинга от частных мнений экспертов и корреспондентов, что негативно влияет на эффективность процесса мониторинга с точки зрения затрат временных, человеческих и финансовых ресурсов, и потенциально приводит к субъективности и недостоверности получаемых результатов мониторинга. Данный метод сложно масштабировать на новые города из-за необходимости поиска подходящих экспертов и корреспондентов на местах. Также к минусам следует отнести обновление раз в год и зависимость от внешних статистических отчетов (например, Всемирного банка), что может негативно сказываться на актуальности результатов.

Подобные проблемы есть и в решениях из рейтингов Monocle's Quality of Life Survey [60], Mercer's Quality of Living Ranking [61] и Movingto GLI [62].

Решение Numbeo's Quality of Life Ranking [63] кардинально отличается от описанных выше. В основе лежит глобальная база данных качества жизни Numbeo, наполняемая за счет конечных пользователей по принципу краудсорсинга. Преимуществом по сравнению с вышеописанными методами является более частое обновление (дважды в год) и использование данных напрямую от пользователей, что должно позитивно влиять на актуальность результатов. К недостаткам следует отнести зависимость актуальности результатов от энтузиазма пользователей-участников вследствие отсутствия регламента обновления данных, что может нивелировать рост актуальности, упомянутый выше. Также сомнительным является подход к валидации данных:

Numbeo «использует мудрость толпы для получения максимально надежных данных» [63], что не дает гарантий достоверности результатов.

Определению количественных показателей состояния городской среды посвящена работа Трояновского В.С. [64], сфокусированная на количественной оценке обеспеченности населения объектами инфраструктуры на основе обязательных к соблюдению показателей, закрепленных в нормативных документах РФ [65, 66]. Данная работа предлагает алгоритм оценки текущей загруженности объектов инфраструктуры на основе их максимальной вместимости, нормативной зоны обслуживания и данных о числе потребителей услуг в зоне обслуживания с учетом возможных пересечений с зонами других объектов. Однако автор не конкретизирует источники данных о числе потребителей услуг, а в качестве источников информации о самих объектах использует значительный объем закрытых данных [64], что может негативно сказываться на актуальности и достоверности получаемых оценок.

До 2026 года наиболее актуальным для РФ решением в области мониторинга состояния городской среды была система мониторинга на основе Индекса качества городской среды, в основе которого лежала оценка состояния 6 типов городских пространств по 6 критериям каждый, что в сумме давало 36 критериев оценки состояния городской среды. Проблемами решения были недостаточные актуальность исходных данных и результатов вследствие низкой частоты обновления данных, сбора первичных показателей через посредника (органы исполнительной власти) без четкого регламента их актуальности, а также полнота и информативность результатов, что было связано с расчетом индекса в целом по городу без возможности оперативно локализовать конкретный элемент городской среды, требующий корректирующего воздействия [16].

Основные группы существующих решений в области мониторинга состояния городской среды как сложной системы, их авторы и ключевые недостатки приведены в Таблице 1.1.

Таблица 1.1 – Основные группы решений в области мониторинга состояния городской среды

Группа решений	Авторы	Недостатки в контексте исследования
Мониторинг с помощью внедрения специализированных устройств-датчиков	Baozhen Yao, Ankun Ma, Rui Feng, Xiaopeng Shen, Mingheng Zhang, Yansheng Yao, Francisco Bautista, Dorian Bautista-Hernández, Aristeo Pacheco, Avto Goguitchaichvili	Мониторинг только отдельно взятых элементов городской среды, без связи с другими. Траты на внедрение датчиков. Рост расходов на внедрение и сопровождение датчиков с увеличением сложности системы.
Мониторинг с помощью опросов населения и/или экспертов	Веселова В.Н., Валеева О.В., Корытный Л.М., Тобиен М.А., Ковайкина Е.С., Лебедев В.В., Лебедев К.В.	Проведение качественных опросов ресурсозатратно. Результаты потенциально субъективны, неполны и/или недостоверны.
Мониторинг по данным официальной статистики	Osman M.A., Tahoona K.Q., Shehata M., Lamee M.A., Трояновский В.С.	Полученные результаты актуальны на дату сбора статистики, но не дают представления о текущем состоянии объекта мониторинга.
Мониторинг на основе рыночных цен недвижимого имущества	Buda E., Broitman D., Czamanski D.	Оценивают улучшение состояния по росту цен, что не всегда достоверно.
Мониторинг на основе пространственных данных	Munywoki M.N., Singh K. Onwuzuligbo Ch., Okeke U.Ch., Udochukwu O.E., Effiom I.U., Kikpoye J.U., Аль Савафи М.Х., Ямашкин С.А., Ямашкин А.А., Бершадский А.М.	Пространственные данные сами по себе не дают достаточно информации для полного покрытия городской среды мониторингом, т.к. не несут детальной атрибутивной информации о её элементах.
Комплексный мониторинг на основе множества показателей	Борисов М.А., Букалова С.В., Корнеичев А.Ю., Шевченко В.А., Погребная О.В., Che L., Yin SH., Jin JU., Wu W., Saeed U., Ahmad S.R., Mohey-ud-din G., Butt H.JA., Ashraf U.	Обычно реализуют сбор и обработку данных силами человека, например, экспертов, что приводит к возможной субъективности, неполноте, недостоверности и неточности результатов.

При изучении анализа неструктурированных открытых данных об объектах городской среды были рассмотрены труды по обработке естественных языков (NLP). Следует выделить работы F. Peng, A. McCallum [67], H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan [68], S. Bird, E. Klein, E. Loper [69] и M. Tomita [70]. Данные труды легли в основу большей части современных решений NLP, от Apache OpenNLP library [71] и до Томита-парсера [72] компании Яндекс.

1.4 Система мониторинга в составе национального проекта

На текущий момент наиболее актуальным для РФ подходом к мониторингу состояния городской среды является его ежемесячная многокритериальная оценка при расчете показателей федеральных проектов в составе национального проекта «Инфраструктура для жизни» [6], целью которого является обеспечение граждан инфраструктурой (жилищной, транспортной, социальной, коммунальной) нового качества [6]. Так, в состав федерального проекта «Развитие инфраструктуры в населенных пунктах» входит показатель «Улучшение качества среды для жизни в опорных населенных пунктах», состоящий из 16 компонентов (далее – показатели федерального проекта), отражающих состояние значимых элементов городской среды, в том числе обеспеченность жильем, благоустроенность общественных территорий, доступ к услугам и инфраструктуре и т.п. [13]. Источником данных для расчета компонентов показателя являются данные федерального статистического наблюдения, внесенные в автоматизированную информационную систему «Фонда развития территорий» (далее – АИС ФРТ) [13].

Применительно к задаче исследования и в контексте существующей системы мониторинга элементы данной системы (см. Рисунок 1.1) следует интерпретировать следующим образом:

- объект мониторинга – городская среда как совокупность множества гетерогенных элементов, формирующих среду жизнедеятельности города на

определенной территории и определяющих комфортность проживания на этой территории [13, 14];

- подсистема сбора первичных данных включает:
 - датчики, функции которых выполняют источники данных [13];
 - механизмы сбора и обработки первичных данных, чьи функции выполняют соответствующие органы исполнительной власти, формирующие данные федерального статистического наблюдения [13].
- подсистема анализа и оценки – в её роли выступают сотрудники Министерства строительства и жилищно-коммунального хозяйства РФ, осуществляющие непосредственный расчет показателей [13];
- подсистема прогнозирования – в проекте не фигурирует [13];
- подсистема представления результатов – сведения о данной подсистеме в нормативных документах отсутствуют;
- подсистема коррекции – в методике показателей проекта не фигурирует [13].

Таким образом, концепция системы мониторинга с Рисунка 1.1 применительно к системе мониторинга, построенной на основе компонентов показателей федерального проекта «Развитие инфраструктуры в населенных пунктах», имеет вид, представленный на Рисунке 1.3.

Нормативная база федерального проекта позволяет выделить следующие ключевые недостатки используемой системы мониторинга:

- Расчет части показателей состояния городской среды производится 1 раз в год, хотя мировая практика показывает, что изменения городской среды происходят чаще и требуют более регулярного мониторинга [73]. Таким образом, значительную часть календарного года оценки показателей состояния городской среды могут быть неактуальны [11, 13].
- Сбор первичных показателей через посредника (органы власти) [13] в сочетании с отсутствием четкого регламента сбора данных может негативно сказываться на актуальности, полноте и достоверности первичных данных и результатов мониторинга, а также делает мониторинг на основе показателей

федерального проекта затратным по ресурсам и затрудняет более регулярную актуализацию для нивелирования предыдущего недостатка.

- Актуальность первичных данных не регламентирована [13]. Возможно использование неактуальных данных, что снижает достоверность результатов мониторинга и его эффективность в целом [15].

- Показатели рассчитываются либо в целом по городу, либо по целому субъекту РФ [13] и при обнаружении нештатных ситуаций не позволяют оперативно локализовать конкретный элемент, требующий корректирующего воздействия, что негативно сказывается на эффективности мониторинга с точки зрения информативности результатов [15].

- Оценивается состояние только городов, отнесенных к опорным населенным пунктам [13], что также негативно сказывается на полноте и информативности результатов мониторинга [15].

- При оценке социальной инфраструктуры акцент делается на состоянии объектов, но не на их способности обеспечить удовлетворение имеющихся потребностей населения [13].

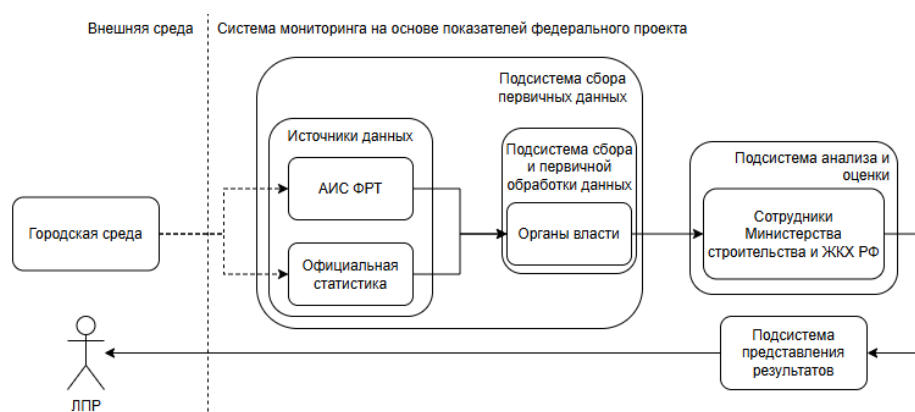


Рисунок 1.3 – Концепция системы мониторинга состояния городской среды на основе показателей федерального проекта

Также используемые показатели состояния не отражают состояние части объектов городской среды, относящихся к пространствам жилья, социальной и общественно-деловой инфраструктуры [13] (соответствующие элементы

городской среды выделены цветом на Рисунке 1.2.). Не учтены следующие значимые показатели состояния городской среды:

1. Потребительские свойства жилья: детали состояния жилья, его окружения, способность удовлетворять потребности населения.
2. Потребительские свойства коммерческих объектов: не отражена способность подобных объектов удовлетворять потребности населения.
3. Нагрузка на существующую социальную инфраструктуру, а именно на учреждения среднего и дошкольного образования, а также здравоохранения: случаи перегрузки объектов, негативно влияющие на качество жизни и состояние городской среды в целом, мониторинг на основе показателей федерального проекта обнаружить не может [13].

В результате показатель определенности состояния объекта мониторинга (1.3) снижается из-за недостаточного охвата значимых свойств мониторингом [13], что снижает общую эффективность мониторинга [15] с точки зрения полноты и достоверности результатов. Перечисленные недостатки ведут к снижению системности и информированности решений, что может приводить к усугублению проблем в результате неверного корректирующего воздействия [15, 20].

Следует отметить, что обеспеченность социальной инфраструктурой (учреждениями среднего и дошкольного образования, а также здравоохранения) оценивается в составе, например Национального рейтинга качества жизни, однако оценка сводится к подсчету долей населения, попадающих в зону обслуживания таких учреждений, однако их загруженность, напрямую влияющая на эффективность удовлетворения ими потребностей населения, проживающего в зоне обслуживания, ни в федеральном проекте, ни в Национальном рейтинге качества жизни не оценивается [10, 13].

Подобные проблемы характерны и для других решений в области комплексного мониторинга состояния городской среды, источником данных для которых являются официальная статистика и др. государственные системы [16, 42].

1.5 Постановка задачи повышения эффективности мониторинга

Исходя из перечисленных выше недостатков действующей системы мониторинга, актуальными задачами в рамках настоящего исследования являются:

1. Повышение определенности состояния объекта мониторинга (городской среды) за счет покрытия мониторингом дополнительных значимых показателей состояния и использования для их оценки дополнительных открытых источников, т.е. повышение полноты результатов мониторинга.

2. Повышение информативности результатов в части локализации объектов городской среды и/или отдельных городских территорий в нештатном состоянии.

Определенность состояния объекта мониторинга, как отмечалось выше, определяется по формуле (1.3). Таким образом, критерий эффективности мониторинга с точки зрения полноты результатов может быть записан как:

$$D = \frac{\sum_{i=1}^c w_{p_i} n_{p_i}}{\sum_{j=1}^d w_{p_j} n_{p_j}} \rightarrow 1, \quad (1.7)$$

где p_i – свойства объекта мониторинга, охваченные системой мониторинга;

p_j – все свойства элементов объекта мониторинга;

w_{p_i} – вклад свойства p_i в функционирование объекта мониторинга;

n_p – число элементов объекта мониторинга, обладающих свойством p .

Недостатки в части информативности результатов мониторинга для действующей системы связаны с оценкой городов в целом, без пространственной детализации, и, как следствие, невозможностью по результатам мониторинга установить конкретные входящие в городскую среду объекты (элементы объекта мониторинга) и/или отдельные городские территории, перешедшие в нештатное состояние. Исходя из вышеописанного, критерий эффективности мониторинга с точки зрения информативности результатов можно выразить через точность

локализации объектов городской среды в нештатном состоянии, что можно быть описано как:

$$K_{loc} = \frac{S_a}{S_l} \rightarrow 1, \quad (1.8)$$

где K_{loc} – коэффициент локализации, $K_{loc} \in [0, 1]$;

S_l – площадь области поиска объекта в нештатном состоянии;

S_a – площадь самого объекта в нештатном состоянии (площадь его земельного участка).

Поскольку перечисленные частные критерии эффективности являются независимыми, разнородными и несопоставимыми, расчет интегрального критерия эффективности мониторинга следует производить на основе функции желательности Харрингтона [74]:

$$d = e^{(-e)^{-z_i}} \rightarrow 1, \quad (1.9)$$

где d – показатель желательности в цифровом выражении $d \in [0, 1]$;

z_i – кодированное значение i -го частного критерия.

Для возрастающих критериев:

$$z_i = a_0 - a_1 x_i, \quad (1.10)$$

где x_i – фактическое значение i -го частного критерия;

коэффициенты a_0 и a_1 определяются из системы уравнений, связывающей граничные значения критерия с соответствующими значениями желательности.

Граничные значения критерия представлены в Таблице 1.2. В качестве левой границы взяты текущие значения для действующей системы мониторинга. Правые границы определены исходя из международной практики и описанных выше недостатков действующей системы мониторинга.

Таблица 1.2 – Границы диапазонов удовлетворительности

Частный критерий	Левая граница x_{i0}	Правая граница x_{i1}
K_{loc}	0	1
D	16	21

Интегральный критерий эффективности принимает вид:

$$Des = \sqrt[n]{\prod_{i=1}^n d_i}, \quad (1.11)$$

где $n = 3$ – мощность множества частных критериев;

d_i – показатель желательности для i -го критерия.

В контексте исследования достичь повышения эффективности по вышеописанным критериям можно за счет воздействия на следующие элементы системы мониторинга:

1. Систему сбора первичных данных путем увеличения числа используемых источников, внедрения новых и/или улучшения существующих процедур сбора, обработки и контроля качества данных.

2. Модули анализа и оценки за счет расширения перечня критериев, а также улучшения существующих и/или внедрения новых методов оценки.

Для этого в рамках исследования предлагается разработать и интегрировать в действующую систему мониторинга элементы для решения задачи оценки состояния по ранее неучтенным значимым показателям:

1. Потребительским свойствам жилых объектов городской среды на основе показателей их внутреннего состояния и окружения – 1 показатель.

2. Потребительским свойствам коммерческих объектов городской среды на основе показателей их внутреннего состояния и окружения – 1 показатель.

3. Обеспеченности социальной инфраструктурой (образованием и здравоохранением) через оценку нагрузки на существующие объекты – 3 показателя.

Реализация предлагаемых изменений предполагает применение методов системного анализа, моделирования, методов искусственного интеллекта (в частности, обработки естественных языков) методов статистики, а также методов планирования эксперимента и обработки экспериментальных данных.

Рассматривая возможные источники данных, способные выполнить в рамках системы мониторинга функцию датчиков, был сделан вывод [75], что наиболее актуальные и подробные данные о состоянии жилых и коммерческих

объектов городской среды содержатся в пользовательских объявлениях об операциях с недвижимым имуществом (жилыми и коммерческими градостроительными объектами), публикуемых в открытых источниках [76]. К их числу относятся тематические сайты, специализированные платформы объявлений, а также тематические группы и сообщества в социальных сетях.

Особенностью является формат представления данных: в виде текстов на естественном (в рамках исследования – русском) языке, содержащих подробные описания объектов в свободном изложении, что приводит к необходимости извлечения из текстов структурированных данных. Решение данной задачи относится к зоне ответственности подсистемы сбора и обработки первичных данных.

При работе с данными, источником которых является человек (авторы объявлений), приходится учитывать вероятность публикации недостоверных сведений. Кроме того, не всякое объявление содержит достаточно полное описание объекта для использования в целях мониторинга: текст бывает ограничен минимумом информации, что делает объявление непригодным для целей исследования. Таким образом, возникает задача контроля качества данных, а именно их очистки и обогащения [76]. Для её решения также выделяется элемент в подсистеме сбора и обработки первичных данных.

Наиболее актуальным открытым источником пространственных данных об объектах городской среды, в частности учреждениях образования и здравоохранения, их взаимном расположении являются ГИС. Методика федерального проекта не задействует ГИС [13], которые, как упоминалось ранее, нашли свое применение при оценке инфраструктурной обеспеченности населения в Национальном рейтинге качества жизни [10]. Аналогично ГИС являются источником данных об окружении жилых и коммерческих объектов, об объектах оказания различных услуг (аптеках, продовольственных магазинах и т.п.), также удовлетворяющих потребности населения. Сбор данных о перечисленных объектах относится к зоне ответственности подсистемы сбора и первичной обработки данных, а интерпретации – к зоне ответственности подсистемы анализа

и оценки. Внедрение описанных решений позволит получать актуальные первичные данные за счет автоматизации процессов сбора и предварительной обработки информации, повысить полноту и информативность результатов, а также потенциально использовать новые инструменты для проактивной оценки состояния.

Таким образом, разрабатываемые в данном исследовании решения, укладываются в представленную на Рисунке 1.4 концепцию системы оценки состояния городской среды по дополнительным критериям.

Ключевые модули системы оценки выполняют следующие функции:

1. Модуль извлечения и трансформации данных – собирает первичную информацию о состоянии жилых и коммерческих объектов из объявлений, а также их окружении и объектах социальной инфраструктуры из ГИС, извлекает структурированные данные для дальнейшей обработки.
2. Модуль очистки и обогащения данных – контролирует полноту, осуществляет прогноз достоверности и очистку данных.
3. Модуль оценки обеспеченности социальной инфраструктурой – выявляет случаи превышения допустимой нагрузки и вычисляет интегральный показатель обеспеченности социальной инфраструктурой.
4. Модуль оценки потребительских свойств жилых и коммерческих объектов – оценивает способность объектов удовлетворять потребности населения на основе данных, ранее полученных из объявлений и ГИС.

Разработанная концепция системы оценки, помимо расширения перечня источников данных, при интеграции в состав системы мониторинга на основе показателей федерального проекта «Развитие инфраструктуры в населенных пунктах», предусматривает функциональное и архитектурное расширение двух её подсистем.

На Рисунке 1.5 приведена концептуальная схема системы мониторинга на основе, дополненной элементами для оценки по расширенному перечню критериев (выделены цветом).

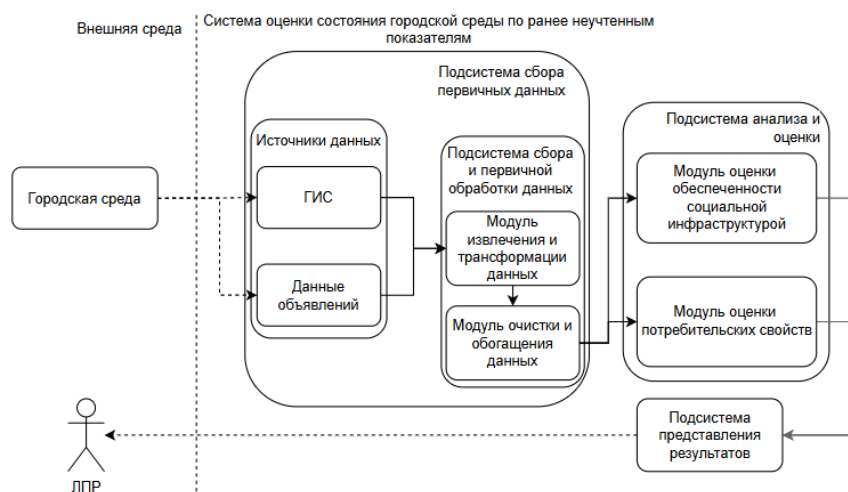


Рисунок 1.4 – Концепция системы оценки состояния городской среды по ранее неучтенным показателям

Целью предлагаемых изменений является повышение эффективности мониторинга состояния городской среды посредством учета дополнительных значимых показателей состояния и применения автоматизированного интеллектуального анализа гетерогенных открытых данных, ранее в процессе мониторинга не использованных.

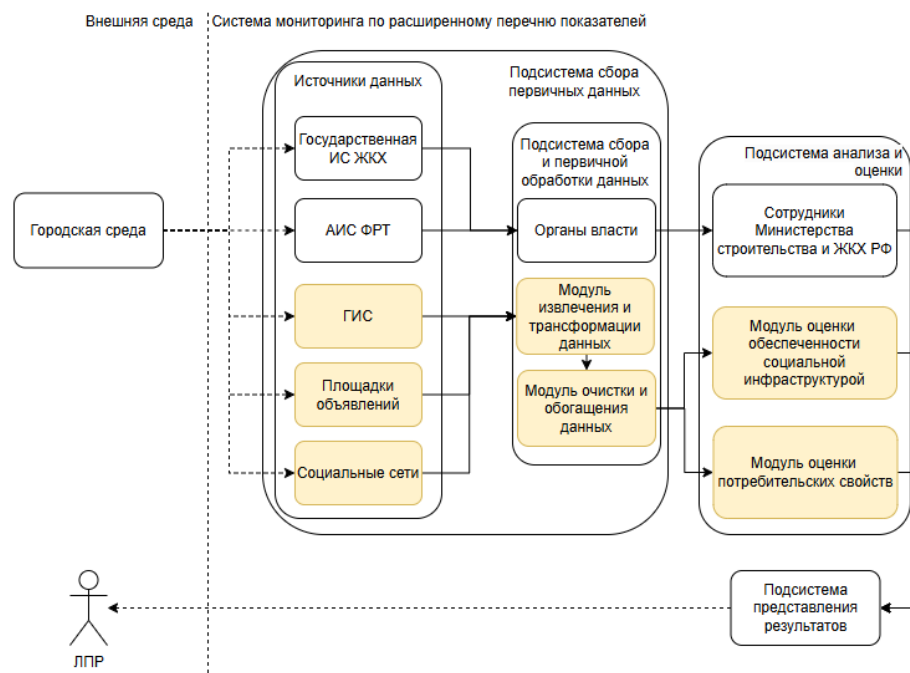


Рисунок 1.5 – Модификация системы мониторинга состояния городской среды на основе показателей федерального проекта

1.6 Выводы по первой главе

1. Анализ исследований и существующих решений позволяет сделать вывод о несовершенстве применяемых в настоящий момент методов мониторинга состояния городской среды. Концентрация на мониторинге отдельно взятых подсистем, использование потенциально недостоверных или неактуальных данных, отсутствие мониторинга части подсистем в комплексных методах мониторинга обуславливают необходимость дальнейшего проведения исследований в данном направлении. Основные группы существующих решений в области мониторинга состояния городской среды как сложной системы, их авторы и ключевые недостатки приведены в Таблице 1.1.

2. Наиболее актуальным для РФ решением, применяемым в настоящий момент, является система мониторинга на основе показателей федерального проекта «Развитие инфраструктуры в населенных пунктах», относящаяся к группе комплексного мониторинга на основе множества показателей из Таблицы 1.1.

3. Ключевыми недостатками, негативно влияющими на эффективность решения на основе показателей федерального проекта, являются полнота и информативность результатов, вытекающие из неполного покрытия значимых показателей состояния городской среды мониторингом, сбора данных в ручном режиме и расчета в целом по городу либо целому субъекту РФ без пространственной детализации.

4. Предложено повышение эффективности существующего решения за счет покрытия мониторингом ранее неучтенных значимых показателей состояния городской среды: потребительских свойств жилых объектов, потребительских свойств коммерческих объектов и обеспеченности социальной инфраструктурой (образованием и здравоохранением). В общей сумме 5 новых показателей.

5. Выделены критерии эффективности мониторинга в контексте исследования: полнота и информативность получаемой информации о состоянии

городской среды, а также интегральный критерий на основе функции желательности Харрингтона.

6. Предложена концепция системы мониторинга по ранее неучтенным показателям состояния (Рисунок 1.4) и соответствующая модификация системы мониторинга на показатели федерального проекта (Рисунок 1.5)

7. Приведена формальная постановка задачи, решение которой необходимо для достижения цели исследования.

ГЛАВА 2. Разработка метода извлечения и трансформации данных

Для оценки состояния сложной пространственно распределенной системы (городской среды) по выделенным ранее неучтенным показателям необходимы инструменты получения данных о состоянии входящих в неё объектов.

Наиболее подробная и актуальная информация о состоянии и внутренних характеристиках жилых и коммерческих объектов городской среды, как упоминалось выше, содержится в пользовательских объявлениях об операциях с недвижимым имуществом, размещаемых на специализированных площадках объявлений [75], а также в тематических разделах в социальных сетях [77]. Наиболее же актуальным открытым источником пространственной информации: взаимном расположении объектов городской среды, инфраструктуры, пешей и транспортной доступности, являются ГИС.

Таким образом, назначением метода является решение двух подзадач ETL [78]: извлечения и трансформации данных из источников. Схема процесса ETL и место в нём разработанного метода продемонстрированы на Рисунке 2.1.

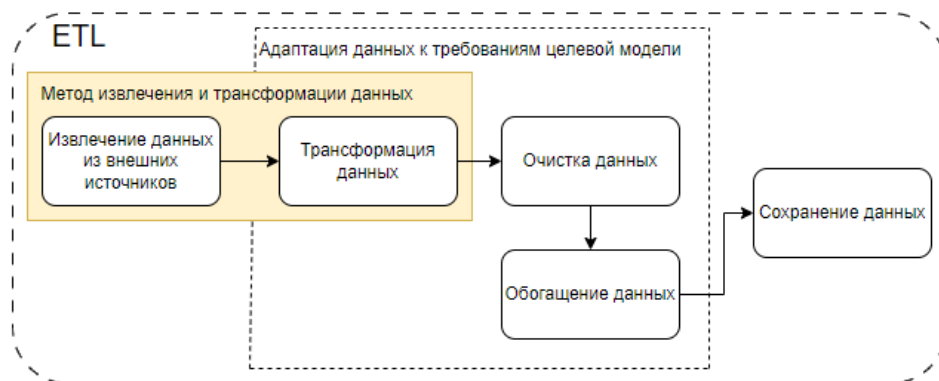


Рисунок 2.1 – Разработанный метод на схеме процесса ETL

Разработанный метод *DEM* (от англ. «Data Extraction Method») имеет следующее формальное представление:

$$DEM = \langle S, C, f_{sc}, M, T_s \rangle, \quad (3.1)$$

где S – множество внешних гетерогенных источников данных о состоянии объектов среды конкретного города;

C – множество доступных функций сбора данных;

f_{sc} – функция применимости элементов множества C к элементам из множества S ;

M – целевое структурированное представление данных из источников S , полученных функциями C ;

T_S – множество функций трансформации данных из источников S к целевым представлениям M .

Назначение метода – сбор и последующая трансформация к целевому структурированному представлению данных о состоянии жилых и коммерческих объектов городской среды, а также их окружении и объектах социальной инфраструктуры из гетерогенных открытых источников.

2.1. Особенности первичных данных

Особенностью объявлений об операциях с недвижимым имуществом (жилыми и коммерческими объектами городской среды) является представление информации в виде текстов на естественном (русском) языке, содержащих подробные описания объектов в свободном изложении [75], т.е. в неструктурированном (см. Рисунок 2.2), а в отдельных случаях – в частично структурированном (см. Рисунок 2.3) виде, что приводит к необходимости решения задачи трансформации данных – извлечения из текстов структурированных данных.

На тематических площадках структура объявлений часто регламентирована внедрением обязательных к заполнению при размещении объявления полей, отражающих ключевые характеристики объекта. Тогда текст предваряется

кратким структурированным блоком ключевых сведений (см. Рисунок 2.4), а в самом тексте данные из блока обычно не дублируются.

Продаю 1-комнатную квартиру. СОБСТВЕННИК.
 Общая площадь квартиры 45,5 м2 .
 Комната 16.8 м2
 Просторная кухня 13.9 м2
 Высота потолков 2.7 м.
 Дом теплый, кирпичный, сдан в декабре 2016 года.
 Планировка квартиры согласно проекту.
 Квартира полностью отделана, лоджия застеклена и утеплена. На окнах в квартире установлены стеклопакеты с тремя стеклами.
 Санузел совмещенный, полностью облицован. Счетчики на ГВС , ХВС и электроэнергию установлены.
 Шикарный вид из окон: частные дома, школа, стадион, 3-ая Продольная. Ангарский пруд.
 В доме 2 лифта: пассажирский и грузовой, есть подземная автостоянка.
 В шаговой доступности: остановки общественного транспорта (трамвай 2, маршрутное такси 98 и 10с), школа, детский сад, поликлиника, почта, СБЕР, стадион.
 В доме с наружной стороны на первом этаже сетевые магазины, аптека, пекарня и пункты выдачи: WB, OZON.
 Удобное месторасположение с развитой инфраструктурой. До центра города на трамвае 10 мин. Открытая парковка для машин, имеется подземный паркинг.
 Квартира приобретена от застройщика без обременений и без материнского капитала. Один собственник.
 Документы в полном порядке, готовы к сделке.
 Звоните, отвечу на все интересующие вас вопросы.

Рисунок 2.2 – Пример неструктурированного описания объекта

#студия #Василеостровская

Квартира сдаётся на период от 1 месяца (до мая)
 Сдается напрямую от собственника двухуровневая большая квартира-студия (45 кв.м.)на Васильевском острове.
 Квартира расположена в топовой локации Санкт-Петербурга. 10 минут пешком до метро. Рядом супермаркет Лента,вкусвилл,магнит и т.д.Выход к Неве. Есть ключ от закрытого двора с парковкой.

В квартире есть все необходимое:

- ✓ Две двуспальные кровати 160 и 140 см
- ✓ Большой жк-телевизор
- ✓ Скоростной wi-fi
- ✓ Полностью оборудованная кухня (плита, холодильник, чайник,микроволновка, посуда)
- ✓ Санузел с ванной
- ✓ Стиральная машинка, утюг, гладильная и сушильная доски
- ✓ Водонагреватель

Стоимость 55 000 +ку (5000р)+залог.

Рисунок 2.3 – Пример частично структурированного описания объекта

Количество комнат: 1	Санузел: совмещенный
Общая площадь: 45.5 м²	Окна: на улицу
Площадь кухни: 13.9 м²	Ремонт: евро
Жилая площадь: 16.8 м²	Мебель: хранение одежды
Этаж: 15 из 18	Способ продажи: свободная
Балкон или лоджия: лоджия	Вид сделки: возможна ипотека
Высота потолков: 2.7 м	

Рисунок 2.4 – Пример структурированного блока описания объекта

Производной от способа представления данных является проблема разного представления эквивалентных значений характеристик объектов: использование синонимов для категориальных характеристик или указание количественных значений в разных единицах измерения и/или в разном формате (числами либо прописью) [75, 77]. В результате, возникает потребность в нормализации найденных значений для единообразия обработки.

В социальных сетях актуальна частичная структуризация объявлений с помощью тегов перед текстом, которые также могут отражать ключевые характеристики описываемого объекта, т.е. также подлежат анализу и интерпретации для обработки всех присутствующих в объявлениях данных.

Также объявления могут сопровождаться фотографиями объекта, изображениями планировки и т.п. Данная информация также полезна для контроля качества данных. Автоматическая обработка изображений требует преобразования их к набору формальных признаков, на основании которых машина сможет сделать вывод о содержимом и подобию [79]. Для оптимизации целесообразно преобразовывать изображения к наборам признаков сразу при извлечении данных и хранить признаки вместе с изображениями.

Цены объектов имеют свойство быстро устаревать, однако данные о них также полезны для контроля достоверности.

Объявления также содержат фотографии объектов. Наличие в разных объявлениях одинаковых фотографий имеет два вероятных объяснения: объявления описывают один и тот же объект или объекты разные, но фотографии в одном объявлении позаимствованы из другого. Т.е. необходим инструмент

эффективного машинного сравнения растровых (формат представления фотографии) изображений для определения совпадающих.

ГИС – класс информационных систем, обеспечивающих сбор, хранение, обработку, доступ, визуализацию и распространение пространственных данных и связанной с ними информации [80].

Данные в ГИС делятся на четыре категории [80]:

1. Пространственные данные (основные данные ГИС). Представлены единичной координатой (или совокупностью координат), определяющих положение объекта в пространстве (и его форму).
2. Атрибутивные данные, отражающие дополнительную информацию об объекте: его качественных и количественных характеристиках (площади земельного участка, отзывах и т.п.).
3. Библиотеки условных знаков для отображения объектов на карте.
4. Метаданные. Как правило, содержат информацию об источниках данных, методах их получения, дате и времени получения и т.п.

В контексте исследования значимыми являются категории 1, 2 и 4.

Пространственные данные содержат информацию о том, какие объекты расположены на конкретной территории, как они расположены относительно друг друга и т.д. Применительно к объектам оказания услуг, для которых существуют актуальные нормативные документы, пространственные данные могут дать информацию о зоне обслуживания и попадающих в эту область объектах.

Пространственные данные не несут информации, является ли объект жилым зданием, образовательным учреждением или городским парком. Определение типа объекта и детализация сведений о нем требует привлечения атрибутивных данных. Особый интерес представляют данные о статусе объекта т.к. в ГИС встречается информация об объектах в статусе «больше не работает», «закрит навсегда» и т.п. Отбрасывать такие объекты на этапе сбора данных целесообразно только если работа ведется с одной ГИС. В противном случае данные о закрытых объектах полезны для перекрестной проверки данных из разных ГИС.

Метаданные представляют интерес с точки зрения контроля актуальности. Различные ГИС обновляются с разной частотой и могут обладать различной информацией о городской среде на конкретных территориях.

2.2. Этапы метода извлечения и трансформации данных

В рамках исследования разработан метод извлечения и трансформации данных о состоянии городской среды, представленный на Рисунке 2.5. Как видно по рисунку, этапы составления словарей и правил поиска, извлечения и нормализации фактов актуальны только для работы с пользовательскими объявлениями. Данные с ГИС уже являются структурированными и в дополнительной предварительной структуризации не нуждаются.

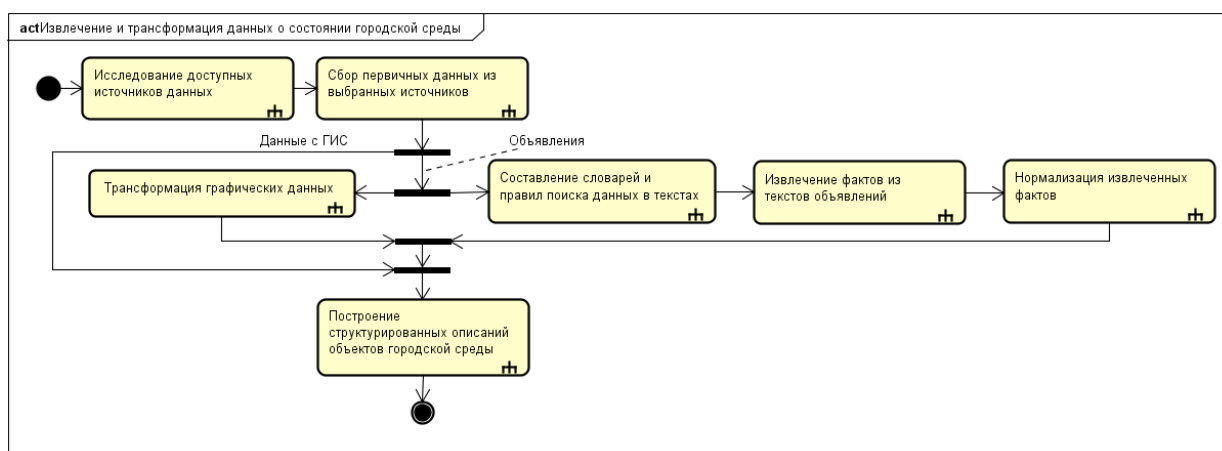


Рисунок 2.5 – Метод извлечения и трансформации данных

2.2.1. Исследование доступных источников первичных данных

Задача получения структурированных данных о состоянии городской среды требует проведения подготовительной аналитической работы, слабо поддающейся автоматизации и требующей участия ЛПР, в интересах которого

производится мониторинг, и экспертов, его консультирующих. Схема алгоритма данного этапа представлена на Рисунке 2.6.

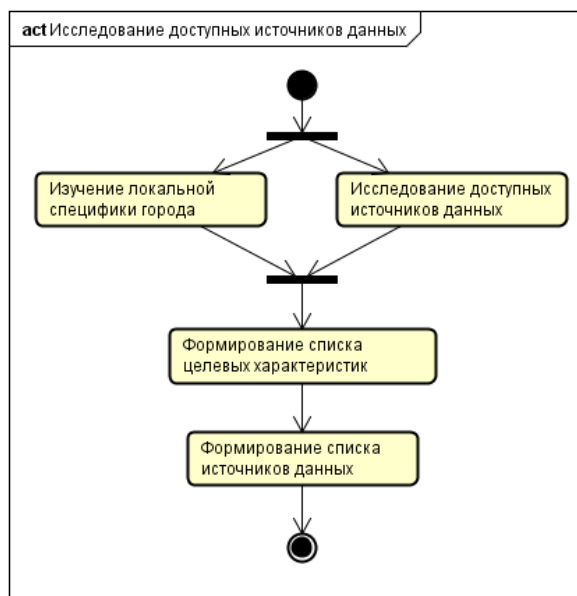


Рисунок 2.6 – Алгоритм исследования доступных источников первичных данных

В первую очередь, на данном этапе выбираются конкретные источники первичных данных, определяется перечень данных, сбор и трансформация которых будут производиться на следующих шагах. Так, по результатам исследования доступных источников делается вывод о структуре данных в каждом из них, производится анализ с точки зрения представленных в них характеристик объектов, результатом которого является список целевых характеристик (фактов) для дальнейшего анализа и способы их целевого структурированного представления.

Список целевых характеристик формируется на основе предпочтений ЛПР, мнений консультирующих экспертов, а также методов статистики, т.к. включение в мониторинг характеристик, информация о которых в источниках почти не встречается, не имеет смысла. Конечный состав характеристик зависит от типа объекта: перечень для, например, квартиры по очевидным причинам отличается от такового перечня для офиса [81].

Также предварительное исследование необходимо из-за возможной локальной специфики, характерной для различных стран [42, 44-47] или отдельных городов в пределах страны, которую необходимо учитывать при сборе и трансформации первичных данных. Примером такой специфики является язык написания объявлений, зависящий от государственных (или наиболее распространенных) языков страны и/или конкретной площадки. Другой пример – популярность конкретных площадок и социальных сетей в стране, определяющая их пригодность к использованию в качестве источника данных. Пример специфики уровня конкретного города – использование обиходных или сокращенных синонимов, например, названий топонимов.

Другой пример специфики – цифры вместо числительных корней при словообразовании: написание «3-хкомнатная» вместо литературной нормы «трёхкомнатная» в объявлениях об операциях с жилыми объектами.

Для получения корректных результатов и эффективного решения задачи оценки состояния городской среды на основе полученных данных, важно нивелировать максимально возможный объём подобных специфик. Обнаруженные в результате проведенного исследования специфики систематизируются и представляются в виде модели знаний (конкретная форма представления не принципиальна, т.к. в дальнейшем машинная обработка модели знаний не предполагается) для использования в дальнейших этапах разработанного метода.

Таким образом, результатами данного этапа являются:

- список открытых источников для дальнейшего использования при сборе первичных данных о состоянии городской среды;
- знания, отражающие локальную специфику ОМ (конкретного города) для использования на дальнейших этапах;
- списки целевых характеристик и производные от них целевые структурированные представления данных для каждого типа объектов.

2.2.2. Сбор первичных данных

Процесс сбора первичных данных в детальном рассмотрении не нуждается, т.к. представляет собой прикладную техническую задачу со специфичными решениями для конкретных источников. Потребность в решении научной задачи в рамках построения процесса сбора данных отсутствует. Единственный момент, который следует отметить – это необходимость получения географических координат объектов из объявлений для дальнейшего взаимодействия с данными ГИС, что реализуется с помощью штатного функционала геокодирования ГИС.

С помощью методов геопространственного анализа из ГИС получают данные об объектах социальной инфраструктуры (учреждениях образования и здравоохранения) и объектах оказания услуг (аптеках, магазинах, местах общественного питания и пр.), их расположении, географических координатах, атрибутах и метаданных, а также о жилых домах как точках концентрации потенциальных потребителей услуг. Процесс сбора данных из ГИС также является технической задачей, решаемой с использованием прикладных программных интерфейсов (англ. Application programming interface, API), предоставляемых ГИС.

2.2.3. Трансформация текстовых первичных данных

Для извлечения значений характеристик объектов (далее – фактов) из текстов их описаний на естественном языке применяются несколько стандартных подходов [80]:

1. Подход на основе регулярных выражений.
2. Использование классификаторов:
 - а. Генеративные (наивный классификатор Байеса).

- б. Линейные классификаторы (логистическая регрессия).
- 3. Статистические модели (марковские, условное случайное поле).

Существуют и гибридные подходы, объединяющие некоторые из перечисленных решений. Однако из доступного на данный момент множества технологий анализа текстов на естественном языке, наиболее популярными свободно распространяемыми являются следующие решения.

GATE (General Architecture for Text Engineering) включает набор модулей: токенизатор, справочник географических названий, разделитель предложений, модуль разметки частей речи и указатель тегов [66].

Natural Language Toolkit (NLTK) – набор библиотек и программ для символической и статистической обработки естественного языка на языке Python, применимый как для английского, так и для русского языка [67]. В перечень задач, решаемых с помощью NLTK, входят:

- 1. Разбиение текста на слова.
- 2. Удаление незначащих слов (стоп-слов).
- 3. Стемминг – сведение слов к их корневой форме через удаление суффиксов и окончаний.
- 4. Лемматизация – морфологический анализ с целью сведения слова к корректной начальной форме.
- 5. Анализ тональности текстов (определение позитивных и негативных настроений).

Однако NLTK не предусматривает решения задачи извлечения фактов (конкретных значений искомых характеристик) из описаний объектов.

Ещё одно решение – набор инструментов машинного обучения OpenNLP для обработки текста на естественном языке. Поддерживает наиболее распространенные задачи NLP: токенизацию, сегментацию предложений, таргетирование на части речи, извлечение именованных сущностей, фрагментацию, разбор и определение языка. Эти задачи обычно необходимы для создания более совершенных служб обработки текста [71].

Кроме того, известными примерами проприетарных прикладных решений для задач анализа текстов являются такие продукты, как IBM Watson [83], суперкомпьютер, ориентированный на понимание вопросов на естественном языке, и поиск ответов на них, и Wolfram Natural Language Understanding (часть системы Wolfram Alpha) [84].

Общим для перечисленных решений является их ориентированность на работу с английским языком. Поддержка других языков является весьма ограниченной или полностью отсутствует. Также затрудняющим фактором для их применения в настоящем исследовании является нацеленность на статистический, но не содержательный анализ текстов.

Решением без данного недостатка являются синтаксические анализаторы (парсеры), работающие на основе формальных грамматик и словарей. Примером является формальный язык описания грамматик Томита, разработанный компанией Яндекс для решения задач по извлечению именованных сущностей из текстов на русском языке, а также работающий с данным языком Томита-парсер [72, 75, 85]. Данный язык поддерживает условия согласованности слов, самостоятельное присвоение словам грамматических характеристик, явное указание способа интерпретации фрагментов текста и многое другое [85].

Таким образом, синтаксические анализаторы являются наиболее подходящим решением задачи извлечения фактов из текстов на естественном языке в контексте настоящего исследования.

При использовании синтаксических парсеров в первую очередь необходимо составить списки названий ключевых топонимов города, а также общепринятых синонимов их названий. Правила образования таких синонимов крайне нечетки и сложно формализуемы. Решение подобной задачу формальными методами затратно по ресурсам, а выгода от решения по сравнению с затратами незначительна в силу локальной специфичности для каждого города. Т.к. количество значимых топонимов в городах относительно невелико (речь идёт о десятках), было решено фиксировать названия и возможные синонимы в соответствующих словарях с явным указанием корректной полной формы

названия для нормализации. Составление словарей также актуально для популярных сокращений значимых ключевых слов, например, «пр-т», «ул.» и т.д. для их замены леммой (полной грамматической формой) при извлечении из текстов.

Представление словарей зависит от средств технической реализации. Рассмотрим представление топонима (Тракторозаводского района г. Волгограда) в словаре на языке Томита [72] – см. Рисунок 2.7.

Поля «key» обозначают возможные представления топонима в тексте, а поле «lemma» содержит полную форму названия. На этапе нормализации извлеченных данных все найденные варианты названия приводятся к лемме.

```
district "тракторозаводский"
{
    key = "ТЗР"
    key = "тракторный"
    key = "тракторозаводский"
    lemma = "тракторозаводский"
}
```

Рисунок 2.7 – Представление инстанции словаря на языке Томита

Следующий этап – составление правил поиска фактов в текстах. Поскольку различные целевые характеристики имеют конечный перечень распространенных способов представления в тексте [75], становится возможным составление конечного набора формальных правил, эти представления описывающих. На каждую целевую характеристику составляются свои правила. Исключения возможны только если характеристики всегда встречаются в тексте вместе. В данном случае корректна следующая постановка задачи: составить формальные правила построения цепочек слов, которыми в текстах на естественном языке может быть представлена каждая из целевых характеристик описываемых объектов, что практически повторяет определение формальных грамматик [86].

Формальная грамматика (англ. Formal grammar) – метод описания формального языка, который описывается следующей четверкой [87]:

$$G = \langle \Sigma, N, S \in N, P \subset ((\Sigma \cup N)^* N (\Sigma \cup N)^*) \rightarrow (\Sigma \cup N)^* \rangle, \quad (3.2)$$

где Σ – конечное непустое множество символов (алфавит), на основе которых строятся грамматически корректные выражения из терминалов (англ. terminals);

N – множество промежуточных символов грамматики – нетерминалов (англ. nonterminals);

S – начальный символ грамматики (англ. start symbol);

P – набор правил вывода корректных цепочек (англ. production rules).

В контексте исследования символами грамматики являются целые слова, а терминалами – различные ключевые слова и символы. В общем виде грамматики представляют собой наборы правил следующего вида:

$$S \rightarrow S_1 S_2 \dots S_i \{Interp_x(S_i)\} \dots S_n, \quad (3.3)$$

где S – нетерминал грамматики (может быть начальным символом);

$S_1 S_2 \dots S_n$ – правая часть (терминалы и/или нетерминалы), описывающая принципы построения цепочек, представляющих характеристику, $S_i, i \in Z$ может быть нетерминалом и иметь в грамматике свои правила вида (3.3);

$Interp_x(S_i)$ – оператор интерпретации, связывающая выводимую из S_i цепочку со значением целевой характеристики x . Фигурные скобки $\{ \}$ обозначают опциональность.

Разработанный алгоритм извлечения фактов из текстов объявлений, представленный на Рисунке 2.8, работает отдельно с каждым объявлением, целевой характеристикой и соответствующим ей набором грамматик и словарей ввиду независимости объявлений и характеристик друг от друга.

При обнаружении соответствия цепочки слов формальной грамматике искомой характеристики цепочка интерпретируется: полностью или частично принимается в качестве значения искомой характеристики. После интерпретации одной характеристики, алгоритм переходит к поиску значений следующей, пока все целевые характеристики не будут обработаны.

Для обеспечения единообразного и единого представления значений каждой целевой характеристики и их корректной машинной

сравнимости требуется нормализация извлеченных фактов. Были выделены следующие случаи, требующие нормализации:

1. Использование синонимов названий городских топонимов, которые требуют приведения к лемме.
2. Представление численных значений в различных единицах измерения, требующие приведения к общей единице измерения – рублям, метрам квадратным или др. в зависимости от характеристики.
3. Словесное представление количественных характеристик, например, количества комнат словами: «трёхкомнатная», «3х-комнатная». Было решено представлять значения таких характеристик числом.

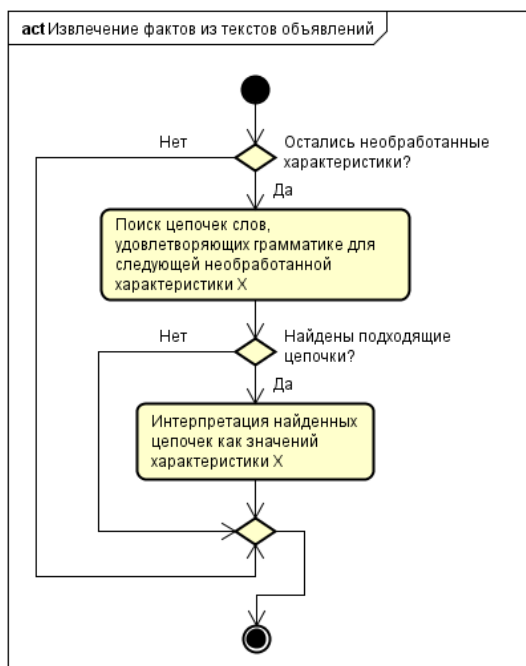


Рисунок 2.8 – Алгоритм извлечения фактов из текстов объявлений

Конкретный способ нормализации выбирается исходя из конкретной целевой характеристики и используемых ей единиц измерения.

2.2.4. Трансформация графических первичных данных

Существуют разнообразные алгоритмы решения задачи сравнения изображений [79, 88-92], из которых наиболее подходящими для целей исследования являются методы на основе перцептивных хэш-алгоритмов [79, 88].

В основе данные методы содержат преобразование исходных изображений к индивидуальным цепочкам бит (перцептивным хэшам), получаемым на основе ключевых характеристик оригинала (общая схема построения хэша представлена на Рисунке 2.9) [79]. Лавинный эффект небольших изменений для перцептивных хэш-алгоритмов неактуален [91], что позволяет сравнивать одинаковые изображения в разных разрешениях или с разными алгоритмами сжатия через расстояние Хэмминга [79].

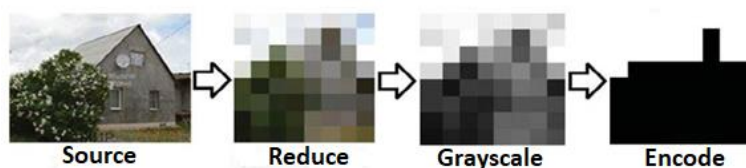


Рисунок 2.9 – Процесс перцептивного хэширования изображения

Однако изображения могут быть опубликованы не в исходном виде, а подвержены некоторым искажениям с целью затруднить их автоматическое сравнение. Были выделены распространенные искажения изображений [93]: сдвиг изображения на n пикселей; цветокоррекция; масштабирование; обрезка; деформация; зеркальное отображение; поворот.

Масштабирование изображения, его незначительные искажения, цветокоррекция не являются затрудняющими факторами выбранного метода сравнения, т.к. перцептивные хэши относительно них инварианты [94]. Проблемой остается зеркальное отображение изображения, т.к. хэш зеркального изображения не является зеркальным по отношению к хэшу исходного [94]. Было решено вычислять и хранить два хэша: исходного изображения и зеркальной

копии. В таком случае совпадением изображение было решено считать совпадение перцептивного хэша исходного первого изображения с зеркальным хэшем второго и наоборот (см. Рисунок 2.10)

Поворот исходного изображения и сдвиг оставались искажениями без определенного способа противодействия. Опираясь на недостатки перцептивных хэш-алгоритмов [79, 88-94], было решено отказаться от использования единственного алгоритма, а прибегнуть к применению ансамбля методов: хэширования на базе Дискретного косинусного преобразования [79, 94] для исходного и зеркального изображений и хэширования на основе преобразования Радона [77, 92], инвариантного к вращениям, для противодействия искажениям через вращение.

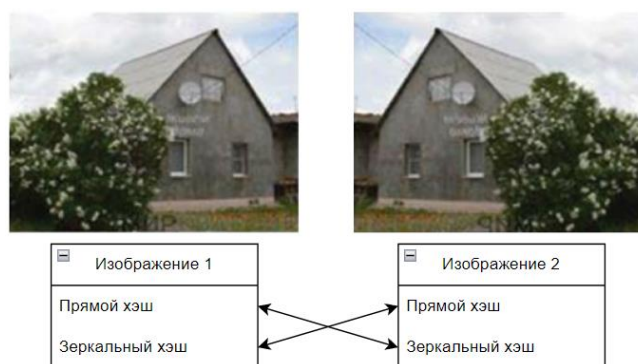


Рисунок 2.10 – Принцип сравнения зеркальных изображений

Полученные перцептивные хэши рассчитываются для каждого нового изображения, и хранятся вместе с ним.

2.2.5. Построение структурированных описаний объектов

Цель этапа – формирование записей данных в соответствии с заранее разработанными структурированными представлениями на основе извлеченных и нормализованных значений целевых характеристик объектов.

В более подробном описании данный этап не нуждается по причине рутинности процесса и его зависимости от технических средств реализации.

2.3. Выводы по второй главе

1. Выделены ключевые особенности исходных данных, получаемых из выбранных гетерогенных источников: пользовательских объявлений об операциях с недвижимым имуществом (основной источник актуальных атрибутивных данных об объектах городской среды) и ГИС (основной источник актуальных пространственных данных об объектах городской среды).

2. Разработан метод автоматизированного извлечения и трансформации данных о состоянии городской среды из гетерогенных источников (см. Рисунок 2.5) на основе обработки графических данных, естественных языков, а также пространственного анализа.

3. Описаны ключевые этапы разработанного метода: исследование доступных источников исходных данных, сбор исходных данных из выбранных источников, извлечение и нормализация значений характеристик объектов городской среды (фактов) из текстовых первичных данных, трансформация графических первичных данных и построение итоговых структурированных описаний объектов.

4. Описаны принципы реализации этапов разработанного метода и обоснован выбор технологий реализации.

Разработанный метод обладает универсальностью с точки зрения применимости к различным городам земного шара, хотя для переноса между городами может потребоваться адаптация отдельных технических элементов. В случае с городами в пределах одной страны либо городов в странах с общим языком, достаточно адаптации словарей ключевых слов, а в случае переноса на новый язык – потребуется обновление грамматик.

Также предлагаемый метод является масштабируемым за счет возможности расширения перечня целевых характеристик объектов, увеличения объёмов словарей и комплексности грамматик и т.д.

Конечной целью разработанных в Главе 2 решений является сбор и последующая трансформация к единому структурированному представлению данных о состоянии и окружении жилых и коммерческих объектов городской среды, а также об объектах социальной инфраструктуры из гетерогенных открытых источников для их последующей обработки машинным способом.

ГЛАВА 3. Разработка метода очистки и обогащения данных

При работе с несколькими источниками возможно получение нескольких записей об одном объекте городской среды (дублирование данных), что негативно влияет на достоверность результатов мониторинга. Требовалось решение задачи связывания данных (англ. Record Linkage) [95] путем удаления или свёртки данных об одном объекте городской среды в его интегральное описание.

Также очистка данных была необходима для исключения неполных данных. В рамках исследования исключению подлежали объекты, которые невозможно было соотнести с пространством города и получить информацию об окружении, т.е. данные без сведений о городе и/или адресе расположения.

При работе с данными, источником которых является человек (авторы объявлений), сохраняется вероятность публикации недостоверных сведений, что требует контроля достоверности данных. Поскольку на основе самих данных однозначный вывод о достоверности невозможен [76], задача свелась к прогнозированию потенциальной недостоверности и обогащения данных метками для обеспечения полноты исходных данных и результатов.

Для решения перечисленных задач в рамках исследования разработан метод очистки и обогащения данных (см. Рисунок 3.1), позволяющий производить: удаление неполных данных (без привязки к городу и/или адресу); связывание данных; прогнозирование потенциальной недостоверности данных.

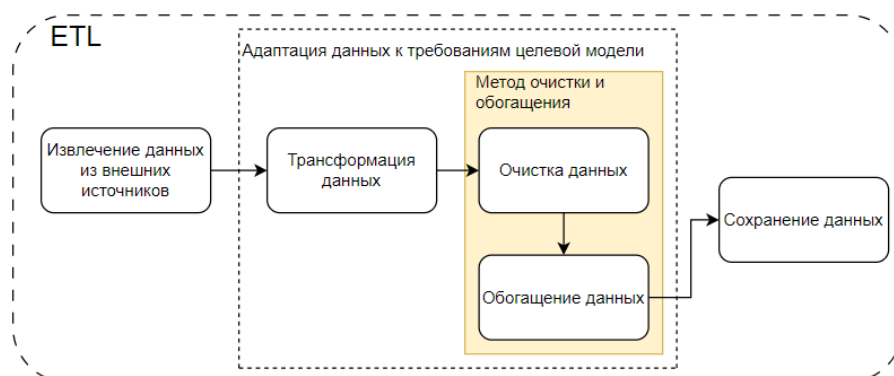


Рисунок 3.1 – Разработанный метод на схеме процесса ETL

Общая схема разработанного метода представлена на Рисунке 3.2.

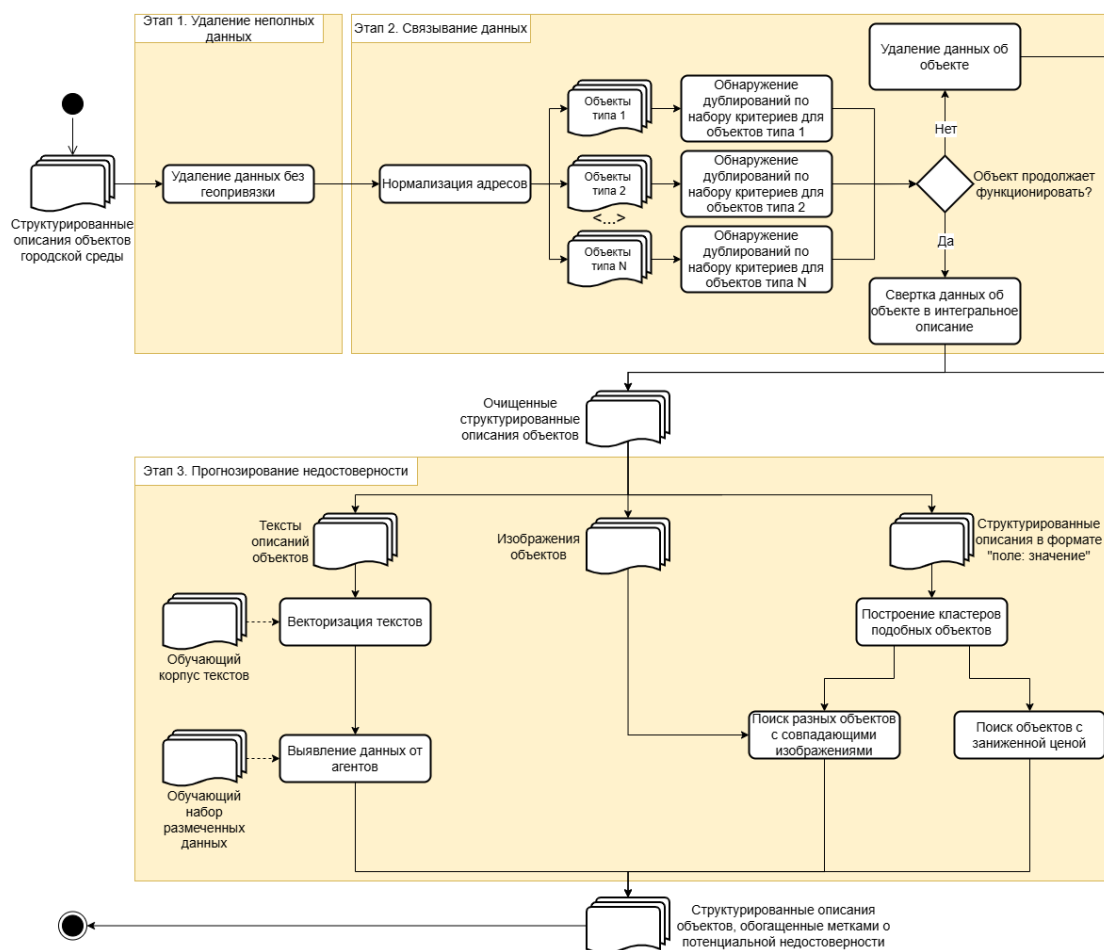


Рисунок 3.2 – Метод очистки и обогащения данных

3.1. Связывание данных

Т.к. разные источники могут содержать разный объем сведений об объекте городской среды, актуально требование: связывание данных не должно приводить к их потере. Т.е. задача связывания состояла в определении дублирования с последующим устранением через свёртку в единое интегральное описание объекта. Исключение – случаи, когда свёртка нецелесообразна по далее описанным причинам. В таком случае данные подлежат очистке (удалению данных об объекте).

Для решения перечисленных задач был разработан алгоритм связывания данных, представленный на Рисунке 3.3.

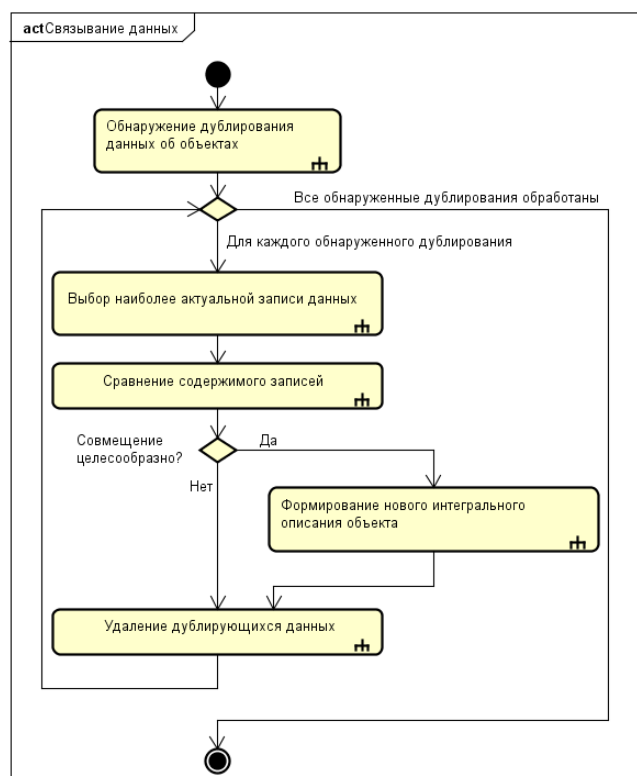


Рисунок 3.3 – Алгоритм связывания данных

3.1.1 Обнаружение дублирования данных

Обнаружение дублирования подразумевает сравнение записей данных по формальным признакам, позволяющим сделать вывод, что они описывают один и тот же объект городской среды. Задача определения множества критериев дублирования сводится к определению характеристик объекта городской среды, способных однозначно идентифицировать его в пределах города. Такие характеристики варьируются в зависимости от типа объекта городской среды (см. Рисунок 3.4).

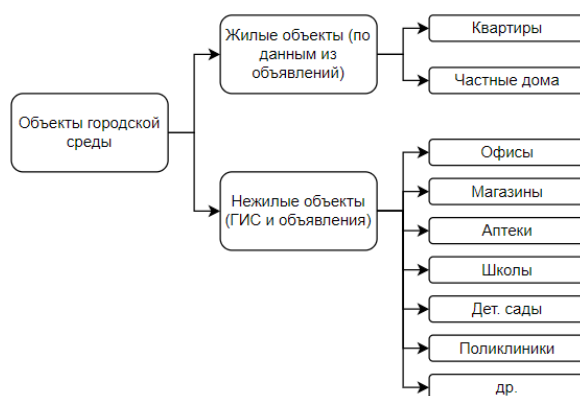


Рисунок 3.4 – Типы объектов городской среды с точки зрения критериев обнаружения дублирования данных

Для идентификации отдельно стоящих зданий (в том числе жилых домов) в пределах города достаточно их адресов (в единой нормальной форме). Совпадение в нескольких записях типа объекта и адреса является достаточным основанием для вывода о дублировании данных.

Особенностью нежилых объектов городской среды является многообразие их форм (см. Рисунок 3.4). Такой объект может быть представлен отдельно стоящим зданием или располагаться внутри другого здания, в котором могут находиться несколько объектов одного типа. Следовательно, необходимым и достаточным условием дублирования данных о нежилых объектах городской среды является совпадение города, адреса, типа объекта и его названия (в нормальной форме), т.к. вероятность нахождения по одному адресу двух однотипных объектов с идентичными названиями стремится к нулю.

Удаление данных о нежилом объекте городской среды целесообразно, если из множества записей данных о нем наиболее актуальная несет сведения о полном прекращении работы (статус «больше не работает» или т.п.). В остальных случаях применяется операция свёртки в интегральное описание. Сравнение актуальности производится на основе метаданных: чем позже дата последнего обновления данных в источнике, тем актуальнее данные.

Специфика квартир в многоквартирных жилых домах такова, что по одному адресу находится множество близких по конфигурации и состоянию объектов,

для которых уникальным (в пределах дома) идентификатором является номер квартиры. Однако в объявлениях адрес квартир указывается с точностью до дома, что делает использование номера невозможным. Остаются актуальными критерии совпадения адреса и типа объекта, позволяющие перейти от поиска дублирующих описаний на всем множестве к работе с группами данных о квартирах, расположенных по одному адресу.

Необходимо было задаться перечнем характеристик квартир, совпадение которых позволит говорить о дублировании с наибольшей вероятностью. Принимая во внимание характеристики квартир, заложенные на уровне архитектурно-планировочных решений, и особенности объявлений, для квартир критериями дублирования были выбраны:

1. Совпадение типа объекта.
2. Совпадение адреса (в нормальной форме).
3. Совпадение общей площади квартир.
4. Совпадение количества комнат.
5. Совпадение этажа расположения.
6. Наличие совпадающих фотографий (с одинаковыми хэшами).
7. Полное совпадение текстов описаний.

Две записи данных о квартирах признаются дублирующими друг друга, если соответствуют одновременно критериям 1-6. Последний критерий является самодостаточным, т.к. полная копия текста возможна только при размещении сведений об одном и том же объекте на разных площадках. Однако использовать его целесообразнее совместно с критериями 1-5 из соображений оптимизации: сравнение текстов только между уже подобными объектами быстрее, чем сравнение каждого с каждым.

Критерий 6 был ограничен совпадением порогового количества изображений в виду отсутствия гарантий, что к объявлениям на разных площадках будет приложено одинаковое количество фотографий и/или что при равном количестве фотографий они будут полностью совпадать.

Итоговый перечень критериев обнаружения дублирования данных приведен в Таблице 3.1.

Таблица 3.1 - Критерии обнаружения дублирования данных

Частные жилые дома	Нежилые объекты	Квартиры в многоквартирных домах
Полностью совпадают: 1. Тип объекта (жилой дом). 2. Город. 3. Адрес (в нормальной форме – см. далее).	Полностью совпадают: 1. Тип объекта (аптека, магазин, школа и др.) 2. Город. 3. Адрес (в нормальной форме – см. далее). 4. Название (в нормальной форме).	Полностью совпадают: 1. Тип объекта (квартира). 2. Город. 3. Адрес (в нормальной форме – см. далее). 4. Общая площадь. 5. Число комнат. 6. Этаж. 7. Тексты объявлений Есть совпадающие: 8. Фотографии.

Схема алгоритма обнаружения дублирующихся данных о квартирах представлена на Рисунке 3.5.

Для сравнения адресов было решено представлять адреса в единой нормальной форме дескриптора – кортежа ключевых составляющих вида:

$$Addr = \langle SD, SN, HN, HL, HC, HB, O \rangle, \quad (3.1)$$

где SD – обозначение улицы (ключевые слова «улица», «проспект» и пр. в полной грамматической форме);

SN – значимые ключевые слова названия улицы в нормальной форме;

HN – номер дома;

HL – литера дома;

HC – корпус дома;

HB – строение;

O – владение (для земельных участков).

Адреса считаются одинаковыми при совпадении дескрипторов.

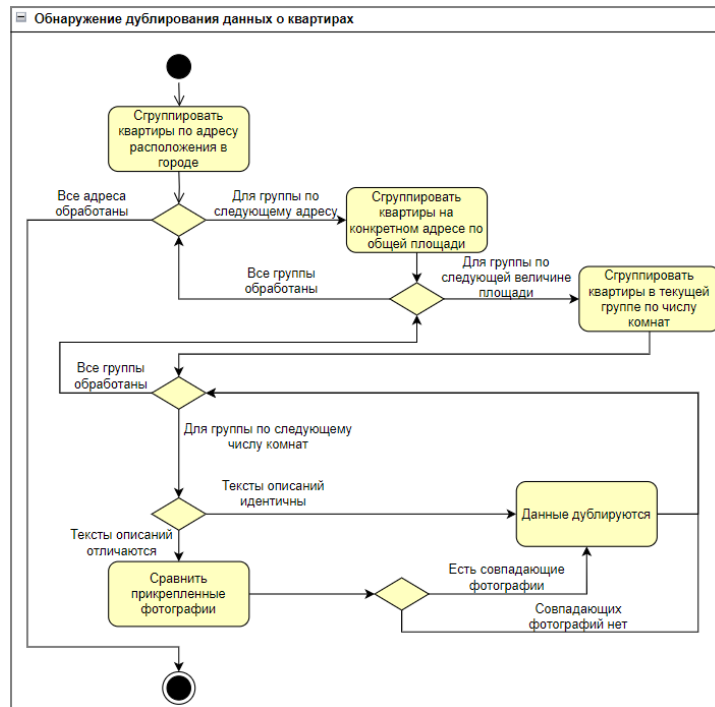


Рисунок 3.5 – Алгоритм обнаружения дублирования данных о квартирах

3.1.2 Устранение дублирования данных

Задачей свёртки дублирующихся данных является совмещение нескольких структурированных описаний одного объекта городской среды к единому интегральному структурированному описанию (см. Рисунок 3.6).

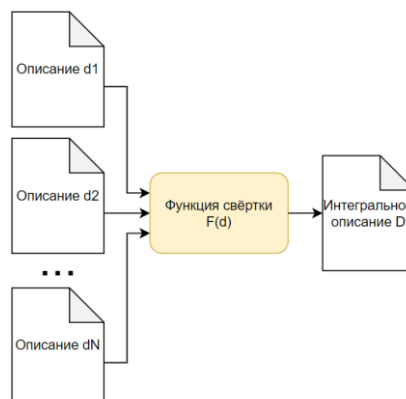


Рисунок 3.6 – Свёртка дублирующихся данных

Функция свёртки F структурированных описаний имеет вид оператора объединения $\cup$, переопределенного для работы со структурированными описаниями объектов городской среды:

$$F(d) = \cup_{i=1}^N d_i = D, \quad (3.2)$$

где D – итоговое интегральное описание объекта городской среды;

d_i – описание объекта из i -го источника;

N – мощность множества источников, из которых была получена информация о данном объекте.

Алгоритм свёртки (см. Рисунок 3.7) сводится к обогащению наиболее актуальной (по метаданным) записи данных: пропуски в ней заполняются наиболее актуальными значениями из прочих записей. Полученное интегральное описание заменяет прочие данные об объекте городской среды.

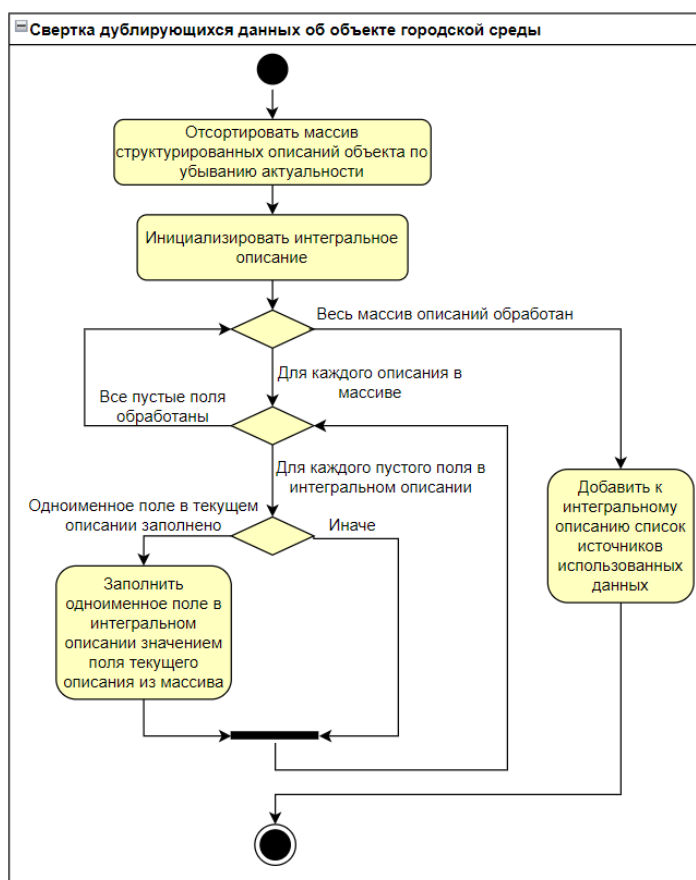


Рисунок 3.7 – Алгоритм свёртки дублирующихся данных

Данные об объектах, прекративших работу, свертке не подвергаются, и подлежат удалению из дальнейшего анализа для обеспечения достоверности результатов.

3.2. Прогнозирование достоверности данных

Потенциально недостоверными следует считать данные из:

- агентских объявлений (от риелторов и риелторских агентств);
- объявлений с заниженными ценами среди аналогичных объектов;
- множественных объявлений от одного автора;
- объявлений с одинаковыми фото, описывающих разные объекты.

Множество недостоверных данных является нечетким: нельзя говорить об однозначной принадлежности к нему конкретной записи данных, лишь о вероятности, выраженной функцией принадлежности [96]. В контексте метода функция была аппроксимирована до двоичной, что позволило свести задачу к двоичной классификации [97].

Однако из-за описанных факторов, удалять выявленные потенциально недостоверные данные из дальнейшего анализа нецелесообразно в виду отсутствия на то однозначных оснований. Альтернативой удалению является отметка принадлежности к множеству потенциально недостоверных с последующим учетом метки на этапе оценки состояния объекта мониторинга.

Таким образом, прогнозирование потенциальной недостоверности решает задачу обогащения данных путем дополнения их двоичными метками о принадлежности к множеству потенциально недостоверных.

3.2.1 Прогнозирование принадлежности данных агентам

Если автором объявления, ставшего источником данных об объекте городской среды, является агент, возможно завышение характеристик относительно реальных либо описание объекта, не существующего в принципе, в маркетинговых целях. Утверждать, что любые объявления агентов недостоверны, нельзя, но и однозначно подтвердить или опровергнуть недостоверность только по открытым данным также невозможно. Возникает задача автоматического разделения записей данных об объектах городской среды, сформированных на основе объявлений, на два множества: данные от агентов и от рядовых пользователей.

Существующие подходы к выявлению недостоверной информации в сети Интернет основаны на возложении ответственности за достоверность на площадку публикации. В качестве методов контроля предлагается применять человеко-ориентированные подходы на основе косвенных критериев с привлечением дополнительных ресурсов и внешних данных [98]. Когда речь заходит о слабо формализуемых подходах на базе неявных и/или нечетких зависимостей, предлагается применение методов машинного обучения [99]. Для перекрестной сверки различных открытых источников исследователи применяют методы парсинга веб-страниц [100], работы с естественными языками методами text mining [101], а для анализа изображений, прилагаемых к объявлениям – нейронных сетей [100] и хэш-алгоритмов [79] с целью выявить дубликаты. Однако данные методы не способны определить потенциально недостоверные данные в отдельно взятых источниках без привлечения дополнительных ресурсов и данных.

При выявлении недостоверных новостей (англ. fake news) в сети Интернет различают методы на основе содержания данных (англ. content-based), и социально-ориентированные (англ. social-based) методы, в основе которых лежит анализ распространения новости, информации об авторе и т.д. [102]. Вторые

слабо применимы к решаемой задаче, т.к. анализируемые ими критерии применительно к объявлениям малозначимы [76]. Первые же методы отталкиваются от текстов, их стилистических особенностей и смысловой нагрузки, а суть их сводится к решению задач классификации методами машинного обучения и обработки естественных языков [103, 104].

Таким образом, задача сводится к разработке метода анализа данных на основе содержания. При этом объявления на некоторых площадках уже отмечены принадлежностью к одному из множеств, что позволяет свести задачу к задаче двоичной классификации из области машинного обучения [76, 97], решаемой с помощью соответствующих моделей [76, 97, 105], обученных на основе собранных структурированных данных.

В составе данных для обучения модели обязательно присутствует поле, отражающее тип автора (принадлежность объявления физическим лицам либо агентам) и являющееся целевым параметром модели. Поля характеристик объектов городской среды не могут быть признаками классификации, т.к. не несут информации об авторе. Однако автор может проявить себя в исходном тексте объявления. Предпосылкой гипотезы является общее структурное и стилистическое подобие объявлений от агентов [76]. Таким образом, целевым параметром модели является поле «Тип автора», а функцию входных параметров выполняет текстовое описание объекта из объявления. Решаемая задача является задачей классификации текстов [106].

Для обучения модели тексты подлежат преобразованию в векторы признаков с обязательным сохранением семантической и стилистической информации. Предварительно осуществляется нормирование и преобразование данных: значения целевого параметра нормируются до двоичного представления, где значение «0» означает принадлежность к множеству не-агентских, а значение «1» – к множеству агентских данных. Перечень возможных первичных значений («Частное лицо», «Агентство», и пр.), а также принцип их нормирования зависит от обучающего набора.

Ключевым для построения модели является выбор способа преобразования текстов в векторы признаков. Предварительно тексты очищаются от стоп-слов, а значимые слова приводятся к нормальной форме [76]. Результат – цепочки только значимых по смысловой нагрузке слов в нормальной форме, которые затем преобразуются в векторы признаков с помощью модели Word2Vec [76, 107, 108], обученной на корпусе нормализованных текстов из объявлений об операциях с недвижимым имуществом, а также корпусе русскоязычных текстов [76]. В качестве входного вектора признаков для обучения модели используется средний вектор значимых слов каждого объявления. Общая схема решения задачи представлена на Рисунке 3.8.

Целевая структура обучающего набора представлена в Таблице 3.2.



Рисунок 3.8 – Выявление потенциально недостоверных данных

Таблица 3.2 – Описание структуры обучающих данных

Название поля	Возможные значения	Описание
Тип автора	0 – не-агент 1 – агент	Указывает на принадлежность конкретному классу, целевой параметр
Описание	Вектор	Средний вектор значимых слов текстового описания, входной параметр

3.2.2 Выявление «сомнительных» данных

Общепринятой практикой является оценка рыночной стоимости недвижимого имущества сравнительным подходом, подразумевающим

вычисление стоимости некоторого объекта X на основе стоимостей подобных ему объектов [109]. Т.е. подобные объекты, размещенные на рынке в один период времени, имеют сопоставимые цены.

В объявлениях возможно завышение характеристик объекта при сохранении стоимости либо занижение цены при высоких характеристиках. Данные из таких объявлений потенциально недостоверны (сомнительны), поскольку описанные действия могут быть связаны с мошенничеством, маркетинговыми уловками и т.п. Подобие объектов устанавливается по их ключевым характеристикам: общей площади, этажности дома, этажа расположения, района города или др. [109, 110]. Формальных метрик подобия не существует, а подбор аналогичных объектов осуществляется приглашенными экспертами.

Возникает задача определения на заданном множестве объектов групп объектов, объединенных неявной метрикой подобия, для сравнения их цен. Данная формулировка задачи соответствует задаче кластеризации из области машинного обучения [111]. Причем количество кластеров заранее неизвестно, что затрудняет применение большинства методов кластерного анализа. В контексте исследования целесообразным является применение восходящей иерархической кластеризации, результатом которой является бинарное дерево, где в листьях находятся отдельно взятые элементы множества, а корнем дерева является кластер (группа), объединяющий в себе совокупность всех элементов [112].

Преимуществом иерархической кластеризации является возможность определить нужное количество кластеров, исследуя свойства полученного дерева. Например, путём выделения в разные группы тех поддеревьев, расстояния между которыми достаточно велики [112].

Каждый элемент анализируемого множества объектов городской среды представляется вектором характеристик (признаков) объекта для подсчета расстояний между ними. Однако разные признаки имеют разные измерительные шкалы, что приводит внесению одними признаками большего вклада в расстояние между объектами, чем другими [111].

Решением, обеспечивающим равновесность всех признаков, является стандартизация данных, что обеспечивают их корректное применение в многомерном анализе [113]. Для стандартизации производится преобразование каждого признака таким образом, чтобы его среднее значение было равно нулю, а стандартное отклонение от среднего — единице, что приводит данные к стандартному нормальному распределению.

После построения оптимального числа кластеров на множестве объектов одного типа (квартир, домов, офисов и пр.) производится статистический анализ цен в каждом кластере с целью поиска аномальных значений – в данном случае заниженных цен. В рамках исследования обнаружение выбросов было решено производить методом межквартильного размаха [113] с учетом только нижней границы, т.к. производился поиск именно заниженных значений (см. Рисунок 3.9). Случаи завышения цен в рамках исследования было решено подозрительными не считать, т.к. завышение цен в контексте мошеннических и маркетинговых уловок не применяется. Поиск аномалий производится в пределах элементов кластера, данные о которых получены из объявлений не старше полугода от даты анализа для нивелирования макроэкономических и подобных факторов, влияющих на цены. Также учтены разные операции с недвижимым имуществом (продажа, аренда), влияющие на цену: поиск аномальных значений производится на множествах объектов из объявлений с одинаковой операцией.

Данные из множественных (более 5) объявлений, размещенных одним пользователем в одном источнике, тоже отмечаются как потенциально недостоверные. На каждой из площадок объявления имеют привязку к автору: контактный номер телефона, идентификатор пользователя или т.п., на основе которых и производится оценка количества объявлений автора.

Ещё одним признаком недостоверности является использование повторяющихся изображений (фотографий), которые встречались в описаниях разных объектов. Разными было решено считать объекты, данные о которых не подверглись свёртке на этапе устранения дублирования. Алгоритм выявления дубликатов изображений с учетом противодействия возможным искажениям

представлен на Рисунке 3.10. Принцип выявления «сомнительных» данных по совпадению изображений иллюстрируется на Рисунке 3.11.

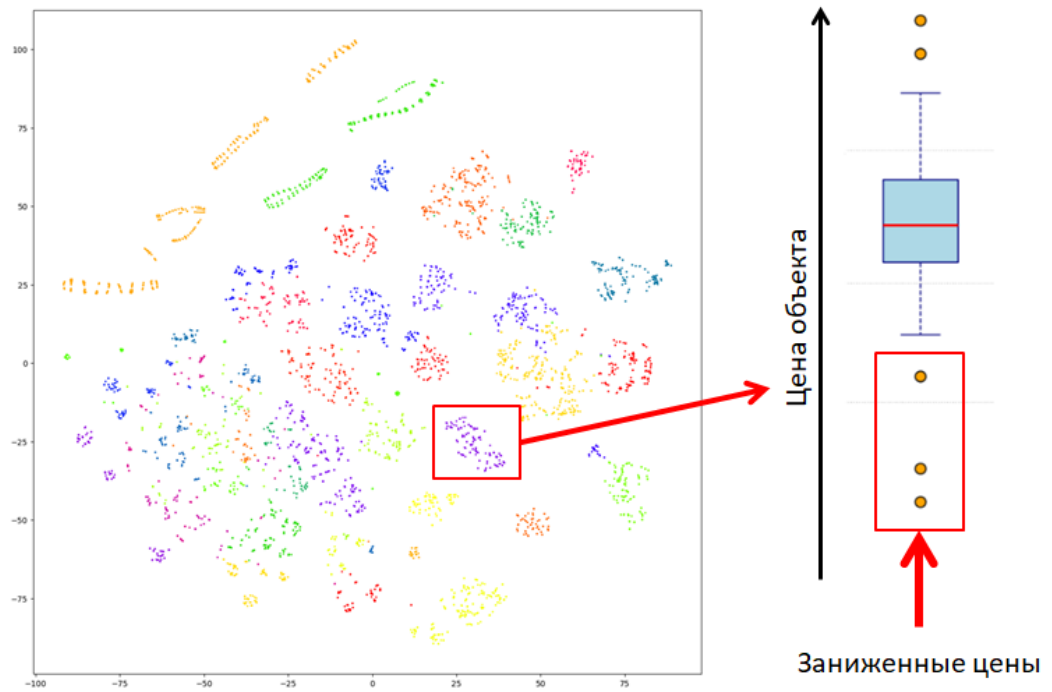


Рисунок 3.9 – Выявление «сомнительных» данных по заниженным ценам

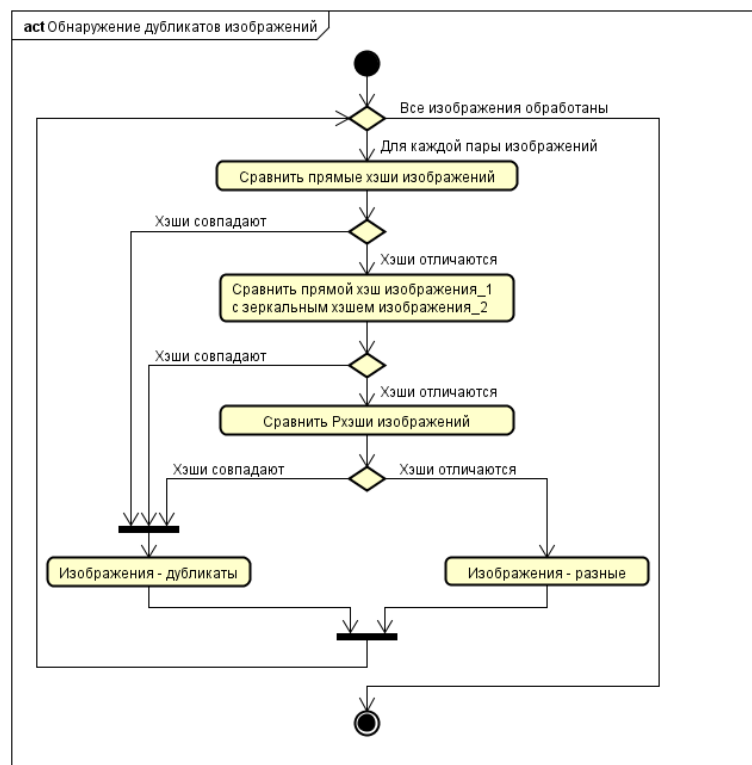


Рисунок 3.10 – Алгоритм выявления дубликатов изображений

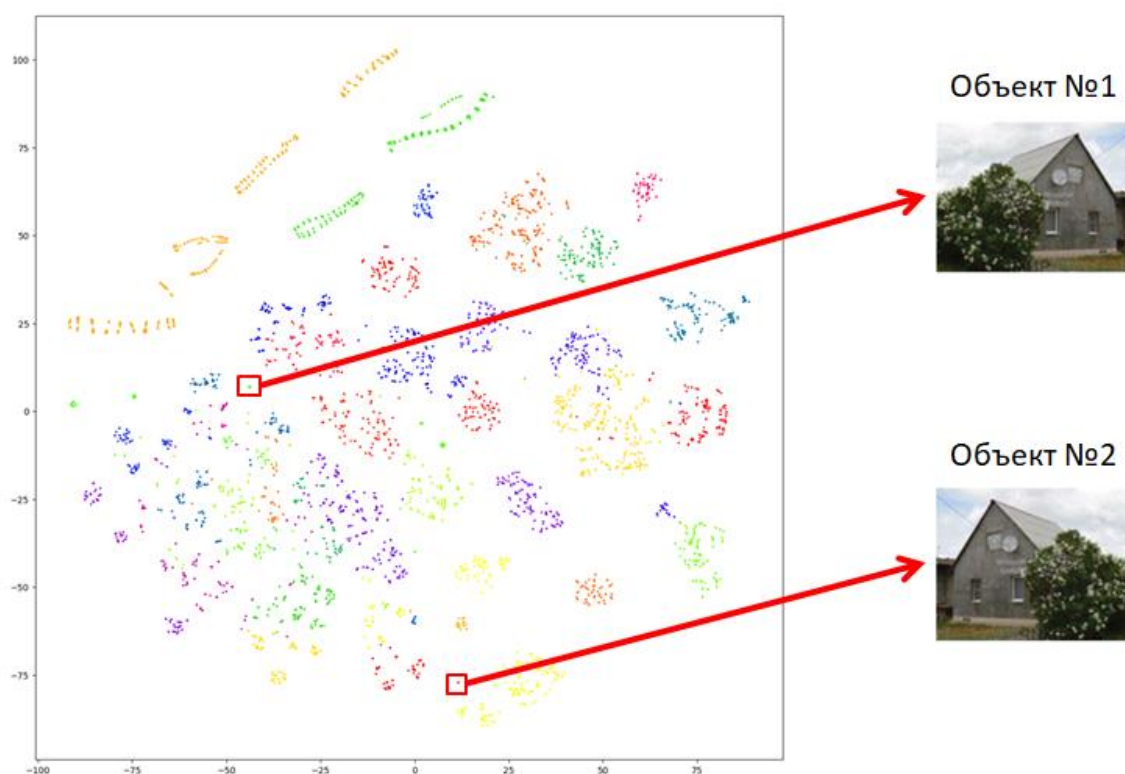


Рисунок 3.11 – Выявление «сомнительных» данных по совпадению изображений

Если по результатам любой из описанных проверок были получены результаты, говорящие о потенциальной недостоверности, запись данных об объекте помечаются как потенциально недостоверная для последующего учета метки в процессе оценки состояния объекта мониторинга.

3.3. Выводы по третьей главе

1. Разработан метод очистки и обогащения данных (см. Рисунок 3.2), включающий следующие основные этапы: удаление неполных данных. связывание данных (устранение дублирований в данных); прогнозирование недостоверности данных (данных от агентов и данных, в отношении которых есть

основания заподозрить недостоверность по признакам занижения цен и совпадения изображений).

2. Разработан принцип обнаружения дублирований в данных об объектах городской среды по формальным признакам, список которых определен для различных типов объектов (см. Таблицу 3.1)

3. Описаны два способа устранения обнаруженных случаев дублирования данных: на основе свёртки путём обогащения наиболее актуальной записи данных об объекте данными менее актуальных записей о нем и на основе удаления данных из дальнейшего анализа в случае, если свёртка нецелесообразна.

4. Разработаны методы прогнозирования недостоверности данных на основе методов машинного обучения (классификации и кластеризации), а также методов статистического исследования данных и сравнения изображений с помощью ансамбля методов перцептивного хэширования (см. Рисунки 3.8-3.11).

Конечной целью описанных в Главе 3 решений является получение наиболее полных, достоверных и актуальных данных об объектах городской среды для дальнейшего анализа.

ГЛАВА 4. Разработка алгоритмов оценки состояния

Формальная постановка задачи оценки E состояния сложной пространственно распределенной системы применительно к оценке состояния городской среды по ранее неучтенным показателям состояния имеет следующий вид:

$$E = \langle L, C, I, F(L), F(C), F(I) \rangle, \quad (4.1)$$

где L – множество жилых объектов городской среды;

C – множество коммерческих объектов городской среды;

I – множество объектов социальной инфраструктуры;

$F(L)$ – функция оценки потребительских свойств жилых объектов;

$F(C)$ – функция оценки потребительских свойств коммерческих объектов;

$F(I)$ – функция оценки обеспеченности социальной инфраструктурой (образованием и здравоохранением).

Говоря об оценке состояния сложной пространственно распределенной системы, корректно дополнить (5.1) функцией интегральной оценки на основе $F(L)$, $F(C)$ и $F(I)$. Однако, совершенствование потребительских свойств жилых объектов, а также появление новых объектов может привести к увеличению нагрузки на объекты социальной инфраструктуры вплоть до перегрузки (нештатного состояния). Поэтому многокритериальная оценка состояния городской среды без интегральной свёртки целесообразна и оправдана. Частные оценки позволят выявить сильные и слабые стороны объекта мониторинга, а также отдельные его элементы, требующие корректирующего воздействия.

Таким образом, разработка алгоритмов оценки состояния сложной пространственно распределенной системы (городской среды) по выделенным ранее неучтенным показателям сводится к разработке алгоритмов оценки для реализации функций $F(L)$, $F(C)$ и $F(I)$. Необходимо отметить, что функции $F(L)$ и $F(C)$ оценивают одну и ту же характеристику объектов городской среды – их потребительских свойств, на которые влияют характеристики внутреннего

состояния и окружения объектов, т.е. алгоритмы $F(L)$ и $F(C)$ имеют одинаковый вид.

Задача же данных алгоритмов сводится к получению численных оценок по выделенным в настоящем исследовании неучтенным показателям состояния городской среды на основе данных, получение которых описано в предыдущих Главах.

4.1 Разработка алгоритма оценки показателя потребительских свойств

4.1.1 Выбор метода оценки состояния

К настоящему моменту разработано множество методов, так или иначе дающих характеристику уровня потребительских свойств объектов городской среды.

Примерами являются Единая классификация многоквартирных жилых новостроек [114] или Классификация офисных помещений [115]. В основе – разделение объектов на классы и подклассы на основе их потребительских свойств, выраженных характеристиками внутреннего состояния (ремонт, качество отделки, и др.), а также окружения и инфраструктуры: остановок общественного транспорта, парковок, магазинов и т.п. Недостатком методов является потребность в информации, доступной только при непосредственном осмотре объектов [116], что делает их непригодными для задач исследования.

Наиболее распространенным примером методов прямой оценки пользователями являются системы подсчета рейтинга товаров и услуг на основе отзывов. Примеры зарубежных решений: Apartments for Rent [117] или Apartments Rating [118]. Однако данный подход субъективен и непригоден для целей исследования.

Еще одним примером являются квалитетрические методы. Основной задачей квалитетрии является разработка методик оценки качества конкретного вида объектов либо процессов числом, характеризующим степень их соответствия заданным требованиям [119]. Понятие «качество» в квалитетрии трактуется как определенная совокупность полезных свойств, подразделяемых по иерархическим уровням [119]. В случае с жилыми объектами качество на верхнем уровне характеризуется сочетанием функциональных характеристик и эстетичности облика, удовлетворяющих потребности жителей. На более низком уровне функциональные характеристики разделяются на свойства, характеризующие удобство и надежность объекта. Далее характеристики удобства рассматриваются как совокупность более конкретных, непосредственно измеримых свойств: площади помещений, наличия теплоизоляции и т.п. [119].

Для разработки квалитетрического алгоритма оценки потребительских свойств жилых и коммерческих объектов городской среды требуется решение следующих задач:

- определить положения и ограничения применения алгоритма;
- выбрать метод декомпозиции качества (в контексте исследования – оценки потребительских свойств) объекта на частные показатели;
- определить методы учета степени влияния частных показателей на итоговый показатель качества исследуемого объекта;
- установить методы количественной оценки частных показателей;
- определить методы объединения частных показателей в интегральный [120].

Наиболее наглядной для человека и широко представленной оценкой потребительских свойств жилых и коммерческих объектов городской среды является их рыночная цена: чем качественнее и привлекательнее для потенциального потребителя объект, т.е. чем лучшей совокупностью потребительских свойств он обладает, тем выше его стоимость на рынке (при условии адекватной оценки). В качестве основных характеристик, влияющих на

итоговые показатели потребительских свойств и итоговую рыночную стоимость, а также применяемых подходов к оценке, обычно выделяют:

- оценку характеристик самого объекта и здания, в котором он расположен (количество комнат, общая площадь, наличие и количество лифтов, состояние ремонта объекта и здания и т.п.) [56];
- оценку характеристик окружения (наличие магазинов, аптек, медицинских учреждений, остановок общественного транспорта в пешей доступности и т.п.) [56].

Для подобной оценки существуют следующие подходы:

- сравнительный – оценка стоимости на основе сравнения аналогичных с подобными объектами [121];
- затратный – метод на основе оценки затрат на восстановление либо замещение объекта аналогичным по функциям [122];
- доходный – оценка на основе ожидаемых доходов [122].

Однако для применения перечисленных подходов необходима специализированная подготовка и знания, а полная их формализация, необходимая для машинной реализации, затруднена эмпиричностью, вытекающей из ориентированности на реализацию походов человеком (экспертом в предметной области). Кроме того, в рамках исследования ставится задача оценки потребительских свойств именно на основе открытых данных для последующей реализации в составе системы мониторинга состояния городской среды. Перечисленные же подходы требуют привлечения значительных массивов внешних закрытых данных, к объектам городской среды напрямую не относящихся и требующих дополнительных процедур сбора с привлечением человека, т.е. обладающих непредсказуемой степенью сложности и трудоемкости.

Однако в рамках данных подходов следует отметить существующий механизм корректировок, основанный на опыте экспертов в определении рыночной стоимости объектов городской среды относительно влияния на неё отдельных характеристик в сравнении с подобными по свойствам объектами [56]. Главным элементом механизма являются выведенные с помощью статистики

численные коэффициенты, задающие отношение цен двух полностью аналогичных объектов с одной отличающейся характеристикой. Например, цена квартиры на первом этаже дома составляет 0,93 от цены аналогичной квартиры на одном из срединных этажей [56].

Наиболее инновационными являются информационные платформы на базе технологии AVM (англ. Automated Valuation Model - «Автоматизированная модель оценки») [123]. В основе лежит математическое моделирование с использованием данных о существующих объектах и истории операций с ними для расчета рыночной стоимости других объектов, жилых и коммерческих [124]. Ключевой особенностью всех систем такого рода является отсутствие трансграничности их предложения, что связано с особенностями законодательства и информационных баз разных стран, а тонкостями местных рынков. Недостатком подхода является оценка на основе условного «среднего» состояния объекта, что далеко не всегда соответствует реальности [124].

Можно выделить несколько существующих и функционирующих в настоящее время реализаций AVM [123]:

- HPIs (House Price Indices) или индексы цен на жилье;
- однокритериальная оценка (англ. Single Parameter Valuation);
- гедонические модели (англ. Hedonic Models).

В основе HPIs лежат временные ряды данных, фиксирующие изменения цен объектов городской среды каждого типа в выделенной географической области [123]. В рамках исследования такие модели неприменимы, поскольку оценку необходимо производить на основе открытых данных, а исторические данные о свойствах объектов за предыдущие периоды не оказывают влияния на их свойства в момент оценки.

Однокритериальные методы оценивают стоимость объекта на базе одной отдельно взятой характеристики. Недостатком является грубость утверждения, что одна отдельно взятая характеристика объекта даст достаточно данных для оценки. Как правило, подобные модели используются в качестве «отправной точки» для дальнейших модификаций [123].

Гедонические модели – это математические выражения с предварительно вычисленными параметрами. В качестве входных данных они принимают множество значений характеристик оцениваемого объекта для расчета итоговой стоимости. Именно такие модели чаще всего обозначают термином «AVM». Они описывают, как каждая характеристика влияет на итоговую стоимость объекта в заданный период времени и в заданной географической области. При этом подход не привязан к конкретным математическим моделям, что оставляет разработчику свободу для выбора исходя из решаемой задачи [123].

Таким образом, для решения задачи оценки потребительских свойств жилых и коммерческих объектов городской среды наиболее подходящей является гедоническая AVM на базе методов квалиметрии. Входными параметрами модели являются значения характеристик объектов. Также целесообразно применение механизма корректировок [56].

4.1.2 Постановка задачи оценки состояния

Задачу оценки потребительских свойств жилых и коммерческих объектов городской среды представим как задачу ранжирования объектов по показателю качества, выраженного оценкой потребительских свойств. Формальная постановка задачи имеет следующий вид [125]:

$$D = \langle F, A, X, G, P \rangle, \quad (4.2)$$

где F – формулировка задачи, состоящая в ранжировании вариантов выбора (объектов городской среды) по оценкам их потребительских свойств;

A – множество жилых и коммерческих объектов городской среды (вариантов выбора);

X – совокупность атрибутов, описывающих характеристики вариантов выбора;

G – совокупность условий, ограничивающих область допустимых решений

(в контексте задачи – пустое множество);

P – предпочтения, определяющие принцип оценки и сравнения качества вариантов выбора.

В контексте задачи предпочтения определяют значимость каждой характеристики из множества X при ранжировании объектов и функциональную модель для оценки и свёртки оценок отдельных атрибутов (частных критериев) с целью получения итогового ранга (оценки) объектов, т.е. целевую функцию.

Таким образом, имеем постановку задачи многокритериальной оценки предпочтительности вариантов выбора, в роли которых выступают жилые и коммерческие объекты городской среды.

В отличие от задачи многокритериальной оценки состояния городской среды в целом, при оценке потребительских свойств отдельных объектов целесообразным является расчет интегральной оценки, вычисляемой на основе нескольких частных оценок по отдельным критериям (атрибутам) [125].

Обозначим целевую функцию как $F(X_f)$, где:

$$F(X_f) \in P, \quad (4.3)$$

$$X_f = \{f_1(x_1), f_2(x_2), \dots, f_n(x_N)\}, \quad (4.4)$$

где X_f – вектор оценок объекта по каждому атрибуту (частному критерию);

$f_i(x_i)$ – функция оценки для i -го атрибута, где $i = \{1, 2, \dots, N\}$;

x_i – значение i -го атрибута, где $i = \{1, 2, \dots, N\}$;

N – мощность множества атрибутов X .

Как следует из (4.4), каждому атрибуту из множества X соответствует функция оценки $f_i(x_i)$, преобразующая его значение к числовой оценке потребительских свойств по данному отдельно взятому атрибуту (частному критерию).

Значения каждого атрибута x_i измеряются по некоторой шкале. В контексте исследования было принято решение отталкиваться от естественных шкал измерения каждого атрибута и соответствующих единиц измерения. В качестве

примера в Таблице 4.1 перечислены возможные типы шкал атрибутов из множества A , применяемых при оценке потребительских свойств квартир.

Таблица 4.1 – Примеры шкал для оценки атрибутов

Атрибут	Вид атрибута	Тип шкалы	Единица измерения\возможные значения
Площадь квартиры	Количественный	Абсолютная (количественная)	м ²
Наличие лифта	Качественный (двоичный)	Номинальная (двоичная)	Есть\Нет
Ремонт	Качественный	Номинальная	Отсутствует\ Косметический\ Евроремонт

Различные шкалы имеют разные принцип и единицы измерения, что привело к проблеме агрегирования оценок для вычисления $F(X_f)$ и необходимости нормирования [125]. Важно было, чтобы нормированные оценки имели один порядок величины, иначе при агрегировании между частными оценками возникало отношение неявного превалирования, не входящее во множество предпочтений P . Для решения задачи нормирования в контексте количественных атрибутов было использовано нормирование на максимум шкалы. Таким образом, функция $f_i(x_i)$ приняла следующий вид:

$$f_i(x_i) = \frac{s(x_i)}{s_{\max}(x_i)} * 10, \quad (4.5)$$

где $s(x_i)$ – оценка по атрибуту x_i (измерение по соответствующей шкале);

$s_{\max}(x_i)$ – максимально возможная оценка (максимум шкалы) атрибута x_i .

Вычисление по (4.5) дает безразмерную оценку в пределах от 0 до 10. Было решено задать для каждой шкалы конечный размер, определяющий некоторый достаточный в контексте задачи и предпочтений уровень значений $s_{\max}(x_i)$. Оценки, превышающие заданный пороговый уровень, было решено нормировать до максимальных $f_i(x_i) = 10$. Формула (4.5) приняла следующий вид:

$$f_i(x_i) = \min\left(\frac{s(x_i)}{s_{\max}(x_i)} * 10; 10\right). \quad (4.6)$$

Для номинальных двоичных шкал каждому значению атрибута ставится в соответствие оценка 0 либо 1. Для номинальных шкал большей размерности (качественных атрибутов с 3 и более значениями) выстраивается градация значений атрибута по их предпочтительности, после чего шкала с размахом от 0 до 1 делится на равные сегменты по количеству возможных значений. Каждому значению атрибута ставится в соответствие оценка с соответствующего сегмента. Альтернативно возможно явное присвоение оценок каждому значению таких атрибутов. Дальнейшее нормирование оценок выполняется по формуле (4.6).

4.1.3 Разработка алгоритма вычисления интегральной оценки состояния

Данные, используемые в настоящем исследовании для оценки потребительских свойств жилых и коммерческих объектов городской среды, обладают четким характером, что позволило отказаться от методов на основе теории нечетких множеств.

Были выделены следующие требования к алгоритму вычисления интегральных оценок потребительских свойств:

- возможность четкой прозрачной интерпретации результатов: пользователь системы мониторинга должен иметь возможность явно проследить влияние каждого частного критерия на интегральную оценку и истолковать полученные значения в контексте предметной области;
- возможность явно определять, корректировать и четко интерпретировать без дополнительных затрат ресурсов значимость частных критериев в составе интегральной оценки;
- возможность точечной корректировки расчета без дополнительных затрат ресурсов, например, при изменении значимости отдельных частных критериев по сравнению с другими;

- масштабируемость: возможность расширения перечня критериев без ущерба быстрдействию и увеличения ресурсозатратности;
- возможность использования статистически известных зависимостей между значениями отдельных частных критериев и интегральной оценкой потребительских свойств объекта [56];
- неизменность ранее полученных результатов оценки с течением времени (при неизменности явно заданных параметров расчета).

Отталкиваясь от требований, была выбрана линейная аддитивная свертка [125], и целевая функция $F(X_f)$ приобрела следующий вид:

$$F(X_f) = \sum_{i=1}^N f_i(x_i) = \sum_{i=1}^N \left(\min \left(\frac{s(x_i)}{s_{max}(x_i)} * 10; 10 \right) \right). \quad (4.7)$$

Для получения более наглядной оценки в диапазоне от 0 до 100% полученное значение целевой функции было решено нормировать на максимально возможное $f_{imax}(x_i)$ при заданных предпочтениях и шкалах. Была получена целевая функция $FN(X_f)$ следующего вида:

$$FN(X_f) = \frac{F(X_f)}{F_{max}(X_f)} * 100\% = \frac{\sum_{i=1}^N f_i(x_i)}{\sum_{i=1}^N f_{imax}(x_i)} * 100\%. \quad (4.8)$$

Значение целевой функции $FN(X_f)$ для конкретного объекта городской среды является интегральной оценкой его потребительских свойств.

Целесообразно было вспомнить о ещё одном интегральном показателе потребительских свойств жилых и коммерческих объектов городской среды: их рыночной цене как недвижимого имущества. Среди методов рыночной оценки выделяется ранее упомянутый сравнительный подход [109], за время существования которого были собраны статистические данные, отражающие разницу цен подобных объектов с отдельными отличающимися характеристиками, в виде корректирующих коэффициентов следующего вида [56]:

$$c = P_{A1} / P_{A2}, \quad (4.9)$$

где c – корректирующий коэффициент;

P_{A1} – цена объекта $A1$;

P_{A2} – цена объекта $A2$, полностью аналогичного объекту $A1$ по всем атрибутам, кроме атрибута x_i , причем соблюдается условие $x_{iA1} > x_{iA2}$, т.е. вариант $A2$ однозначно превосходит вариант $A1$ по потребительским свойствам, связанным с атрибутом x_i .

Такие коэффициенты с помощью методов статистики рассчитаны для различных атрибутов различных типов объектов, и опубликованы в формате справочников [56], используемых экспертами на практике.

Т.к. функция $FN(X_f)$ и рыночная цена объекта являются мерилем одной характеристики, совокупной оценки потребительских свойств, определяющих его привлекательность для потенциальных потребителей, была выдвинута гипотеза, что с ростом целевой функции $FN(X_f)$, должна расти и рыночная цена, т.е. оба показателя должны коррелировать. Руководствуясь данной гипотезой и имеющимся множеством корректировок C , формула (5.8) была приведена к виду [126, 127]:

$$FNC(X_f) = \frac{\sum_{i=1}^N f_i(x_i)}{\sum_{i=1}^N f_{imax}(x_i)} * \prod_{j=1}^M c_j * 100\%, \quad (4.10)$$

где M – мощность множества корректировок C ;

c_j – корректирующий коэффициент для j -й корректировки.

Использование корректировок позволило использовать проверенные на практике методы и статистические данные [56] и внести вклад в достоверность результатов без вреда для интерпретируемости.

4.1.4 Выбор метода определения весов частных критериев оценки состояния

Для учета разной значимости частных критериев оценки потребительских свойств, целевая функция $FNC(X_f)$ была модифицирована для учета степени значимости (важности) каждого из частных критериев [128]. Сведения о важности

частных критериев были включены в состав множества предпочтения P . Важность каждого критерия представляет собой числовой коэффициент w_i , отражающий значимость или «вес» i -го частного критерия (атрибута объекта городской среды). Коэффициент w_i должен удовлетворять следующим условиям [127]:

- $w_i \in R$;
- $w_i \in [0, 1]$;
- $\sum_i w_i = 1$.

Существуют различные методы определения весов частных критериев в задаче многокритериальной оценки, как на основе субъективных мнений и предпочтений приглашенных экспертов, так и на основе объективных данных (рыночных показателей, статистики продаж и др.) [125]. Выбор метода зависит от доступных данных, а также от целей оценки.

Объективные методы позволяют определять веса, отталкиваясь от существующего массива объективных данных. В качестве примера был рассмотрен метод энтропии, основанный на анализе изменчивости значений критериев в наборе данных [129]. Подобные принципы определения весов актуальны и для других объективных методов [130, 131]. Их недостатком в контексте исследования является зависимость от данных: при поступлении в систему мониторинга новых данных вновь рассчитанные веса могут отличаться от использованных на предыдущей итерации оценки состояния объекта мониторинга. Описанная ситуация для задач исследования не является корректной, т.к. оценки состояния объекта мониторинга, полученные в разные периоды времени, должны быть сравнимы между собой для прослеживаемости динамики изменения его состояния [14]. Изменение же весов критериев между разными итерациями оценки приведет к нарушению прослеживаемости вследствие неявных изменений в принципах формирования оценок. Если же транслировать изменения весов на прошлые итерации оценки состояния объекта мониторинга, будут изменяться результаты, что также не является корректным.

Был сделан вывод, что для использования в настоящем исследовании подходят методы, позволяющие задать абсолютные (неизменные с течением времени и поступлением новых данных) веса критериев. Такой результат способны дать субъективные методы, которые, однако, требуют привлечения экспертных оценок [125]. Что нельзя считать непреодолимым препятствием, т.к. веса критериев подлежат определению один раз: перед запуском системы мониторинга.

Единственным случаем, когда пересчет весов критериев оправдан – это расширение перечня критериев оценки состояния.

Среди субъективных методов определения весов критериев были выделены следующие [125]:

- метод явного назначения весов;
- метод последовательного сравнения критериев по абсолютной важности;
- метод последовательного сравнения по суммарной важности;
- метод парного сравнения по абсолютной важности;
- метод парного сравнения по относительной важности Т. Саати.

Метод явного назначения весов является наиболее тривиальным при небольшом числе критериев, однако усложняется вместе с ростом их числа, а корректировка весов оказывается затруднительной и трудоемкой из-за назначения весов и контроля их корректности в ручном режиме [125].

Методы последовательного сравнения критериев по абсолютной и суммарной важности являются более прозрачными и допускают более простую корректировку весов, но принцип назначения важности в них не формализован [125], что затрудняет сопровождение системы мониторинга, использующей данные методы.

Наиболее прозрачными и простыми в части внесения корректировок в веса являются методы на основе попарных сравнений важности критериев [125]. В основе лежит построение матрицы $A = (a_{ij})_{h \times h}$ парных сравнений критериев

$f_1, \dots, f_h$ по их важности, в качестве элементов a_{ij} которой указывается предпочтительность i -го критерия по сравнению с некоторым j -м критерием [125]. В случае парных сравнений по абсолютной важности возможны только два значения a_{ij} : $a_{ij} = 0$ означает, что i -й критерий менее важен по сравнению с j -м ($f_i < f_j$), а $a_{ij} = 1$ – наоборот ($f_i > f_j$).

Данный метод обладает достаточной гибкостью и прозрачностью по сравнению с ранее описанными, однако его модификация, метод парных сравнений по относительной важности Т.Саати [43, 125, 132] допускает задавать в качестве значений матрицы a_{ij} не просто факт большей важности одного критерия над другим, но также степень предпочтительности, оперируя при этом строго формализованными значениями, что приводит к большей прозрачности, по сравнению с методом последовательных сравнений.

Возможные значения элементов a_{ji} для матрицы парных сравнений Саати приведены в Таблице 4.2.

Таблица 4.2 – Значения элементов матрицы парных сравнений Саати

a_{ij}	Комментарий
1/9	j -ая альтернатива намного значимее i -ой
1/7	j -ая альтернатива явно значимее i -ой
1/5	j -ая альтернатива значимее i -ой
1/3	j -ая альтернатива немного значимее i -ой
1	Альтернативы примерно равноценны
3	i -ая альтернатива немного значимее j -ой
5	i -ая альтернатива значимее j -ой
7	i -ая альтернатива явно значимее j -ой
9	i -ая альтернатива намного значимее j -ой

Для матрицы парных сравнений Саати обязательно к исполнению правило [43, 132]:

$$a_{ij} = 1/a_{ji}. \quad (4.11)$$

Исходя из условия (4.11), заполнение матриц парных сравнения Саати было ограничено заполнением одного из двух треугольников матрицы относительно

главной диагонали. Второй же треугольник вычисляется на основе первого. Общий вид матрицы парных сравнений Саати приведен в Таблице 4.3.

Таблица 4.3 – Матрица парных сравнений Саати

A	f_1	f_2	...	f_h	a_k
f_1	1	a_{12}	...	a_{1h}	a_1
f_2	$a_{21} = \frac{1}{a_{12}}$	1	...	a_{2h}	a_2
...	...	...	1	...	...
f_h	$a_{h1} = \frac{1}{a_{1h}}$	$a_{h2} = \frac{1}{a_{2h}}$	...	1	a_h

«Ценность» a_i критерия вычисляется по следующей формуле [43, 132]:

$$a_i = \sqrt[h]{\prod_{j=1}^h a_{ij}}, \quad (4.12)$$

где a_{ij} – важность i -го критерия относительно j -ого;

h – мощность множества критериев.

Итоговый вес критерия вычисляется по следующей формуле [43, 132]:

$$w_i = a_i / \sum_{j=1}^h a_j, \quad (4.13)$$

где w_i – итоговый вес i -го критерия;

$\sum_{j=1}^h a_j$ – сумма ценностей всех критериев.

4.1.5 Целевая функция оценки состояния

Итоговая целевая функция интегральной оценки потребительских свойств для жилых и коммерческих объектов городской среды принимает следующий вид:

$$FNCW(X_f) = \frac{\sum_{i=1}^N w_i f_i(x_i)}{\sum_{i=1}^N w_i f_{imax}(x_i)} * \prod_{j=1}^M c_j * 100\%, \quad (4.14)$$

где $f_i(x_i)$ – нормированная безразмерная оценка объекта по частному критерию

(атрибуту) x_i , вычисленная по формуле (4.6);

$f_{imax}(x_i)$ – максимально возможная безразмерная оценка объекта по частному критерию (атрибуту) x_i при заданных предпочтениях P ;

w_i – вес частного критерия (атрибута) x_i ;

N – мощность множества критериев (атрибутов) оценки;

M – мощность множества корректировок C ;

c_j – корректирующий коэффициент для j -й корректировки.

Формула (4.14) является модифицированной формулой взвешенной суммы [125], дополненной нормированием на максимально возможную оценку и использованием механизма корректирующих коэффициентов [56].

4.1.6 Алгоритм оценки показателя потребительских свойств жилых и коммерческих объектов

До сих пор задача ранжирования (4.2) была рассмотрена в первом приближении: с допущением, что множества X и P универсальны для любого объекта городской среды из множества A . Допущение не влияет на характер целевой функции, принципы оценки и свёртки критериев, но в контексте предметной области не соответствует действительности, т.к. для объектов городской среды разных типов актуальны разные множества значимых атрибутов и предпочтений. Таким образом, постановка задачи ранжирования (4.2) была скорректирована следующим образом:

$$D = \langle F, A, X, G, P, T, C, \mu(t, x, p), \varphi(t, c, p) \rangle, \quad (4.15)$$

где F – формулировка задачи, состоящая в ранжировании вариантов выбора (объектов городской среды) по оценкам их потребительских свойств;

A – множество жилых и коммерческих объектов городской среды (вариантов выбора);

X – совокупность атрибутов, описывающих характеристики вариантов выбора;

G – совокупность условий, ограничивающих область допустимых решений (в контексте задачи – пустое множество);

P – предпочтения, определяющие принцип оценки и сравнения качества вариантов выбора;

T – множество типов объектов городской среды, участвующих в процедуре оценки;

C – множество корректировок, используемых в процессе оценки;

$\mu(t, x, p)$ – функция применимости атрибута $x \in X$ и связанных предпочтений $p \in P$ к объектам типа $t \in T$;

$\varphi(t, c, p)$ – функция применимости корректировки $c \in C$ и связанных предпочтений $p \in P$ к объектам типа $t \in T$.

К функциям $\mu(t, x, p)$ и $\varphi(t, c, p)$ предъявляются следующие требования:

- $\mu(t, x, p) \in [0; 1], \mu(t, x, p) \in Z$;
- $\varphi(t, c, p) \in [0; 1], \varphi(t, c, p) \in Z$.

Таким образом, функции $\mu(t, x, p)$ и $\varphi(t, c, p)$ являются четкими дискретными целочисленными функциями с двумя возможными значениями: 0 и 1. $\mu(t, x, p) = 0$ означает, что атрибут $x \in X$ и связанные предпочтения $p \in P$ к объектам типа $t \in T$ при оценке неприменимы. $\mu(t, x, p) = 1$ означает, что атрибут и связанные предпочтения к объектам типа $t \in T$ при оценке потребительских свойств применяются в обязательном порядке. Аналогично для функции $\varphi(t, c, p)$. Частичная применимость (т.е. ситуация $\mu(t, x, p) \in R$ и $\varphi(t, c, p) \in R$, превращающая их в функции принадлежности [125]) не рассматривается, т.к. примеров подобных ситуаций на практике обнаружено не было [56]. Учитывая бинарный характер функций $\mu(t, x, p)$ и $\varphi(t, c, p)$, в рамках разрабатываемой системы оценки состояния городской среды они были представлены через эксплицитные связи атрибутов и корректировок с типами

объектов городской среды, к которым данные атрибуты и корректировки применимы.

Интегральная оценка FI состояния городской среды по критерию потребительских свойств входящих в неё жилых объектов вычисляется как:

$$FI = \frac{\sum_{i=1}^N (r_i * FNCW(a_i))}{\sum_{i=1}^N r_i}, \quad (4.16)$$

где r_i – коэффициент достоверности данных об объекте $a_i \in A$, равный 0,5 в случае, если данные отмечены как потенциально недостоверные, и 1 в других случаях;

$FNCW(A_i)$ – оценка потребительских свойств объекта $a_i \in A$;

N – мощность множества A .

Аналогично по формуле (4.16) вычисляется оценка состояния городской среды по критерию потребительских свойств входящих в неё коммерческих объектов.

Таким образом, разработанный алгоритм оценки потребительских свойств жилых и коммерческих объектов городской среды принимает вид, представленный на Рисунке 4.1.

4.2 Разработка алгоритма оценки показателя обеспеченности социальной инфраструктурой

Показатели состояния городской среды, используемые в федеральном проекте «Развитие инфраструктуры в населенных пунктах» и Национальном рейтинге качества жизни, не отражают обеспеченность социальной инфраструктурой (учреждениями дошкольного и среднего образования, а также здравоохранения) с точки зрения удовлетворенности потребностей населения в пределах территории объекта мониторинга, в услугах образования (школ, детских садов) и здравоохранения (медицинских учреждений) [133-135].

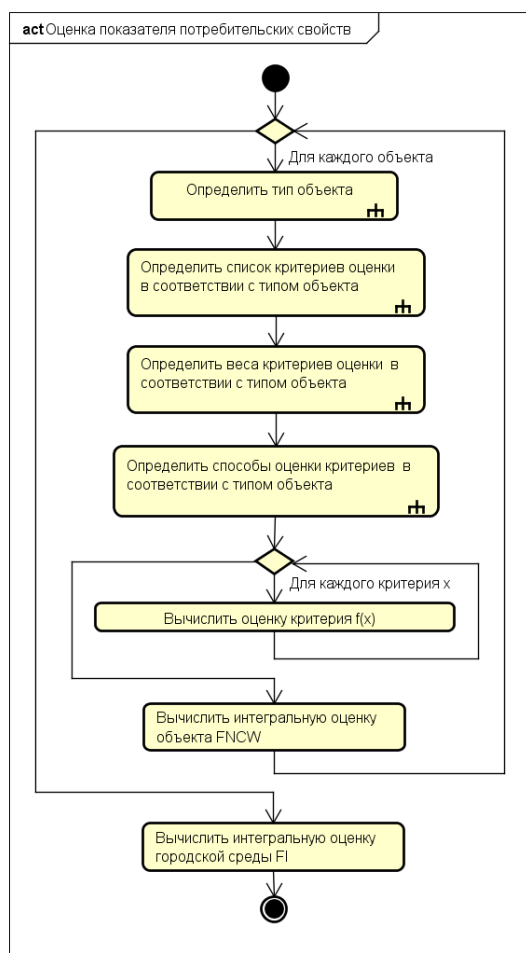


Рисунок 4.1 – Алгоритм оценки потребительских свойств жилого или коммерческого объекта городской среды

При строительстве жилья и объектов социальной инфраструктуры ответственные лица руководствуются сводами принятых на государственном и муниципальном уровне нормативных документов [65, 66], устанавливающих нормативные величины атрибутов объектов, обязательные для соблюдения [64]. При достижении утвержденных нормативных величин объекты социальной инфраструктуры признаются удовлетворяющими потребности населения. Таким образом, значимая в контексте задачи мониторинга состояния городской среды целевая характеристика объектов социальной инфраструктуры, а именно вместимость (проектная мощность [64]), зависит от конфигурации окружающей городской среды, в особенности жилья, от которого зависит допустимая

нормативами численность населения, а от нее – необходимая вместимость объектов социальной инфраструктуры.

Необходимая вместимость объектов социальной инфраструктуры, согласно нормативным документам [65, 66], определяет целевые значения характеристик, которым должны удовлетворять объекты для обеспечения заданного уровня вместимости.

Из описанных выше зависимостей (назовем их прямыми) следует вывод о возможности обратных расчетов: зная площадные характеристики объекта социальной инфраструктуры и обладая соответствующими нормативными документами, можно вычислить его нормативную вместимость. Зная же конфигурацию окружающего объект жилья, можно вычислить вместимость объектов социальной инфраструктуры заданного типа, необходимую для удовлетворения потребностей населения. Подобные расчеты позволяют получить приближенные нормативные значения требуемой и фактической вместимости, но не позволят оценить конфигурацию, существующую в конкретном городе или на конкретном участке его территорий здесь и сейчас. Для точных расчетов необходима фактическая точечная информация о текущей конфигурации населения, получение которой весьма затруднительно по многим причинам. Однако в контексте задачи мониторинга состояния сложной пространственно распределенной системы (городской среды), целью которого является выявление конкурентных преимуществ городов и ограничений, препятствующих их развитию, актуальных проблем, перспективных направлений развития городов [16], подобных нормативных значений достаточно. Вычисление нормативных показателей позволяет выявить нештатные состояния элементов городской среды, допущенные на уровне городского планирования.

Формальная постановка задачи оценки нагрузки ES на объекты социальной инфраструктуры заданного типа в рамках настоящего исследования выглядит как:

$$ES = \langle O, L, N, Al(o, l, n) \rangle, \quad (4.17)$$

где O – множество объектов социальной инфраструктуры заданного типа

(например, школ) в составе городской среды по данным ГИС;

L – множество жилых объектов по данным ГИС;

N – актуальные нормативы для рассматриваемого типа объектов социальной инфраструктуры;

$Al(o, l, n)$ – алгоритм оценки нагрузки на объект вида:

$$Al(o, l, n) = (o, l, n) \xrightarrow{F} v, \quad (4.18)$$

где v – количественная оценка нагрузки на объект o на основе данных об окружающих жилых объектах l и нормативов n .

Задача оценки обеспеченности SI городской среды объектами социальной инфраструктуры заданного типа, таким образом, была представлена в следующем виде:

$$SI = \langle O, Al, V(Al, O), F(V) \rangle, \quad (4.19)$$

где O – множество объектов социальной инфраструктуры заданного типа (например, школ) в составе городской среды по данным ГИС;

Al – алгоритм оценки нагрузки на один объект из множества O ;

$V(Al, O)$ – множество оценок нагрузки, полученных алгоритмом Al для каждого элемента множества O ;

$F(V)$ – функция свёртки оценок из множества V в интегральную оценку обеспеченности.

4.2.1 Оценка максимальной вместимости объектов

Нормативные документы, в частности, Местные нормативы градостроительного проектирования (МНГП) [66] каждого города РФ, устанавливают расчетные показатели минимально допустимого уровня обеспеченности населения объектами социальной инфраструктуры. Данные показатели представляют собой [66]:

- нормативные значения, определяющие зону обслуживания для каждого из рассматриваемых типов объектов социальной инфраструктуры;
- минимальные значения вместимости учреждений образования или здравоохранения в расчете на 1000 человек населения в зоне обслуживания;
- нормативные величины площади земельного участка на единицу вместимости для каждого из типов объектов (опционально – различные нормативные величины для различных диапазонов значений вместимости).

Способ выражения вместимости меняется в зависимости от типа объекта: для школ и детских садов он выражен в минимальном количестве мест на тысячу человек, для учреждений здравоохранения амбулаторного характера (поликлиник) – в минимальной пропускной способности (посещений за смену) на тысячу человек населения в зоне обслуживания и т.д. [66]. Нормативные величины площади земельного участка также зависят от типа объекта [66].

Таким образом, на основе имеющихся нормативных документов, максимальная вместимость C_{max} каждого объекта социальной инфраструктуры рассчитывается как:

$$C_{max} = \frac{S_a}{N_s}, \quad (4.20)$$

где S_a – площадь земельного участка объекта (по данным ГИС);

N_s – нормативная величина площади участка на единицу вместимости.

Значения C_{max} являются нормативными пороговыми значениями, относительно которых в дальнейшем можно судить о перегрузке объектов.

4.2.2 Оценка необходимой вместимости объектов

На основе нормативных величин МНГП [66] и численности населения, проживающего в пределах оцениваемой городской среды и в зоне обслуживания, необходимая нормативная вместимость C каждого объекта социальной

инфраструктуры при заданной конфигурации городской среды рассчитывается следующим образом:

$$C = \text{ceil}(\frac{P_R}{1000} * N_c), \quad (4.21)$$

где P_R – численность населения в зоне обслуживания R объекта;

N_c – норматив вместимости на тысячу человек населения;

$\text{ceil}(x)$ – оператор округления аргумента (x) в большую сторону для получения целочисленных значений C .

При отсутствии детальных данных о численности населения P_R с использованием нормативных документов и собранных с помощью описанных в Главах 2 и 3 методов данных можно вычислить нормативную численность населения, на которую рассчитано жилье в зоне обслуживания согласно утвержденным нормативным документам. Вычисленная на основе такого показателя необходимая вместимость объектов социальной инфраструктуры позволяет определить, сбалансирована ли в принципе конфигурация жилых и инфраструктурных объектов в составе городской среды и способны ли существующие объекты социальной инфраструктуры (образования и здравоохранения) удовлетворить потребности максимальной численности населения, которая может проживать на территории объекта мониторинга согласно действующим нормативным документам [65].

Своды правил (СП) регламентируют нормативные показатели площади жилья на человека в соответствии уровнем комфорта (бизнес, комфорт или муниципальный) конкретного жилого объекта [65].

На основе данных об общей площади жилья в зоне обслуживания объекта социальной инфраструктуры, а также нормативных величин площади на человека для каждого уровня комфорта, нормативная численность населения в зоне обслуживания PRN рассчитывается как:

$$PRN = \sum_{i=1}^K (\frac{S_{li}}{N_i}), \quad (4.22)$$

где S_{li} – общая площадь всех жилых объектов i -го уровня комфорта в зоне

обслуживания объекта инфраструктуры;

N_i – норматив площади на человека для i -го уровня комфорта жилых объектов;

K – мощность множества уровней комфорта жилых объектов (согласно действующему на момент проведения исследования СП [65], $K = 3$).

Более приблизительный расчет по сравнению с (4.22) можно выполнить, располагая данными об общей площади жилых объектов в зоне обслуживания и оперируя рекомендуемыми СП долями жилья каждого уровня комфорта в общем объёме жилищного строительства [65] исходя из допущения, что при строительстве жилых объектов, входящих в оцениваемую городскую среду, данные нормативы были соблюдены. В таком случае формула (4.22) принимает следующий вид:

$$PRN = \sum_{i=1}^K (S_l * \frac{p_i}{N_i}), \quad (4.23)$$

где p_i – рекомендуемая доля i -го уровня комфорта жилых объектов;

N_i – норматив площади на человека для i -го уровня комфорта жилого объекта;

S_l – общая площадь всех жилых объектов в зоне обслуживания объекта инфраструктуры.

Процесс оценки усложняется, когда зоны обслуживания разных объектов социальной инфраструктуры пересекаются, и несколько жилых объектов находятся в зонах обслуживания нескольких объектов социальной инфраструктуры. В таком случае производится оценка распределения спроса между объектами обслуживания. Без такой оценки сложно определить степень загрузки каждого отдельного объекта социальной инфраструктуры (особенно, планируемого к размещению), что затрудняет принятие решения об оптимальном размещении объектов [64]. Вычислительное решение данной подзадачи на основе имитации последовательного заполнения объектов социальной инфраструктуры, пригодное для использования в рамках настоящего исследования, было разработано в работе [64].

4.2.3 Определение перегруженных объектов

Доля загрузки U объекта социальной инфраструктуры рассчитывается на основе значений C_{max} и C при заданной конфигурации городской среды как:

$$U = \frac{C}{C_{max}}. \quad (4.24)$$

Если $U > 1$, объект социальной инфраструктуры является перегруженным, т.е. находится в нештатном состоянии [14, 15] и не способен обеспечить удовлетворение потребностей населения в своей функции при заданной конфигурации городской среды.

4.2.4 Оценка обеспеченности социальной инфраструктурой

Зная долю загрузки каждого объекта социальной инфраструктуры, входящего городскую среду, рассчитанный по формуле (4.24), обеспеченность объектами социальной инфраструктуры заданного типа рассчитывается как:

$$S = \frac{Q_r + \text{ceil}[(1/u) * Q_o]}{Q_t} * 100\%, \quad (4.25)$$

где Q_t – общая потребность в социальной инфраструктуры заданного типа в единицах вместимости;

Q_r – потребность, удовлетворенная объектами без перегрузки;

Q_o – потребность, удовлетворенная перегруженными объектами социальной инфраструктуры ($U > 1$);

u – коэффициент перегрузки, рассчитываемый следующим образом:

$$u = \frac{\sum_{i=1}^{|I_o|} (U_{oi})}{|I_o|}, \quad (4.26)$$

где $|I_o|$ – мощность множества I_o перегруженных объектов социальной инфраструктуры заданного типа, для которых $U > 1$;

U_{oi} – значение U для i -го перегруженного объекта из множества I_o .

4.2.5 Алгоритм оценки показателя обеспеченности социальной инфраструктурой

Алгоритм оценки показателя обеспеченности социальной инфраструктурой заданного типа представлен на Рисунке 4.2.

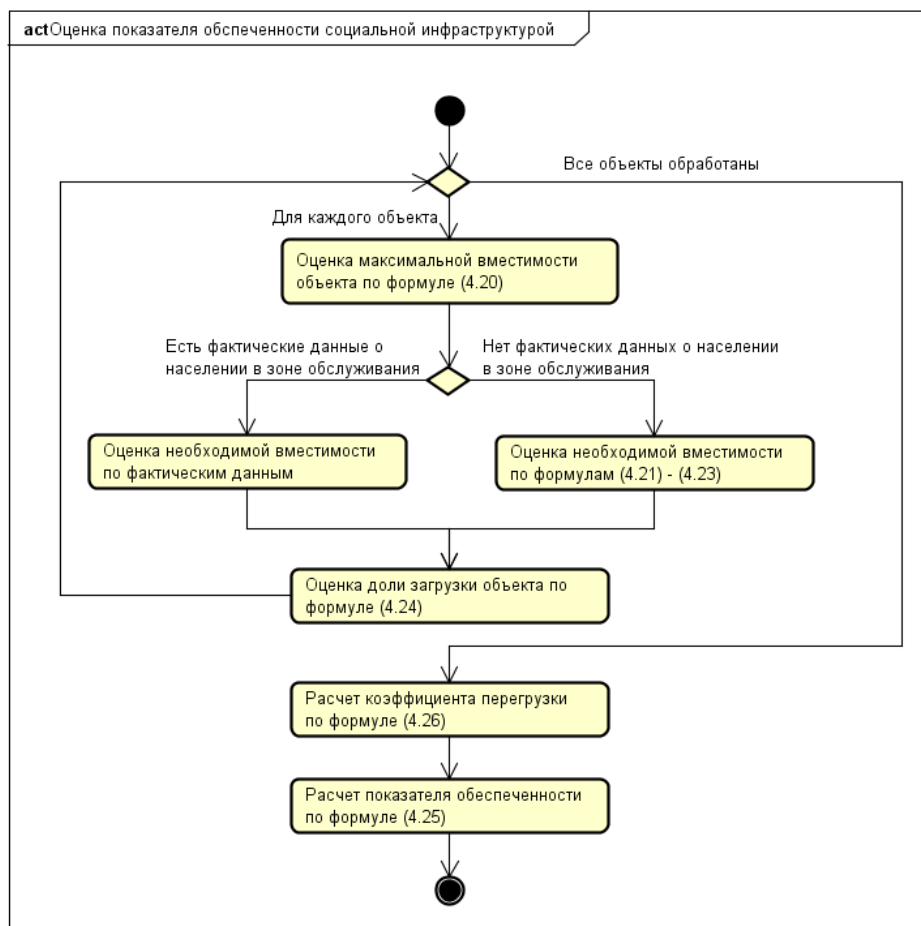


Рисунок 4.2 - Алгоритм оценки показателя обеспеченности социальной инфраструктурой заданного типа

4.3 Выводы по четвертой главе

1. Обоснован выбор методов оценки состояния системы (городской среды) для использования в рамках исследования.

2. Разработан алгоритм оценки показателя потребительских свойств жилых и коммерческих объектов состояния сложной пространственно распределенной системы (городской среды) на основе сформированных данных о состоянии объектов городской среды и их окружении. В основе алгоритма лежит многокритериальная взвешенная оценка, модифицированная для учета существующих статистических зависимостей: корректировок и корректирующих коэффициентов (см. Рисунок 4.1).

3. Разработан алгоритм оценки показателя обеспеченности городской среды объектами социальной инфраструктуры (учреждениями образования и здравоохранения) на основе оценки их вместимости согласно действующим нормативным документам и загруженности с учетом численности населения в зоне обслуживания (см. Рисунок 4.2).

4. Для оценки обеспеченности инфраструктурой также предложены расчетные формулы, позволяющие получить приближенные нормативные величины вместимости и загруженности для определения перегруженных объектов в случае отсутствия детальных данных о численности населения на исследуемой территории.

Разработанные алгоритмы инварианты относительно размеров исследуемых городских территорий и количества объектов на них.

ГЛАВА 5. Разработка системы оценки и обоснование эффективности решений

Для реализации предложенных методов и алгоритмов был разработан прототип программного решения для оценки состояния сложной пространственно распределенной системы (городской среды) по критериям потребительских свойств жилых и коммерческих объектов городской среды, а также обеспеченности объектами социальной инфраструктуры. В процесс оценки, реализуемый разработанным прототипом, были включены объекты городской среды следующих типов:

- жилые объекты:
 - квартиры в многоквартирных жилых домах;
 - частные жилые дома;
- коммерческие объекты:
 - офисы;
 - торговые площади;
 - отдельно стоящие здания;
- объекты социальной инфраструктуры:
 - школы;
 - детские сады;
 - поликлиники.

5.1 Разработка архитектуры прототипа

Была разработана архитектура прототипа системы оценки состояния сложной пространственно распределенной системы (городской среды) по предложенным критериям, представленная диаграммой компонентов на

Рисунке 5.1. Различные модули сбора данных объединены на диаграмме в единые компоненты в соответствии с типом источников данных.

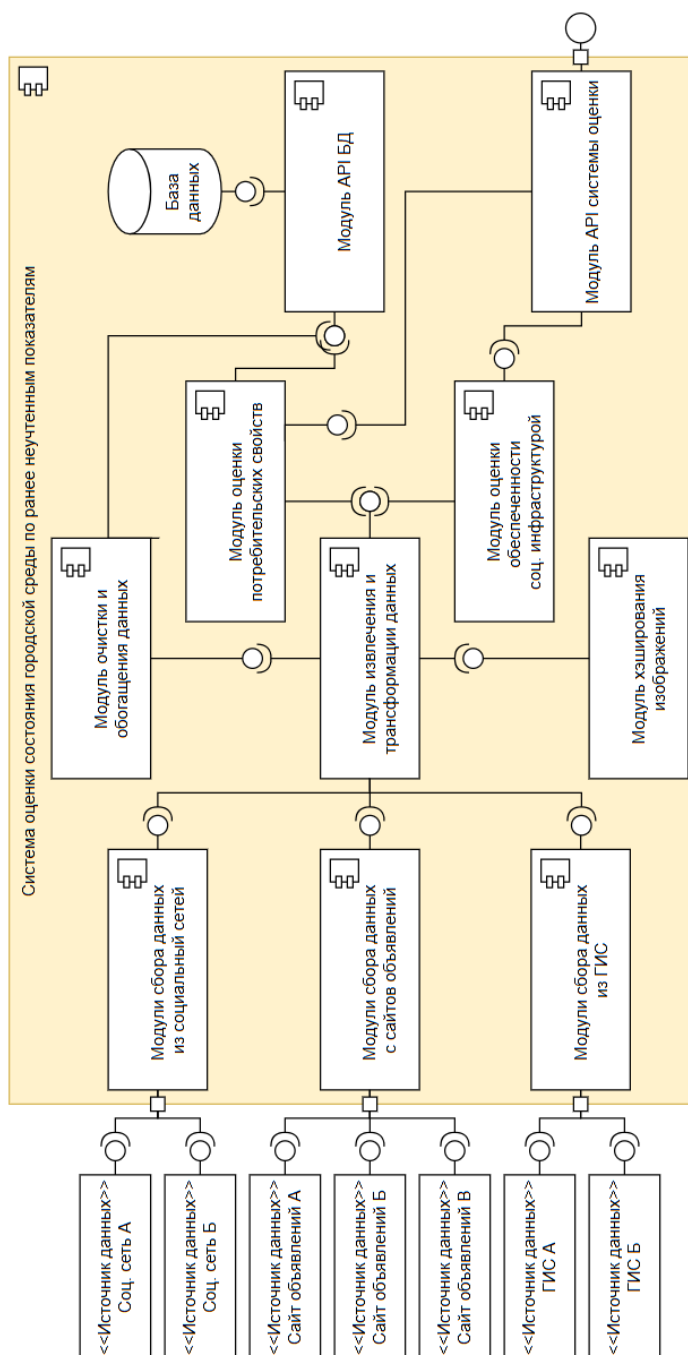


Рисунок 5.1 – Архитектура разработанной системы оценки

Единой точкой доступа к функциям оценки разработанной системы является модуль API, предоставляющий прикладной программный интерфейс (API) для возможной интеграции с внешними техническими решениями.

Разработанная архитектура обеспечивает масштабируемость в части расширения перечня источников данных: интеграция новой социальной сети, тематической площадки с объявлениями или ГИС потребует только разработки и интеграции нового модуля сбора данных из нового источника через его API.

5.2 Проектирование базы данных

Была спроектирована схема базы данных (далее – БД) разработанной, представленная в Приложении А на Рисунке А.1. БД системы хранит информацию только о жилых и коммерческих объектах городской среды в соответствии с типами, учтенными в рамках прототипа, поскольку данные о их окружении и объектах социальной инфраструктуры более изменчивы, а потребность в использовании исторических данных отсутствует.

5.3 Реализация модулей сбора первичных данных

Модули сбора данных реализованы на базе общего интерфейса, определяющего три метода: метод получения сырых данных посредством отправки запроса к API источника (`extract_raw`), метод трансформации сырых данных к общему JSON-формату для дальнейшей обработки (`transform_raw`) и публичный метод `get_data`, который вызывает два предыдущих и возвращает собранные данные в формате JSON.

Таким образом, каждый модуль сбора представляет собой отдельный класс, реализующий вышеописанный интерфейс, определяя собственную реализацию для трёх перечисленных методов с учетом специфики конкретного источника данных и его API (формата запросов, возвращаемых данных и т.п.).

Поскольку реализация публичного метода `get_data` для различных модулей сбора является идентичной (состоит из вызова методов `extract_raw` и `transform_raw`), было решено определить абстрактный класс, реализующий вышеописанный интерфейс, для которого методы `extract_raw` и `transform_raw` являются абстрактными, а метод `get_data` имеет реализацию по умолчанию. Классы модулей сбора данных, таким образом, были реализованы как наследники абстрактного класса, реализующие его абстрактные методы с учетом специфики соответствующего источника данных.

Результирующая диаграмма классов представлена на Рисунке 5.2.

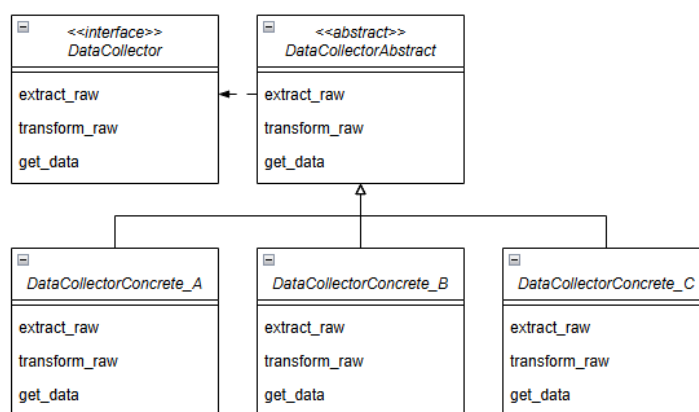


Рисунок 5.2 – Диаграмма классов для модулей сбора первичных данных

Результатами работы каждого модуля являются данные об объектах городской среды в формате JSON. На Рисунках 5.3 и 5.4 представлены примеры JSON-файлов для разного вида объектов: квартиры в многоквартирном доме и объекта социальной инфраструктуры (детском саде). Общая целевая структура JSON-файла представлена в Таблице 5.1.

```

{
  "type": "KINDER",
  "metadata": {
    "source": "2gis.com",
    "issue_date": "2025-03-18",
    "author": {}
  },
  "obj_data": {
    "address_name": "Улица Константина Симонова, 42",
    "name": "Детский сад №17",
    "area": 11772
  }
}
  
```

Рисунок 5.3 – Пример JSON-файла с данными об объекте социальной инфраструктуры

```

{
  "obj_type": "APART",
  "metadata": {
    "source": "VK.com",
    "issue_date": "2025-03-16",
    "author": {
      "name": "Ольга Т.",
      "phone": "+7-9XX-XXX-XX-XX",
      "registered": "2019-12-04",
      "type": 1
    }
  },
  "obj_data": {
    "address": "Волгоградская обл., Волгоград, территория Рабочий посёлок Горьковский, Волгоградская ул., 7",
    "district": "р-н Советский",
    "rooms": 3,
    "area_total": "58,7",
    "area_kitchen": "6",
    "floors_total": 5,
    "floor": 5,
    "balcony": true,
    "repair": "косметический",
    "appliances": "кондиционер, водонагреватель",
    "furniture": "кухня, хранение одежды, спальные места",
    "walls": "кирпичный",
    "description": "Просторная квартира в спокойном уютном районе. С развитой инфраструктурой. Рядом школа,
  }
}

```

Рисунок 5.4 – Пример JSON-файла с данными о жилом объекте

Таблица 5.1 – Общая структура JSON-файла с данными об объекте ГС

Название поля	Тип данных	Описание
obj_type	enum	APART – квартира HOUSE – частный дом COMM – коммерческий объект SCHOOL – школа KINDER – детский сад PCLIN – поликлиника
metadata	Вложенная структура	Содержит сведения о: дате публикации, авторе, контактах, источнике данных.
obj_data	Вложенная структура	Содержит сведения об объекте городской среды в формате «ключ-значение».

5.4 Реализация модуля извлечения и трансформации данных

Модуль извлечения и трансформации данных обеспечивает выполнение двух целевых функций:

- функции извлечения фактов (значений целевых характеристик объектов городской среды) из подаваемых на вход модуля текстовых описаний

объектов на естественном (русском) языке в соответствии с типом этих объектов с помощью Томита-парсера;

- функции нормализации извлеченных фактов и формирования структурированных данных об объектах, чьи описания были поданы модулю в качестве входных данных.

Томита-парсер поставляется в виде бинарного файла, однако чтобы использовать его для решения любой задачи требовалось предварительно создать соответствующую задаче конфигурацию. Основой такой конфигурации является главный настроечный файл (см. Рисунок 5.5) – текстовый файл с расширением .proto, в котором указываются:

- местоположение исходных текстов для обработки (параметр «Input»);
- путь к корневому словарю (параметр «Dictionary»);
- статьи корневого словаря для запуска (параметр «Articles»);
- список фактов для извлечения из текстов (параметр «Facts»);
- способ и место сохранения извлеченных фактов (параметр «Output»);
- несколько вспомогательных параметров.

```
encoding "utf8";
TTextMinerConfig {
    Dictionary = "./tomita/dic.gzt"
    Input = {Dir = "./input"}
    Output = {File = "./output/output.xml"}
    PrettyOutput = "./output/pretty.html"

    Articles = [
        {Name = "Ид"}
        {Name = "Адрес"}
        {Name = "Район"}
        {Name = "Город"}
    ]

    Facts = [
        {Name = "ID"}
        {Name = "Address"}
        {Name = "District"}
        {Name = "City"}
    ]
}
```

Рисунок 5.5 – Фрагмент главного файла настройки Томита-парсера

С точки зрения конфигурации Томита-парсера факты представляют собой именованные наборы типизированных полей для записи данных. Перечень и

состав таких наборов описывается файлом настройки извлекаемых фактов `fact_types.proto` (см. Рисунок 5.6).

```
import "base.proto";
import "facttypes_base.proto";

message ID: NFactType.TFact
{
    optional string ID = 1;
}

message City: NFactType.TFact
{
    optional string City = 1;
}

message Address: NFactType.TFact
{
    optional string Microdistrict = 1;
    optional string Street = 2;
    optional string House = 3;
    optional string Korpus = 4;
    optional string Apartment = 5;
}
```

Рисунок 5.6 – Фрагмент файла настройки фактов для извлечения из текстов

Корневой словарь – это перечень всех используемых в проекте словарей и грамматик. Для каждой грамматики в корневом словаре определяется отдельная статья – множество всех цепочек слов, которые могут быть выведены на основе правил данной грамматики. Корневой словарь описывается файлом настройки `dic.gzt` (см. Рисунок 5.7). Кроме прочего, корневой словарь допускает включение дополнительных словарей директивой «`import`». Сюда же подключаются настроенные файлы фактов и дополнительных ключевых типов (при наличии таковых).

```
encoding "utf8";

import "base.proto";
import "articles_base.proto";
import "fact_types.proto";
import "kwtypes.proto";
import "../gramms/district/district.gzt";
import "../gramms/city/city.gzt";
import "../gramms/type/type.gzt";
import "../gramms/kind/kind.gzt";
import "../gramms/heating/heating.gzt";
import "../gramms/floors/floors.gzt";
import "../gramms/rooms/rooms.gzt";
import "../gramms/_number/_number.gzt";
import "../gramms/officelayout/officelayout.gzt";
import "../gramms/floor/floor.gzt";

TAuxDicArticle "ид"
{
    key = {"tomita:../gramms/id/id.cxx" type=CUSTOM}
}

TAuxDicArticle "Адрес"
{
    key = {"tomita:../gramms/address/address.cxx" type=CUSTOM}
}
```

Рисунок 5.7 – Фрагмент корневого словаря

Для извлечения структурированных данных из текстов на русском языке Томита-парсер использует механизм формальных КС-грамматик – наборов правил, которыми описывается множество искомых цепочек слов в тексте. Для каждой искомой характеристики объектов ГС была написана своя КС-грамматика на формальном языке Томита, описывающая цепочки слов, которыми эта характеристика может быть представлена в текстах.

Пример грамматики представлен на Рисунке 5.8. Данной грамматикой предусмотрен вывод цепочек, включающих ключевое слово «ремонт» или «отделка», сопровождаемое дескриптором типа ремонта («косметический», «евро», «дизайнерский» и т.п.) и маркером наличия ремонта (ключевые слова «выполнен», «не требуется» и т.д.), либо маркером отсутствия ремонта (ключевые слова «отсутствует», «требуется» и т.п.) без дескриптора типа. При этом интерпретации подлежат только цепочки, в которых указан маркер наличия ремонта, а факт «Repair» получает значение, соответствующее входящему в интерпретируемую цепочку дескриптору типа ремонта без нормализации, т.е. в том виде, в каком эти цепочки были найдены в тексте.

```
#encoding "utf8"

RepairW -> 'ремонт'|'отделка';

RepairType -> 'типовой'|'косметический'|'евро'|'европейский'|'капитальный'|'небольшой'|'кап'|
'dизайнерский'|'чистовой'|'касметический'|'офисный';

RepairEn -> 'сделать'|'выполнить'|'есть'|'не'|'требовать'|'присутствовать';

RepairDis -> 'требовать'|'рекомендовать'|'отсутствовать'|'нет'|'нету'|'не'|'сделать'|
'следовать'|'сделать'|'следовать'|'проводить'|'нужный'|'нуждаться'|'не'|'быть';

Repair -> ('под') RepairType RepairW|'евроремонт'|'капремонт';

RepairStr -> RepairEn interp (Repair.RepairEn::not_norm) AnyWord* Repair interp (Repair.Repair::not_norm);

RepairStr -> RepairDis AnyWord* Repair;

RepairStr -> RepairW AnyWord* RepairDis;
```

Рисунок 5.8 – Пример формальной грамматики на языке Томита

В рамках разработанного модуля для Томита-парсера был определен следующий состав конфигурационных файлов:

- набор главных настроечных файлов – по одному на каждый из обрабатываемых типов объектов городской среды;
- набор КС-грамматик – по одной на целевую характеристику;
- словари ключевых слов для отдельных грамматик (применение данных словарей подробно описано в Главе 2 исследования);
- корневой словарь `dic.gzt`, в котором описаны статьи для каждой созданной грамматики;
- файл фактов `fact_types.proto` с описанием полей всех искомым в текстах фактов (характеристик объектов городской среды).

Томиита-парсер был настроен следующим образом:

- источником данных для обработки являются тексты объявлений, подаваемые парсеру через системный поток ввода-вывода;
- результат работы парсера представляется в виде XML-дерева и возвращается по месту вызова парсера через системный поток ввода-вывода;
- необходимые настроечные файлы располагаются в той же директории, что исполняемый файл самого парсера;
- файлы грамматик и дополнительных словарей ключевых слов расположены в отдельной папке «gramms» и размещены в подпапках в соответствии с описываемыми характеристиками.

Структура конфигурации представлена на Рисунке 5.9. На Рисунке 5.10 в визуальной форме представлены результаты работы Томиита-парсера с конфигурацией для обработки объявлений об операциях с квартирами.

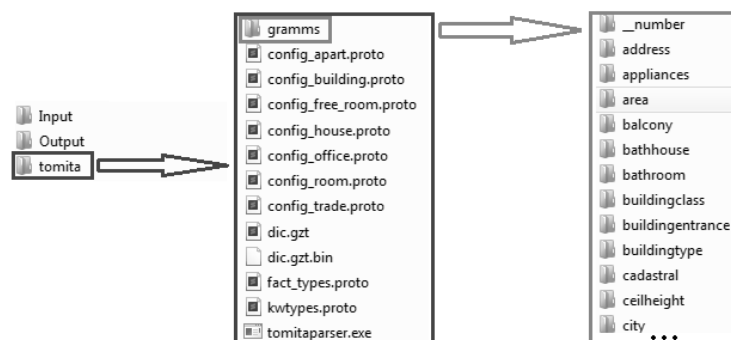


Рисунок 5.9 – Структура конфигурации Томиита-парсера

\ 0.txt EOS

```
{ "id" : 12 , "description" : "сдам трехкомнатную квартиру по адресу проспект ленина 8 с раздельным
санузлом , балконом и кондиционером . кирпичная новостройка бизнес класса . есть собственное
парковочное место . цена 25 тр в месяц" } EOS
```

Rooms	Kind	Type	HouseStatus	Bathroom		ParkingPlace		Conditioner	
Rooms	Kind	Type	Status	Bathroom	Number	ParkingPlace	Number	Conditioner	Number
3	аренда	квартира	новостройка	раздельный	1	парковочное место	1	кондиционер	1

HouseClass	ID	Address				Price		
Class	ID	Microdistrict	Street	House	Korpus	Apartment	Amount	Addon
бизнес	12		проспект ленина	8			25 тр	

Рисунок 5.10 – Пример результатов работы Томита-парсера

Как показал проведенный анализ источников данных, однозначно гарантировать наличие значений для каждой целевой характеристики в каждом объявлении нельзя. Поэтому было решено использовать значения по умолчанию, если в объявлении данные о характеристике не были найдены: “0” для числовых полей и пустая строка для строчных.

Алгоритм работы модуля включает следующие шаги:

- определение конфигурации Томиты в зависимости от типа объекта;
- запуск Томита-парсера с выбранной конфигурацией и передача ему текстов из поля «description» раздела «obj_data» для анализа;
- обработка XML-дерева разбора, полученного от Томита-парсера;
- нормализация данных: приведение количественных характеристик к заранее определенным единым шкалам измерений (площади – в метрах квадратных, цены – в рублях и т.д.);
- построение структурированных описаний объектов городской среды на основе входных данных и извлеченной из текстов информации.

Перечень характеристик различных типов жилых и коммерческих объектов городской среды, данные о которых способен извлекать разработанный модуль, а также их техническое описание, сформированные исходя из схемы БД (см. Приложение А, Рисунок А.1), представлен в Таблице Б.1 Приложения Б.

5.5 Реализация модуля хэширования изображений

Функция модуля – получение на вход массива изображений (структура элемента массива приведена в Таблице 5.2) и расчет перцептивных хэшей от каждого изображения: dHash (хэш на базе DCT) [136] и rHash (хэш на базе преобразования Радона) [137] исходного изображения, а также dHash зеркального изображения. Затем – возвращение полученного на вход массива, дополненного рассчитанными значениями хэшей (структура элементов массива приведена в Таблице 5.3). Пример результатов работы модуля приведен на Рисунке 5.11.

Таблица 5.2 – Структура элементов входного массива изображений

Поле	Тип данных	Описание
_id	int	Идентификатор
url	string	URL-адрес изображения в формате «http://<адрес изображения>»

Таблица 5.3 – Структура элементов выходного массива изображений

Поле	Тип данных	Описание
_id	int	Идентификатор
url	string	URL-адрес изображения в формате «http://<адрес изображения>»
dHash	string	dHash исходного изображения
mirror_dHash	string	dHash зеркального изображения
rHash	string	rHash исходного изображения

```
[
  {
    'id': 1,
    'URL': 'https://f.mirkvartir.me/custom/f7/f7513582-e482-49cd-98ad-1796a23a5bd0.jpg'
  }
]

↓

[
  {
    'id': 1,
    'URL': 'https://f.mirkvartir.me/custom/f7/f7513582-e482-49cd-98ad-1796a23a5bd0.jpg',
    'dHash': 'd861313939293333',
    'mirror_dHash': 'e4797363636b3333',
    'rHash': 'bM\\x00\\xff>Z\\xac^Is53MHF\\x85KGmTX0reAMVMOFVOK@SDQK_U'
  }
]
```

Рисунок 5.11 – Пример результатов работы модуля перцептивного хэширования

5.6 Реализация модуля очистки и обогащения данных

Модуль очистки и обогащения данных в рамках разработанного прототипа системы обрабатывает не только новые поступившие в систему данные, но и данные, сохраненные в БД ранее. Такое решение было принято, чтобы ранее записанные в БД данные также подвергались контролю качества, что позволяет, например, проводить связывание данных при поступлении дубликатов объектов городской среды, ранее сохраненных в БД.

В первую очередь модуль удаляет данные, для которых не выполнено условие полноты: не заполнена информация о городе и адресе расположения.

Затем модуль осуществляет связывание данных. Для объектов разных типов связывание ведется параллельно для оптимизации процесса и повышения быстродействия модуля.

Прогнозирование достоверности данных (поиск данных из «сомнительных» и агентских объявлений) также выполняется в параллельных процессах для повышения быстродействия модуля.

5.7.1 Прогнозирование принадлежности данных агентам

В рамках прототипа для обучения модели классификации были использованы данные из открытых источников, представленные пользовательскими объявлениями об операциях с квартирами (сдаче, продаже) [76]. Размер набора составил 6457 записей (3940 от агентов и 2520 от обычных пользователей). Структура набора приведена в Таблице 5.4.

В составе набора было выделено поле «Тип автора», отражающее принадлежность объявления, из которого были получены данные, физическим лицам («Арендодатель»/«Частное лицо») либо агентам

(«Агентство»/«Застройщик»/«Риелтор»). Данное поле стало целевым параметром после нормирования значений до двоичного представления, где 0 – не-агент («Арендодатель»/«Частное лицо»), 1 – агент («Агентство»/«Застройщик»/«Риелтор»).

Таблица 5.4 – Описание структуры набора данных

Название поля	Возможные значения	Описание
Тип автора	«Агентство», «Арендодатель», «Частное лицо», «Застройщик», «Риелтор»	Указывает, кем является автор объявления, из которого были взяты данные
Тип операции	«Сдам», «Продам»	Операция с объектом
Заголовок	Текст на русском языке в свободном изложении	Содержит краткую информацию об объекте
Описание	Текст на русском языке в свободном изложении	Подробное описание объекта
Цена	Числа	Сумма сделки (в рублях)
Регион	Названия регионов РФ	Регион расположения
Адрес	Текст на русском языке	Адрес с указанием города
Дата объявления	Дата в формате ДД.ММ.ГГГГ	Дата публикации объявления
Широта	Число с плавающей точкой	Географические координаты квартиры
Долгота		
Количество комнат	Целое число	Число комнат
Площадь	Число с плавающей точкой	Площадь в м ²
Этаж	Целое число	Этаж, на котором расположен объект
Этажей	Целое число	Количество этажей в доме
Техника	Текст на русском языке	Список через запятую
Интернет и ТВ	Текст на русском языке	Список через запятую
Балкон	«Балкон», «Лоджия»	Список через запятую
Парковка	Текст на русском языке	Перечислены имеющиеся опции парковки машины
Санузел	«Совмещенный», «Раздельный»	Тип санузла
Материал стен	«Панели», «Кирпич», «Монолит»	Материал стен дома
Лифты	Целое число	Количество лифтов
Год	Год ГГГГ	Год постройки дома

Для дальнейшего приведения набора к целевой структуре, описанной в Таблице 3.2, тексты из поля «Описание» были приведены к нормальной форме и очищены от стоп-слов. В результате были получены цепочки только значимых по смысловой нагрузке слов, представленных в нормальной форме.

Для преобразования в векторы признаков была использована модель Word2Vec [109], обученная на корпусе текстов объявлений из исследуемого набора данных, а также корпусе нормализованных текстов классической русской литературы [77]. Итоговым вектором признаков для обучения модели классификации стал средний вектор по всем значимым словам.

В обучающую выборку попали 3142 записей данных от агентов и 2023 от обычных пользователей. В тестовую – 798 и 494 соответственно. Были обучены и протестированы несколько алгоритмов ансамблевого машинного обучения с целью определить модель с наилучшими метриками качества классификации. Результаты эксперимента представлены в Таблице 5.5.

Таблица 5.5 – Результаты эксперимента с моделями классификации

Алгоритм	Accuracy	Precision	Recall	F1-score	ROC AUC
Random Forest (RF)	0.91	0.92	0.92	0.92	0.91
Stochastic Gradient Descent (SGD)	0.73	0.70	0.97	0.82	0.66
Logistic Regression (LR)	0.82	0.85	0.87	0.86	0.81
Gaussian Naive Bayes (GNB)	0.70	0.73	0.80	0.77	0.66
Gradient Boosting Classifier (GBC)	0.88	0.9	0.91	0.91	0.87
K Nearest Neighbours (KNN)	0.88	0.89	0.91	0.90	0.87

Как видно из результатов (см. Рисунок 5.12), лучшими показателями обладает модель Random Forest (accuracy = 0,95). Модели GBC и KNN также показали достойную точность (0,88), однако в было решено сосредоточиться на модели Random Forest [77].

После оптимизации гиперпараметров модели Random Forest, точность модели удалось поднять до 0,94. Также важными метриками являются Precision (показывает долю истинно-положительных выводов классификатора) и полнота Recall (показывает, какую долю положительных меток классификатор распознал корректно). Для модели Random Forest после оптимизации гиперпараметров эти метрики равны 0,96 и 0,95 соответственно [76].

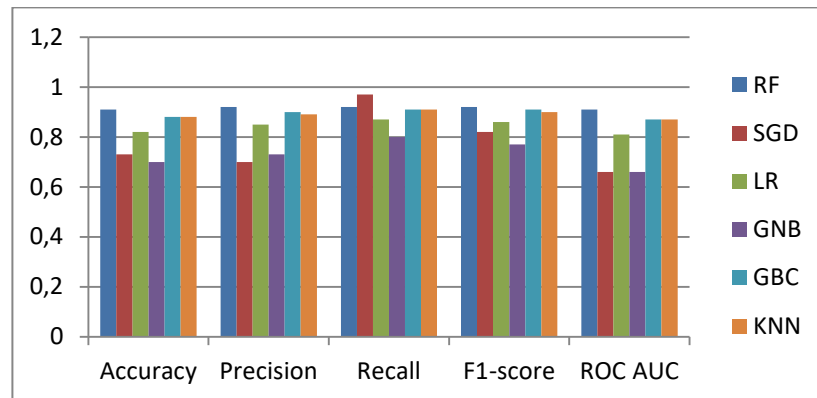


Рисунок 5.12 – Сравнение обученных моделей классификации объявлений по метрикам

Обученная таким образом модель классификации легла в основу прогнозирования принадлежности данных агентам. Результатом работы классификатора является обогащение структурированных данных об объекте городской среды (в рамках разработанного прототипа речь идет именно о квартирах в многоквартирных домах), полученных разработанным модулем на вход, отметкой о принадлежности автора к агентам, а самих данных – к множеству потенциально недостоверных. При записи данных в БД описанные метки попадут в поля «agent» таблицы «Author» и «unreliable» таблицы «Object» соответственно (см. Таблицу А.1 Приложения А).

При поступлении в БД новых данных от автора, отмеченного в таблице «Author» как агент, новые поступающие данные также маркируются как потенциально недостоверные в таблице «Object» даже в том случае, если модель не распознала их как данные от агента.

5.7.2 Выявление «сомнительных» данных

В целях оптимизации и повышения быстродействия модуля все три способа выявления «сомнительных» данных, описанных в Главе 3, выполняются параллельно. Если данные уже отмечены как потенциально недостоверные, описанной далее проверке они не подвергаются.

Разработанный прототип отмечает в качестве потенциально недостоверных данные об объектах, полученные из множественных объявлений (5 и более объявлений от одного автора), которые ранее прошли этап связывания и не были подвергнуты свёртке.

Также недостоверными признаются данные о разных объектах, для которых обнаружены совпадающие изображения (рассмотрению подлежат как новые данные, поступающие в систему, так и ранее записанные в БД). Разными считаются объекты, данные о которых ранее прошли этап связывания и не были подвергнуты свёртке.

Для выявления данных с заниженными ценами в рамках прототипа был реализован кластерный анализ для подмножества жилых объектов городской среды: квартир в многоквартирных домах. Структура одной записи в наборе данных, подвергаемых кластеризации, представляла собой совмещение таблиц Object и Apartment БД (см. Таблицу A.1 Приложения A). В кластерном анализе участвуют как новые поступающие в систему данные, так и ранее сохраненные в БД, дата публикации которых отстает от даты анализа не более, чем на полгода.

Для стандартизации данных все категориальные атрибуты объектов были подвергнуты двоичному кодированию. Атрибуты с типом boolean также были приведены к двоичному представлению. Количественные характеристики были подвергнуты стандартизации.

Кластеризация выполняется по методу Уорда [138]. В результате проводимой подготовки и стандартизации данных удалось достичь значения коэффициента кофенетической корреляции [139] на уровне 0.75, что показывает,

насколько дендрограмма кластеризации отражает отличия между объектами. Чем ближе значение к единице, тем лучше [139]. На Рисунке 5.13 представлена визуализация результатов кластеризации квартир (размерность уменьшена tSNE [140]). Увеличение количества данных должно улучшить результаты.

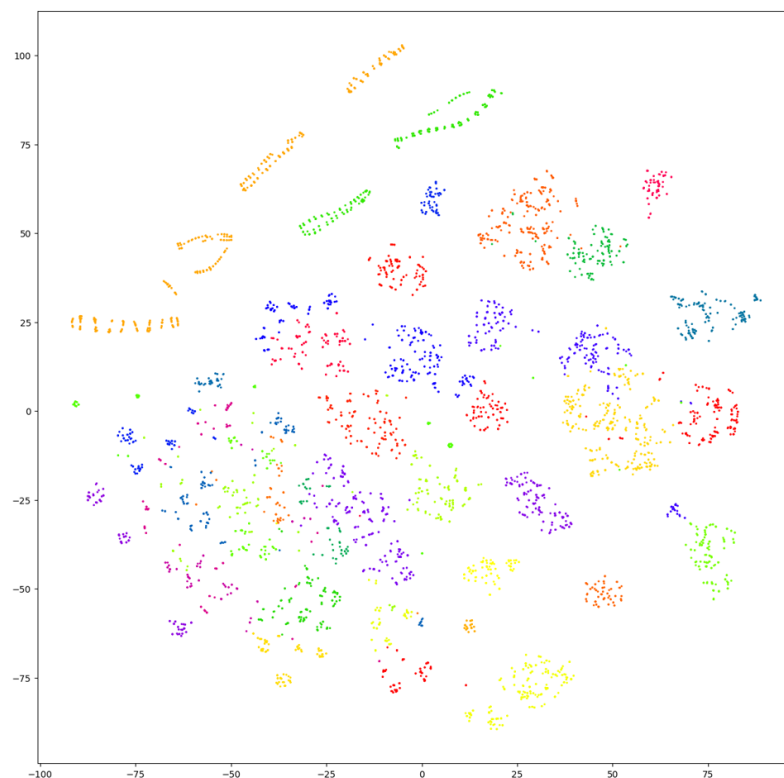


Рисунок 5.13 – Визуализация результатов кластеризации

Для объектов в каждом кластере методом межквартильного размаха [113] производится поиск выбросов по значению цены объекта. Если цена объекта ниже, чем первый квартиль минус 1,5 межквартильных размаха, данные о нём отмечаются как потенциально недостоверные в поле «unreliable» таблицы «Object» (см. Таблицу А.1 Приложения А).

Оптимальное число кластеров выбирается исходя из изменения ускорения – быстроты изменения расстояния между объединяемыми кластерами. Для этого строится совместный график зависимости расстояния между кластерами и ускорения от количества кластеров. Оптимальным считается количество кластеров, при котором ускорение начинает убывать монотонно. Для

реализуемого прототипа оптимальное число кластеров было определено равным 50 (см. Рисунок 5.14).

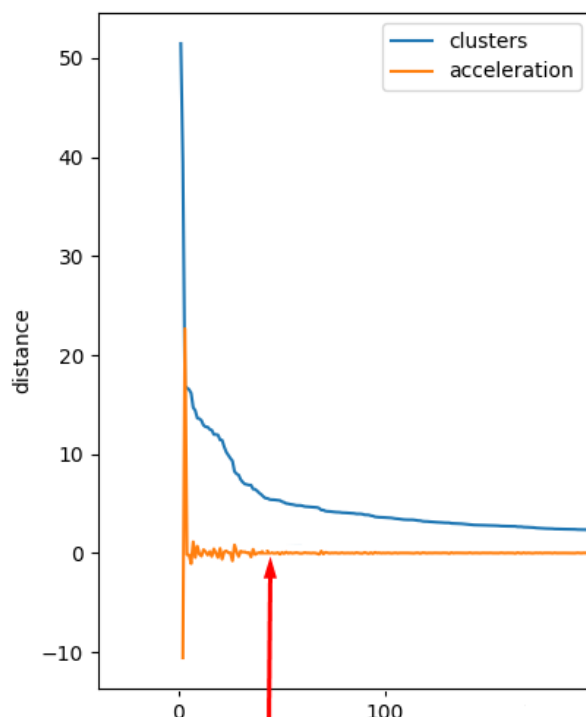


Рисунок 5.14 – Определение оптимального числа кластеров

5.7.3 Обнаружение дубликатов изображений

На этапах обогащения данных, описанных в предыдущих разделах Главы, используется сравнение изображений с целью обнаружения дубликатов. Для сравнения изображений на основе dHash было использовано расстояние Хэмминга. С учетом возможных искажений исходных изображений [79], было решено считать совпадающими изображения, расстояние Хэмминга для которых не превышает заданного порога. Для определения оптимального порога был проведен эксперимент [79] на выборке из 2150 изображений жилых квартир. Часть изображений не были уникальными в пределах выборки. Каждое из изображений подвергалось одному из следующих основных типов искажения:

- обрезке на 5%;
- цветокоррекции.

Прочие искажения, такие как зеркальное отражение и поворот на угол, в ходе эксперимента не рассматривались, т.к. для них предусмотрены альтернативные механизмы противодействия (см. Главу 3).

Каждое из искаженных изображений сравнивалось с исходным. Помимо истинно-положительных (ТР) выводов о совпадении изображений фиксировалось также число ложно-положительных (ФР). Результаты эксперимента представлены в Таблице 5.6 [79].

Таблица 5.6 – Результаты эксперимента по поиску оптимального порогового значения расстояния Хэмминга

	= 0	<=5	<=10	<=15
ТР: Обрезка	1	1	1080	1861
ТР: Цветокоррекция	1867	2150	2150	2150
ФР	0	0	6	186

В качестве оптимального для дальнейшего использования при поиске дубликатов изображений было выбрано пороговое значение расстояния Хэмминга равное 10 [79]. Величины порога расстояния Хэмминга более 15 не рассматривались в связи с ростом количества ложно-положительных результатов сравнения [79].

Сравнение же gHash производится с помощью коэффициента корреляции Пирсона [141]. Чем ближе значение к 1, тем выше степень подобия изображений. В рамках прототипа совпадающими было решено считать изображения, для которых коэффициент равен 1.

Алгоритм, реализуемый модулем, представлен на Рисунке 3.9 в Главе 3.

5.7 Разработка модуля оценки показателя потребительских свойств

Т.к. алгоритмы оценки для жилых и коммерческих объектов идентичны, реализации их также были объединены в одном модуле. Для гибкости и масштабируемости модуля было решено задавать следующие параметры оценки потребительских свойств объектов городской среды через файлы конфигурации:

- список типов жилых и коммерческих объектов для оценки;
- список частных критериев для оценки объектов каждого типа;
- пороговые (достаточные) значения для частных критериев;
- заполненные матрицы парных сравнений Саати (см. Главу 4);
- список возможных корректировок по каждому типу объектов;
- значения корректирующих коэффициентов и условия их применения для каждой корректировки.

Для ведения корректирующих коэффициентов был предусмотрен перечень их возможных типов, приведенный в Таблице 5.7 [142]. В Таблице 5.8 представлен состав файлов конфигурации для оценки потребительских свойств жилых и коммерческих объектов [143].

Расчет оценок потребительских свойств жилых и коммерческих объектов городской среды реализован по каждому типу объектов, учтенному в рамках прототипа, согласно алгоритмам, описанным в Главе 4 настоящего исследования и в соответствии с конфигурациями, приведенными в Таблице В.1 Приложения В.

Достаточный уровень для шкалы оценки площади жилья в конфигурациях для квартир и жилых частных домов был рассчитан на основе средней численности семьи в РФ равной 2,2 человека [144] и целевого показателя обеспеченности населения жильем в 33 кв. м. на человека согласно Указу Президента от 07.05.2024 г. №309 [4]. Достаточный уровень для шкалы оценки площади участка частного жилого дома в соответствующей конфигурации был рассчитан на основе действующего законодательства [143].

Таблица 5.7 – Список доступных типов поправочных коэффициентов

Тип корректировки	Структура параметров
Для числовых характеристик (точное равенство)	fieldname – поле частного критерия val_1...n – значение поля и применяемый коэффициент other – коэффициент в случае, если значение поля не предусмотрено в val_1...n
Для числовых характеристик (попадание в интервалы значений)	fieldname – поле частного критерия range_1...n – диапазон значений поля и применяемый коэффициент other – коэффициент в случае, если значение поля не предусмотрено в range_1...n
Для учета факторов, проявляющихся при равенстве двух полей	fieldname_1 – первое поле fieldname_2 – второе поле True – коэффициент в случае равенства False – коэффициент в случае неравенства
Для учета факторов, задаваемых через текстовые поля	fieldname – поле частного критерия val_1...n – значения и соответствующий коэффициент other – коэффициент в случае, если значение поля не предусмотрено в val_1...n
Для учета факторов, задаваемых через логические поля	fieldname – поле частного критерия True – коэффициент для значений «истина» False – коэффициент для значения «ложь»

Таблица 5.8 – Состав файлов конфигурации для оценки потребительских свойств жилых и коммерческих объектов

Имя файла	Тип файла	Содержимое
criteria.ini	.ini	Список доступных для оценки типов объектов и соответствующие частные критерии оценки для каждого типа
corrections.ini	.ini	Список учитываемых корректировок для каждого типа объектов
scales.ini	.ini	Параметры шкал для частных критериев
correction_coeffs.ini	.ini	Параметры корректировок и значения поправочных коэффициентов
presets.ini	.ini	Содержит полные имена файлов с заполненными матрицами парных сравнений критериев для каждого типа объектов ГС
В соответствии с файлом presets.ini	.csv	Матрица парных сравнений критериев оценки для одного типа объектов

Результатом работы является массив данных об объектах, полученный модулем на вход, дополненный рассчитанными значениями оценок, а также значения интегральных оценок состояния сложной пространственно распределенной системы (городской среды) по ранее неучтенным показателям состояния: потребительским свойствам жилых объектов и потребительским свойствам коммерческих объектов городской среды.

5.8 Разработка модуля оценки показателя обеспеченности социальной инфраструктурой

Аналогичное решение по определению значимых параметров оценки посредством файлов конфигурации было принято для модуля оценки показателя обеспеченности социальной инфраструктурой. Файлами конфигурации было решено определять следующие параметры:

- список типов социальной инфраструктурой;
- нормативные значения вместимости на 1000 человек для каждого типа социальной инфраструктурой;
- нормативные значения площади участка земли на единицу вместимости для каждого типа социальной инфраструктурой;
- зона обслуживания для каждого типа социальной инфраструктурой (радиус обслуживания и/или нормативная транспортная доступность);
- рекомендуемые доли жилья каждого уровня комфорта (см. Главу 4);
- нормативные значения площади жилья на человека по каждому уровню комфорта (см. Главу 4);
- этажность различных типов жилых объектов (многоквартирных и частных домов) для случая, когда данная информация в результате формирования данных получена не была.

Полный состав файлов конфигурации для модуля оценки обеспеченности социальной инфраструктурой представлен в Таблице 5.9.

В рамках прототипа для оценки обеспеченности социальной инфраструктурой через файлы конфигурации были на основе действующих СП [65] и МНГП [66] заданы параметры, приведенные в Таблицах 5.10–5.12.

Таблица 5.9 – Состав файлов конфигурации для оценки обеспеченности СИ

Имя файла	Тип файла	Содержимое
infrastructure.ini	.ini	Список доступных для оценки типов объектов инфраструктуры и нормативные величины для каждого типа
living.ini	.ini	Рекомендуемые доли уровней комфорта жилья, нормативные величины площади жилья на человека для каждого уровня
default.ini	.ini	Этажность по умолчанию для каждого типа жилых объектов

Таблица 5.10 – Нормативные значения для объектов инфраструктуры

Тип объекта инфраструктуры	Зона обслуживания	Норматив площади территории на единицу вместимости	Норматив вместимости на 1000 человек
Детский сад	300 м	30 кв. м.	52 места
Школа	500 м или 15 минут на общ. транспорте	20 кв. м.	112 мест
Поликлиника	1000 м	10 кв. м.	9,68 посещений в смену

Таблица 5.11 – Доли и нормативы площади для уровней комфорта

Уровень комфорта	Норматив площади на человека, кв. м.	Доля в жилом фонде
Бизнес-класс	40	10
Эконом-класс	30	50
Муниципальный	20	40

Таблица 5.12 – Этажность жилых объектов по умолчанию

Тип жилого объекта	Этажность здания по умолчанию
Многоквартирный дом	8
Частный дом	1

Результатом работы является массив данных об объектах городской среды, полученный модулем на вход, дополненный оценками вместимости и загруженности, а также интегральная оценка показателя обеспеченности городской среды объектами социальной инфраструктурой как по каждому типу в отдельности, так и в целом по всем типам.

5.9 Обоснование эффективности предложенных решений

Для проверки работоспособности прототипа системы и эффективности разработанных методов на предмет удовлетворения предъявляемым критериям эффективности, были проведены вычислительные эксперименты.

Проверка достоверности алгоритма оценки потребительских свойств проводилась на примере анализа объявлений с завершёнными сделками по продаже квартир в г. Волгограде, полученных в ходе исследований на кафедре САПРиПК ВолгГТУ в 2018-2019 гг. Были выделены записи о продаже квартир в 2019 г., количество записей составило 197 штук. Фрагмент полученного набора данных представлен в Таблице 5.13. Прочерк в ячейке таблицы означает отсутствие данных.

Поскольку объекты имели разную жилую площадь и количество комнат, что также влияло на их рыночную цену, было решено оперировать величинами цены за квадратный метр, а не абсолютными ценами.

Согласно гипотезе (см. Главу 4), оценки потребительских свойств должны коррелировать с ценами, по которым продается (или сдается в аренду) объект городской среды, т.к. оба показателя (цена и оценка городской среды) прямо или

опосредованно являются мерилom одной характеристики – качества объекта. Сравнение уровней оценок, полученных на основе этих данных разработанными методами, с уровнями нормированных цен, по которым были совершены сделки, представлено на графиках на Рисунке 5.15. Для представления оценок и цен за единицу площади, измеряемых десятками тысяч рублей, на одном графике в диапазоне от 0 до 100, величины цен были нормированы на максимальное значение в выборке и затем умножены на 100, что позволило привести значения цен к безразмерным в диапазоне от 0 до 100. Таким образом, верхний график отображает нормированные цены объектов за единицу площади (метр квадратный) в рублях, отсортированные по возрастанию, а нижний – оценки потребительских свойств тех же объектов в порядке возрастания их цен.

Таблица 5.13 – Фрагмент данных для проверки достоверности оценки показателя потребительских свойств

Частный критерий	Квартира 1	Квартира 2	Квартира 3	...	Квартира 197
Общая площадь	79.6	66	68	...	62
Лифты	1	1	2	...	1
Цена за кв.м.	80 124	69 211	73 665	...	70 743
Ремонт	–	косметический	евроремонт	...	–
Этаж	2 из 9	1 из 9	6 из 12	...	2 из 10
Санузел	раздельный	раздельный	раздельный	...	совмещенный
Кондиционер	есть	–	есть	...	–
Материал стен	панели	кирпич	панели	...	монолит
Интернет	есть	есть	есть	...	нет
Лифт	есть	нет	есть	...	есть
Грузовой лифт	есть	нет	нет	...	нет
Мебель	есть	есть	есть	...	есть
Техника	есть	есть	есть	...	есть
Балкон	есть	есть	есть	...	–
Аптека рядом	есть	есть	есть	...	нет
Продукты рядом	есть	есть	есть	...	есть
Остановки рядом	есть	есть	есть	...	есть
Школы рядом	есть	нет	есть	...	нет
Парки рядом	есть	нет	нет	...	есть
Дет. сады рядом	есть	нет	есть	...	есть

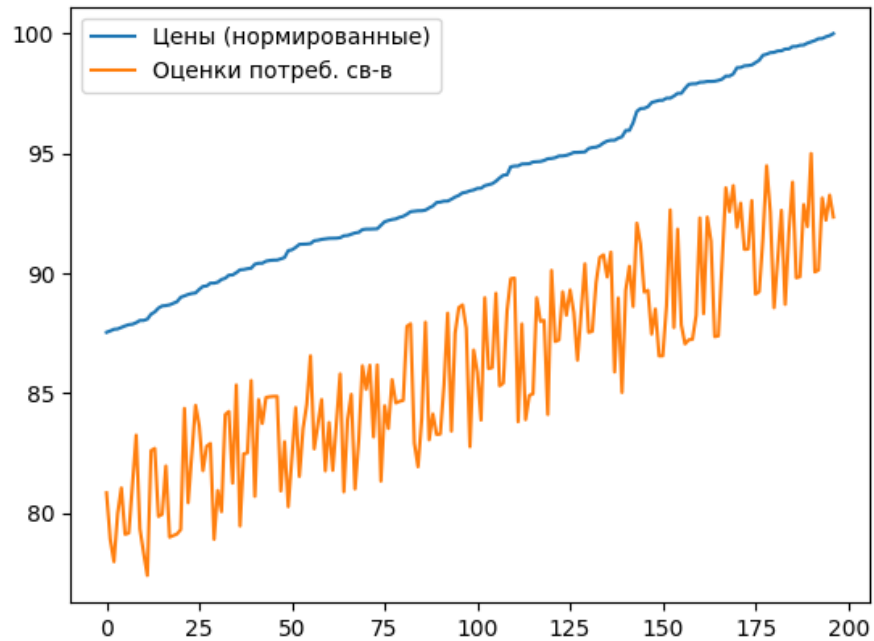


Рисунок 5.15 – Сравнение уровней нормированных цен жилых объектов и оценок их потребительских свойств

Коэффициент корреляции Пирсона двух величин составил 0,89, что свидетельствует о наличии устойчивой связи, хотя определенная доля флуктуаций случайного характера присутствует. Связано их наличие с неполнотой данных об отдельных частных критериях оценки, а также с возможностью назначения более высоких или более низких цен относительно объективных характеристик конкретного объекта в силу различных причин, макроэкономические колебания же были исключены за счет использования данных, собранных в течение периода стабильности цен. Результаты статистически значимы по t-критерию Стьюдента при заданном размере выборки, величине корреляции и уровне значимости 0,05.

Для проверки работоспособности и достоверности результатов работы разработанного прототипа системы оценки был рассмотрен Дзержинский район г. Волгограда. Район является густонаселенным и наиболее активно застраиваемым на момент проведения исследования, поэтому ситуации локального дисбаланса потребностей жителей и обеспеченности этих потребностей возникали достаточно регулярно. Был произведен сбор, контроль и повышение качества данных с последующим расчетом оценок состояния сложной

пространственно распределенной системы (городской среды) по трем предлагаемым критериям. Оценка обеспеченности социальной инфраструктуры производилась на основе расчетных нормативных данных о численности населения. В результате оценки были получены оценки по критериям, представленные диаграммой на Рисунке 5.16.

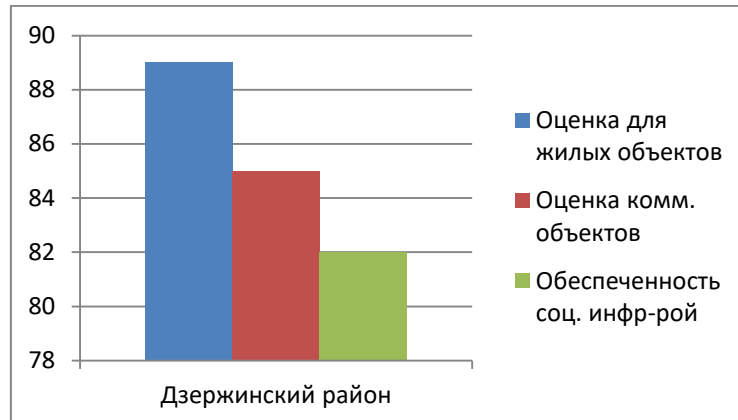


Рисунок 5.16 – Результаты оценки состояния ГС для Дзержинского района г. Волгограда

Как видно по Рисунку 5.16, наименьшая оценка для района была получена по критерию обеспеченности социальной инфраструктурой (на графике отображена интегральная средняя оценка обеспеченности по каждому из рассматриваемых в рамках прототипа типов объектов СИ). Раздельная оценка по каждому типу инфраструктуры представлена на Рисунке 5.17.

Меньшей из полученных оценок являлась оценка обеспеченности района школами. Детальный анализ загрузки конкретных объектов показал среднюю перегрузку школ на уровне 19% (пример перегруженного объекта приведен на Рисунке 5.18) при наличии участков жилой застройки, не обеспеченных школой в пешей доступности (см. Рисунок 5.19).

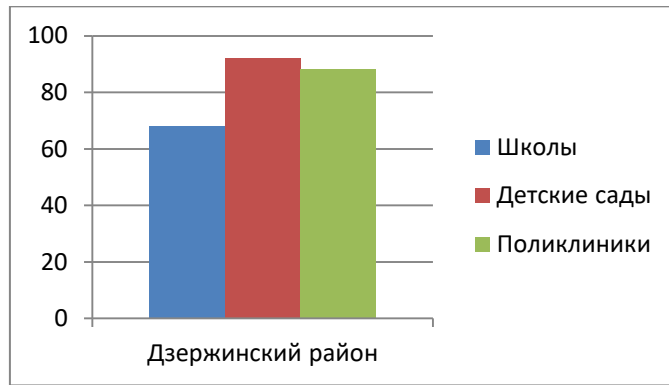


Рисунок 5.17 – Результаты оценки показателя обеспеченности социальной инфраструктурой для Дзержинского района г. Волгограда

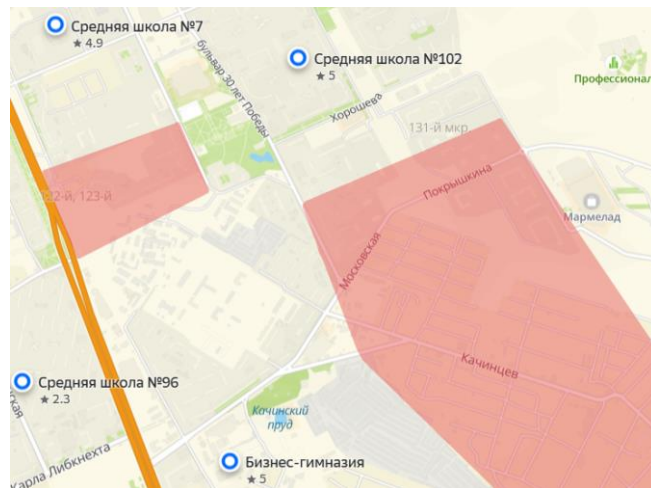


Рисунок 5.18 – Территории, не обеспеченные школами в пешей доступности



Рисунок 5.19 – Пример перегруженной школы на карте района

Действующие решения по оценке состояния городской среды позволяют получать оценки в масштабах всего города, что приводит к бесконечно малым коэффициентам локализации K_{loc} (1.9) при S_l = площадь города. Оценка же с помощью разработанных алгоритмов позволяет обнаруживать объекты в нештатных состояниях с точностью до площади конкретного объекта, т.е. позволяет достичь значений $K_{loc} \approx 1$. Таким образом, разработанные решения позволяют повысить эффективность мониторинга с точки зрения информативности результатов (локализации объектов городской среды в нештатном состоянии).

Среднее время работы прототипа от момента запуска до момента завершения процесса оценки по результатам 10 независимых запусков для одного района (Дзержинского) г. Волгограда составило 1,43 часа. При масштабировании на весь город (8 районов) среднее время работы по результатам 10 независимых запусков составило 7,91 часа, что говорит о зависимости времени работы от размеров территории объекта мониторинга и количества объектов городской среды, в него входящих. Также с целью фиксации времени работы прототип системы был запущен для оценки городской среды г. Москва как самого крупного по площади города РФ. Была использована конфигурация оценки для г. Волгограда (см. Таблицу В.1 Приложения В), из которой была исключена оценка района расположения объекта. Время работы прототипа составило в среднем 16,31 часа.

Степень определенности состояния городской среды как сложной пространственно распределенной системы при оценке по 16 показателям федерального проекта, дополненным тремя предложенными в рамках исследования критериями, выраженная через показатель покрытия мониторингом (1.3), по сравнению с оценкой на основе показателей федерального проекта повышается на ~31% (21 показатель против 16), что подтверждает повышение полноты результатов мониторинга.

Сравнение значений критериев эффективности мониторинга, определенных в Главе 1, подтверждающее эффективность разработанных решений, представлено в Таблице 5.14.

Таблица 5.14 – Результаты апробации разработанных решений

Критерий эффективности	Метрика	Изменение	
		Было	Стало
Информативность результатов в части локализации объектов в нештатном состоянии	Коэффициент локализации K_{loc} (1.8)	$K_{loc} \approx 0$	$K_{loc} \approx 1$
Полнота информации о состоянии объекта мониторинга	Показатель покрытия мониторингом $D(1.7)$	Показатель покрытия повышен на ~31% относительно системы мониторинга на основе показателей федерального проекта	
Интегральный критерий Des (1.12)		0,95 («Отлично» по шкале Харрингтона)	

5.10 Выводы по пятой главе

1. Разработан прототип системы автоматизированной оценки состояния городской среды по ранее неучтенным показателям: потребительским свойствам жилых, коммерческих объектов и обеспеченности социальной инфраструктурой

(образованием и здравоохранением). Описаны детали реализации каждого из разработанных методов и алгоритмов в составе прототипа.

2. По результатам апробации разработанного прототипа получено подтверждение повышения эффективности действующей системы мониторинга состояния городской среды как сложной пространственно распределенной системы с точки зрения информативности и полноты результатов вследствие использования разработанных решений (методов и алгоритмов). В частности, удалось добиться:

- повышения информативности результатов мониторинга с точки зрения локализации объектов в нештатном состоянии (коэффициент локализации $K_{\text{loc}} \approx 1$ против бесконечно малых значений при использовании только действующей системы мониторинга);

- повышения полноты данных о состоянии сложной пространственно распределенной системы (городской среды) (улучшения определенности состояния на $\sim 31\%$).

Заключение

Выполненные исследования вносят вклад в развитие методов мониторинга состояния сложных систем. В рамках работы получены следующие результаты и выводы:

1. Анализ исследований и существующих решений в области мониторинга состояния городской среды как сложной пространственно распределенной системы позволил сделать вывод о несовершенстве применяемых в настоящий момент методов мониторинга состояния городской среды. Концентрация на мониторинге отдельно взятых подсистем, использование потенциально недостоверных или неактуальных данных, отсутствие мониторинга части подсистем в комплексных методах мониторинга негативно влияет на эффективность мониторинга в целом.

2. Выделено наиболее актуальное для РФ решение, лежащее в основе действующей системы мониторинга состояния городской среды, и его ключевые недостатки: полнота и информативность результатов. Поставлена цель по повышению эффективности мониторинга за счет автоматизированного сбора данных из гетерогенных открытых источников и их автоматической обработки для оценки состояния городской среды по ранее неучтенным значимым показателям: потребительским свойствам жилых и коммерческих объектов, а также обеспеченности социальной инфраструктурой (образованием и здравоохранением).

3. Разработан новый метод автоматизированного извлечения и трансформации данных о состоянии городской среды из гетерогенных источников на основе обработки графических данных, естественных языков, а также пространственного анализа для получения наиболее актуальных и полных первичных данных. Разработанный метод обладает универсальностью с точки зрения применимости к различным городам земного шара, а также масштабируемостью за счет возможности расширения перечня целевых

характеристик объектов, увеличения объёмов словарей и комплексности грамматик и т.д.

4. Разработан новый метод очистки и обогащения данных, реализующий удаление неполных данных, связывание данных, а также прогнозирование достоверности данных на основе сочетания методов статистики, обработки графических данных, естественных языков и машинного обучения.

5. Разработаны алгоритмы оценки состояния городской среды как сложной пространственно распределенной системы по ранее неучтенным значимым показателям. Отличительной особенностью алгоритмов является использование универсальных методов, адаптированных к задачам оценки состояния городской среды и принятия решений по её преобразованию.

6. На основе предложенных методов и алгоритмов разработан прототип системы оценки состояния городской среды по ранее неучтенным значимым показателям. Проведена апробация разработанных методов и алгоритмов, доказывающая достижение поставленной цели.

7. Диссертация представляет собой научно-квалификационную работу, главным результатом которой является повышение эффективности мониторинга состояния городской среды путем повышения полноты (на ~31%) и информативности (улучшение локализации объектов городской среды в нештатном состоянии с территории всего города до конкретного объекта) результатов.

8. Получены 3 свидетельства о регистрации программ для ЭВМ, приведенные в Приложении Г.

9. Результаты работы успешно внедрены в ООО «СтройБилд» с целью поддержки принятия решений по выбору места строительства и конфигурации новых жилых и нежилых зданий, что позволило повысить эффективность и обоснованность принимаемых решений и сократить затраты времени на анализ состояния ГС. Акт внедрения приведен в Приложении Д.

В ходе выполнения диссертационной работы были решены все поставленные задачи, цель исследования достигнута. В качестве направления дальнейшей работы можно выделить изучение перспектив использования предложенных методов и алгоритмов в качестве основы системы поддержки принятия решений в области комплексного развития городских территорий.

Список используемых сокращений и условных обозначений

GLI	Global Liveability Index (глобальный индекс качества жизни)
КС-грамматики	контекстно-свободные грамматики
РФ	Российская Федерация
АИС ФРТ	Автоматизированная информационная система «Фонда развития территорий»
ГИС	геоинформационная система
БД	база данных
ЖКХ	жилищно-коммунальное хозяйство
Рхэш	хэш на базе преобразования Радона
ООН	Организация объединенных наций
ДЗЗ	дистанционное зондирование земли
AVM	automated valuation model
HPIs	house price indices
МНГП	местные нормативы градостроительного проектирования
СП	Своды правил
JSON	JavaScript Object Notation (текстовый формат обмена данными, основанный на JavaScript)
API	Application Programming Interface (программный интерфейс приложения)
TP	true-positive (истинно-положительное)
FP	false- positive (ложно-положительное)

Список используемых источников

1. Buda, E., Broitman, D., Czamanski, D., Urban Structure in Troubled Times: The Evolution of Principal and Secondary Core/Periphery Gaps through the Prism of Residential Land Values [Электронный ресурс] // Sustainability. – 2021. – Vol. 13, № 10. – 5722. – Режим доступа : <https://doi.org/10.3390/su13105722> (дата обращения 16.05.2026).
2. "Градостроительный кодекс Российской Федерации" от 29.12.2004 N 190-ФЗ (ред. от 04.05.2026) [Электронный ресурс]. – Режим доступа : https://www.consultant.ru/document/cons_doc_LAW_51040/ (дата обращения 16.05.2026).
3. Федеральный закон "О внесении изменений в Градостроительный кодекс Российской Федерации и отдельные законодательные акты Российской Федерации в целях обеспечения комплексного развития территорий" от 30.12.2020 N 494-ФЗ [Электронный ресурс]. – Режим доступа : https://www.consultant.ru/document/cons_doc_LAW_372677/ (дата обращения 16.05.2026).
4. Указ Президента Российской Федерации от 07.05.2024 г. №309 «О национальных целях развития российской федерации на период до 2030 года и на перспективу до 2036 года» [Электронный ресурс]. – Режим доступа : <http://www.kremlin.ru/acts/bank/50542> (дата обращения 16.05.2026).
5. Перечень поручений по итогам заседания наблюдательного совета АСИ N Пр-1247 от 06 августа 2020 года [Электронный ресурс]. – Режим доступа : <http://www.kremlin.ru/acts/assignments/orders/copy/63873> (дата обращения 16.05.2026).
6. Паспорт национального проекта, Национальный проект «Инфраструктура для жизни» (Министерство строительства и жилищно-коммунального хозяйства Российской Федерации, 2025 г.) [Электронный ресурс]. – Режим доступа :

<https://demo.garant.ru/#/document/411792769/paragraph/24:0> (дата обращения 16.05.2026).

7. Цели в области устойчивого развития [Электронный ресурс] // ООН. – Режим доступа : <https://www.un.org/sustainabledevelopment/ru/> (дата обращения 16.05.2026).

8. Масленникова, Н. Ю. Понятие и сущность мониторинга с позиции системного подхода / Н. Ю. Масленникова, О. К. Слинкова // Science Time. – 2014. – № 6. – С. 110-121.

9. Фаттахов, Р. В., Низамутдинов, М. М., Орешников В. В. Методология оценки привлекательности крупных городов России для жителей, туристов и бизнеса // Регион: Экономика и Социология. – 2019. – № 4(104). – С. 268-294. – doi: 10.15372/REG20190412

10. Национальный рейтинг качества жизни [Электронный ресурс] // Агентство стратегических инициатив. – Режим доступа : https://asi.ru/government_officials/quality-of-life-ranking/ (дата обращения 16.05.2026).

11. Паспорт федерального проекта, Федеральный проект «Развитие инфраструктуры в населенных пунктах» (Министерство строительства и жилищно-коммунального хозяйства Российской Федерации, 2025 г.) [Электронный ресурс]. – Режим доступа : <https://base.garant.ru/411792767/> (дата обращения 16.05.2026).

12. Паспорт федерального проекта, Федеральный проект «Формирование комфортной городской среды» (Министерство строительства и жилищно-коммунального хозяйства Российской Федерации, 13 января 2025 г.) [Электронный ресурс]. – Режим доступа : <https://demo.garant.ru/#/document/411792753/paragraph/12439:0> (дата обращения 16.05.2026).

13. Приказ Минстроя России от 09.06.2026 N 363/пр «Об утверждении методики расчета показателя «Улучшение качества среды для жизни в опорных

населенных пунктах» федерального проекта «Развитие инфраструктуры в населенных пунктах» национального проекта «Инфраструктура для жизни».

14. Кореньков, В. В., Дмитриенко П. В. Архитектура и пути реализации системы локального мониторинга ресурсного центра // Системный анализ в науке и образовании. – 2011. – № 3. – С. 36-47.

15. Дмитриенко, П. В. Методика оценки эффективности систем мониторинга вычислительных ресурсов // Компьютерные исследования и моделирование. – 2012. – Т. 4, № 3. – С. 661-668.

16. Распоряжение Правительства РФ «Об утверждении Методики формирования индекса качества городской среды» от 23.03.2019 N 510-р (ред. от 20.09.2023) [Электронный ресурс]. – Режим доступа : https://www.consultant.ru/document/cons_doc_LAW_320891/ (дата обращения 16.05.2026).

17. Тихонов, В. Е. Категории анализа городских центров как сложных систем // Наука, образование и экспериментальное проектирование в МАРХИ : тезисы докладов международной научно-практической конференции профессорско-преподавательского состава, молодых ученых и студентов. — Москва : МАРХИ, 2013. – С. 47–49.

18. Геложина, Л. М. Комфортная городская среда: понятие и роль общественного участия в развитии городской среды // Экономика и социум. – 2021. – № 9(88). – С. 325-329.

19. Кравченко, Э. В., Женетль, З. Н. Мониторинг состояния городской среды для целей размещения и строительства объектов недвижимости // Электронный сетевой политематический журнал "Научные труды КубГТУ". – 2020. – № 8. – С. 775-781.

20. Адамов, О. И., Афолина, М. И., Коробейникова, А. Е., Привезенцева, С. В. Формирование комфортной городской среды : учебно-методическое пособие. – Москва : МИСИ-МГСУ 2022. – 32 с. – ISBN 978-5-7264-3048-5.

21. Гаврилина, Т. О. Социокультурное пространство города: структура, функции // Вестник Сургутского государственного педагогического университета. – 2019. – № 2(59). – С. 26. – doi: 10.26105/SSPU.2019.59.2.026.
22. Катаева, Ю. В., Кадырова, Я. И. Городская среда как пространство жизнедеятельности местных сообществ // Экономика и управление в XXI веке: тенденции развития. – 2013. – № 10. – С. 76-81.
23. Кочнев, Д. С. Жилая среда в социокультурном пространстве современного российского города // Ярославский педагогический вестник. – 2015. – № 6. – С. 364-367.
24. Глумнушина, П. С. Развитие открытых общественных пространств при строительстве коммерческой и жилой недвижимости // Весенние дни науки : сборник докладов, Екатеринбург, 21–23 апреля 2022 года. – Екатеринбург: Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, 2022. – С. 626-631.
25. Хомич, В. А. Экология городской среды: Учеб. пособие для вузов. – Омск: Изд-во СиБАДИ, 2002. – 267 с.
26. Алибаева, Н. Т. Мониторинг состояния городской среды // Международный журнал гуманитарных и естественных наук. – 2023. – № 9-2(84). – С. 156-159. – doi: 10.24412/2500-1000-2023-9-2-156-159.
27. Kitchin, R. The Data Revolution: A Critical Analysis of Big Data, Open Data & Data Infrastructures. – SAGE Publications, Limited, 2022. – 376 p.
28. Scott, A. J. Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier // Economic Geography. – 2012. – Vol. 88. – P. 97-100. – doi: 10.2307/41474071.
29. World Cities Report 2016 [Электронный ресурс] // UN-Habitat. – Режим доступа : <https://unhabitat.org/sites/default/files/download-manager-files/WCR-2016-WEB.pdf> (дата обращения 16.08.2025).
30. Batty, M. The New Science of Cities. – The MIT Press, 2017. – 520 p.

31. A Deep Learning Framework About Traffic Flow Forecasting for Urban Traffic Emission Monitoring System / B. Yao, A. Ma, R. Feng [и др.] // *Frontiers in Public Health*. – 2022. – Vol. 9. – DOI 10.3389/fpubh.2021.804298.
32. An automated monitoring system of urban pollution using geochemical and magnetic parameters / F. Bautista, D. Bautista-Hernández, A. Pacheco [и др.] // *Boletín de la Sociedad Geológica Mexicana*. – 2023. – Vol. 75, № 2. – A100523. – DOI 10.18268/bsgm2023v75n2a100523.
33. Sreelatha, R., Roopa Lakshmi R. Deep Learning-based Detection System for Heavy-Construction Vehicles and Urban Traffic Monitoring // *International Journal of Advanced Computer Science and Applications*. – 2022. – Vol. 13, № 3. – DOI 10.14569/ijacsa.2022.0130344.
34. Nojeng, S., Jaya, R., Murniati R. The Real Time Monitoring System and Early Detection of Unbalanced Load for the Three Phase Distribution Systems in Urban Areas // *International Journal Of Electrical Engineering And Intelligent Computing*. – 2023. – Vol. 1, № 1. – P. 1-6. – doi: 10.33387/ijeeic.v1i1.5623.
35. Конвергентный подход к мультиагентной обработке данных в процессе мониторинга дорожно-транспортной инфраструктуры / А. Г. Финогеев, А. А. Финогеев, А. М. Ляпин [и др.] // *Труды международного симпозиума "Надежность и качество"*. – 2018. – Т. 1. – С. 320-323.
36. Беляева, Е. Л. Проблемы социальной оценки проектов благоустройства (на примере пляжа в Строгино) // *Социология города*. – 2023. – № 1. – С. 62-82. – doi: 10.35211/19943520_2023_1_62.
37. Система поддержки принятия решений для рационального озеленения города на примере г. Ереван, Республика Армения / А. С. Акопов, А. Л. Бекларян, А. К. Сагателян [и др.] // *Программная инженерия*. – 2019. – Т. 10, № 2. – С. 87-96. – doi: 10.17587/prin.10.87-96.
38. Интеллектуальные технологии для поддержки принятия управленческих решений в сфере экологической безопасности транспорта в больших городах-портах / О. В. Ложкина, Г. Г. Рогозинский, В. Н. Ложкин [и др.]

др.] // Морские интеллектуальные технологии. – 2020. – № 2-1(48). – С. 133-141. – doi: 10.37220/MIT.2020.48.2.010.

39. Sztubecka, M., Skiba, M., Mrówczyńska, M., Bazan-Krzywoszanska, A. An innovative decision support system to improve the energy efficiency of buildings in urban areas [Электронный ресурс] // Remote Sensing. – 2020. – Vol. 12, № 2. – 259. – doi: 10.3390/rs12020259. – Режим доступа : <https://doi.org/10.3390/rs12020259> (дата обращения 16.08.2025).

40. Данилина, Н. В. Улицы малоэтажной застройки городов: проблемы и перспективы развития / Н. В. Данилина, И. Д. Теплова // Архитектура и строительство России. – 2022. – № 3(243). – С. 24-29.

41. Данилина, Н. В. Оценка доступности остановочных пунктов наземного городского пассажирского транспорта с использованием географических информационных систем / Н. В. Данилина, С. С. Руденко, Г. А. Язбердиева // Недвижимость: экономика, управление. – 2020. – № 1. – С. 36-41.

42. Spatial Monitoring and Classifying of Urban Deterioration in the Egyptian Cities Using Geographic Information System (GIS) Approach: A Case Study of Mansoura City, Egypt / M. A. Osman, K. Q. Tahooun, M. Shehata, M. A. Lamee // Civil Engineering and Architecture. – 2022. – Vol. 10, No. 5A. – P. 176-188. – doi: 10.13189/cea.2022.101409.

43. Саати, Т. Л. Об измерении неосязаемого. Подход к относительным измерениям на основе главного собственного вектора матрицы парных сравнений // Cloud of Science. – 2015. – Т. 2, № 1. – С. 5-39.

44. Munywoki, M. N., Singh, K. Application of GIS to monitor infrastructural development in Mombasa County, Kenya // South African Journal of Geomatics. – 2023. – Vol. 12, No. 2. – P. 145-157. – doi: 10.4314/sajg.v12i.2.3.

45. Ejikeme, J., Kenekukwu A., Onwuzuligbo Ch. Monitoring of Spatio-Temporal Dynamics of Air Pollution in Enugu Urban using Remote Sensing and GIS // European Journal of Environment and Earth Sciences. – 2023. – Vol. 4, № 2. – P. 22-31.

46. Olorunfemi, I. E., Fasinmirin, J. T., Olufayo, A. A., Komolafe, A. A. GIS and remote sensing-based analysis of the impacts of land use/land cover change (LULCC) on the environmental sustainability of Ekiti State, southwestern Nigeria // *Environment, Development and Sustainability*. – 2020. – Vol. 22, № 2. – P. 661-692. – doi: 10.1007/s10668-018-0214-z.
47. Modelling urban sprawl of the greater Port Harcourt city using remote sensing and geographic information system techniques / Ch. Onwuzuligbo, U. Ch. Okeke, O. E. Udochukwu [и др.] // *Earth Science Malaysia*. – 2023. – Vol. 7, № 1. – P. 36-42. – doi: 10.26480/esmy.01.2023.36.42.
48. Аль Савафи, М. Х. Геоинформационные технологии в градостроительной деятельности // *Вестник Белгородского государственного технологического университета им. В.Г. Шухова*. – 2021. – № 6. – С. 52-62. – doi: 10.34031/2071-7318-2021-6-6-52-62.
49. Каргашина, М. А. Система геоинформационного обеспечения градостроительной деятельности: современное состояние и ключевые аспекты развития // *ИнтерКарто. ИнтерГИС*. – 2018. – Т. 24, № 1. – С. 253-261. – doi: 10.24057/2414-9179-2018-1-24-253-261.
50. Ямашкин, С. А. Управление природно-социально-производственными системами посредством геопорталов // *Современные наукоемкие технологии*. – 2023. – № 7. – С. 122-128. – doi: 10.17513/snt.39705.
51. Анализ опыта в области проектирования, разработки, внедрения и эффективного использования цифровых инфраструктур пространственных данных в области устойчивого развития территорий / С. А. Ямашкин, А. А. Ямашкин, В. В. Занозин, А. Н. Бармин // *Астраханский вестник экологического образования*. – 2020. – № 1(55). – С. 107-128. – doi: 10.36698/2304-5957-2020-19-1-107-128.
52. Бершадский, А. М., Ямашкин С. А. Управление природно-социально-производственными системами на основе пространственных данных // *Современные наукоемкие технологии*. – 2023. – № 10. – С. 10-17. – doi: 10.17513/snt.39785.

53. Иващенко, А. В. Цифровой мониторинг социально-экономических показателей развития региона / А. В. Иващенко, Е. А. Додонова, И. Н. Дубинина // Перспективные информационные технологии (ПИТ 2022) : труды Международной научно-технической конференции, Самара, 18–21 апреля 2022 года / под редакцией С.А. Прохорова. – Самара: Издательство Самарского научного центра РАН, 2022. – С. 35-38.

54. Иващенко, А. В. Цифровая платформа для мониторинга, анализа и оценки показателей развития региона / А. В. Иващенко, И. Н. Дубинина, Е. А. Додонова // Информатика и вычислительная техника : Сборник научных трудов XIV Всероссийской научно-технической конференции аспирантов, студентов и молодых ученых, Ульяновск, 15–16 июня 2022 года. – Ульяновск: Ульяновский государственный технический университет, 2022. – С. 100-104.

55. Гедонистическая модель стоимости квартиры: Владивосток [Электронный ресурс] // Национальный исследовательский университет «Высшая школа экономики. – Режим доступа : <https://economics.hse.ru/io/news/843318942.html> (дата обращения 16.08.2025).

56. Лейфер, Л. А. Справочник оценщика недвижимости. В 4 т. / Л. А. Лейфер, Т. В. Крайникова. – Изд. 4-е, актуализированное и расш. – Нижний Новгород : ИНФОРМ-Оценка, 2016. – Т. 1-4.

57. Astafiev, S. A., Yi G. Real Estate Investment DECISION Support System // Известия вузов. Инвестиции. Строительство. Недвижимость. – 2018. – Vol. 8, № 1(24). – Р. 23-27.

58. Global Liveability Index 2024 [Электронный ресурс] // Economist Intelligence Unit, EIU. – Режим доступа : <https://www.eiu.com/n/campaigns/global-liveability-index-2024/> (дата обращения 16.05.2025).

59. Global Liveability Index 2023 Summary [Электронный ресурс] // Economist Intelligence Unit, EIU. – Режим доступа : https://ethz.ch/content/dam/ethz/special-interest/baug/irl/irl-dam/lehrveranstaltungen/bsc/site-management/Literatur/Economist_Jun-Global-Liveability-Index-2023.pdf (дата обращения 16.08.2025).

60. Monocle's Quality of Life Survey [Электронный ресурс] // Monocle. – Режим доступа : <https://monocle.com/search/quality-of-life-survey/> (дата обращения 16.08.2025).
61. Quality of Living City Ranking 2024 [Электронный ресурс] // Mercer. – Режим доступа : <https://www.mercer.com/insights/total-rewards/talent-mobility-insights/quality-of-living-city-ranking/> (дата обращения 16.05.2025).
62. Global Liveability Index 2025 [Электронный ресурс] // movingto. – Режим доступа : <https://www.movingto.io/global-liveability-index> (дата обращения 16.05.2025).
63. Cost of Living [Электронный ресурс] // Numbeo. – Режим доступа : <https://www.numbeo.com/cost-of-living/> (дата обращения 16.05.2025).
64. Трояновский В.С. Количественное определение показателей состояния городской среды [Электронный ресурс] // Мирпроект. – Режим доступа : <https://mirproekt.ru/wp-content/uploads/2022/08/Troyanovskij-V.S.-Kolichestvennoe-opredelenie-pokazatelej-kachestva-gorodskoj-sredy.pdf> (дата обращения 16.08.2025).
65. СП 42.13330.2016 «СНиП 2.07.01-89* Градостроительство. Планировка и застройка городских и сельских поселений» (Приказ Минстроя России от 30 декабря 2016 г. № 1034/пр) [Электронный ресурс] // Минстрой России. – Режим доступа : <https://minstroyrf.gov.ru/docs/14465/> (дата обращения 15.05.2026).
66. Местные нормативы градостроительного проектирования городского округа город-герой Волгоград [Электронный ресурс] // Официальный сайт администрации Волгограда. – Режим доступа : <https://www.volgadmin.ru/d/branches/grad/citybuilding/PP3> (дата обращения 15.05.2026).
67. Peng, F., McCallum A. Information extraction from research papers using conditional random fields // Information Processing & Management – 2006. – P. 56-77.

68. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications // Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics. – 2002. – P. 145-158.
69. Bird, S., Klein, E., Loper E. Natural Language Processing with Python // O'Reilly Media Inc. – 2009.
70. Tomita, M. LR Parsers for Natural Languages / M. Tomita // COLIGN : 10th International Conference on Computational Linguistics : Proceedings of COLIGN 84. – 1984. – P. 354-357.
71. Apache OpenNLP library [Электронный ресурс] // Apache OpenNLP. – Режим доступа : <https://opennlp.apache.org/> (дата обращения 16.05.2026).
72. Томита-парсер – Технологии Яндекса [Электронный ресурс] // Яндекс. – Режим доступа : <https://tech.yandex.ru/tomita> (дата обращения 16.06.2026).
73. Monthly monitoring of urban development and renewal at the block-level in China using Sentinel-2 time series / H. He, J. Yan, L. Liu [et al.] // Remote Sensing of Environment. – 2026. – Vol. 332. – P. 115070. – doi: 10.1016/j.rse.2025.115070.
74. Корнилова, А. В. Опрос жителей многоквартирных жилых домов о качестве капитального ремонта / А. В. Корнилова, В. А. Ермаков, А. В. Баулин // Строительство: наука и образование. – 2025. – Т. 15, № 3. – С. 126-140. – doi: 10.22227/2305-5502.2025.3.8.
75. Извлечение структурированного описания объектов недвижимости из пользовательских записей на естественном языке / И. С. Зеленский, Д. С. Донченко, Д. С. Парыгин [и др.] // Известия Волгоградского государственного технического университета. – 2017. – № 14(209). – С. 41-46.
76. Зеленский, И.С. Распознавание агентских объявлений об операциях с недвижимостью с использованием методов машинного обучения // International Journal of Open Information Technologies. – 2025. – Т. 13 , № 4. – С. 124-130.

77. Зеленский, И. С., Донченко, Д. С. Программный модуль сбора и анализа нечеткой информации о недвижимости из социальных сетей // XXII Региональная конференция молодых ученых Волгоградской области : Тезисы докладов, Волгоград, 21–24 ноября 2017 года / Ответственный редактор А.В. Навроцкий. – Волгоград: Волгоградский государственный технический университет, 2017. – С. 182-183.

78. Loshin, D. ETL (Extract, Transform, Load) // Business Intelligence. – 2nd. – Morgan Kaufmann, 2012. – 400 p. – ISBN 978-0-12-385890-0.

79. Validation of real estate ads based on the identification of identical images / A. Golubev, I. Zelenskiy, D. Parygin [и др.] // Proceedings of the 2018 International Conference on System Modeling and Advancement in Research Trends, SMART 2018 : 7, Moradabad, 23–24 ноября 2018 года. – Moradabad, 2018. – P. 274-279. – doi: 10.1109/SYSMART.2018.8746926.

80. Ананьев, Ю. С. Геоинформационные системы. Учеб. пособие. – Томск: Изд. ТПУ, 2003. – 70 с.

81. Интеллектуальная поддержка решений по использованию объектов недвижимости для управления урбанизированными территориями / И. С. Зеленский, Д. С. Парыгин, О. В. Савина [и др.] // International Journal of Open Information Technologies. – 2020. – Т. 8, № 11. – С. 13-29.

82. Никулин, В. В., Ладанова, Е. О. Использование Томита-парсера для извлечения фактов из тестовых данных // XLVII Огарёвские чтения : Материалы научной конференции. В 3-х частях, Саранск, 06–13 декабря 2018 года / Составитель А.В. Столяров. Ответственный за выпуск П.В. Сенин. Том Часть 1. – Саранск: Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва, 2019. – С. 343-345.

83. IBM Watson [Электронный ресурс] // IBM. – Режим доступа : <https://www.ibm.com/watson/> (дата обращения 16.05.2026).

84. Wolfram Natural Language Understanding System [Электронный ресурс] // Wolfram. – Режим доступа : <https://www.wolfram.com/natural-language-understanding/> (дата обращения 16.05.2026).

85. Томита-парсер, руководство разработчика [Электронный ресурс] // Яндекс. – Режим доступа : <https://tech.yandex.ru/tomita/doc/dg/concept/about-docpage/> (дата обращения 16.05.2026).
86. Хопкрофт, Д., Мотвани, Р., Ульман, Д. Введение в теорию автоматов, языков и вычислений, 2-е изд. : Пер. с англ. – Москва, Издательский дом «Вильямс», 2002. – 528 с. – ISBN 5-8459-0261-4.
87. Chomsky, N. Three models for the description of language // IRE Transactions on Information Theory. – 1956. – Vol. 2, № 3. – p. 113–124. – doi:10.1109/TIT.1956.1056813.
88. Ситникова, Л. К., Еланцев, М. О., Султанов, Р. О. Двухэтапный метод поиска похожих изображений в большой базе изображений // Выставка инноваций - 2022 (весенняя сессия) : Сборник материалов XXXIII Республиканской выставки-сессии студенческих инновационных проектов, Ижевск, 29 апреля 2022 года. – Ижевск: Ижевский государственный технический университет имени М.Т. Калашникова, 2022. – С. 304-309. – doi: 10.22213/ie022143.
89. Новицкая, А. Д., Жилик, Н. А. Методы распознавания и отслеживания изменения объекта с течением времени // Труды БГТУ. Серия 3: Физико-математические науки и информатика. – 2017. – № 1 (194). – С. 79–81.
90. Gonzalez, R. C., Woods R. E., Eddins, S. L. Digital image processing using matlab. – Editorial Pearson-Prentice Hall, 2004. – 597 p. – ISBN 0-13-008519-7.
91. Трефилов, П. А. Хранение и поиск схожих изображений в темпоральных базах данных с использованием перцептивных хэш-строк / П. А. Трефилов // Труды международного симпозиума "Надежность и качество". – 2020. – Т. 1. – С. 192-196.
92. Огневой, Г. Д. Методы и алгоритмы поиска изображений / Г. Д. Огневой // Информационные технологии в образовании, науке и производстве : II Международная научно-техническая интернет-конференция, 4 декабря 2014 г. Секция: Информационные технологии в производстве и научных

исследованиях [Электронный ресурс] // Репозиторий Белорусского национального технического университета. – Режим доступа : <https://rep.bntu.by/handle/data/12606> (дата обращения 16.05.2026).

93. Parsing of Data on Real Estate Objects from Network Resource / V. N. Cherkesov, V. P. Malikov, A. V. Golubev [и др.] // Information Technologies in Science, Management, Social Sphere and Medicine : Proceedings of the IV International research conference "Information technologies in Science, Management, Social sphere and Medicine" (ITSMSSM 2017), Tomsk, 05–07 декабря 2017 года. Vol. 72. – Shenzhen: Atlantis Press, 2017. – P. 385-388. – doi: 10.2991/itsmssm-17.2017.80.

94. Валишин, А. А., Запривода, А. В., Цухло, С. С. Моделирование и сравнительный анализ эффективности перцептивных хеш-функций для поиска сегментированных изображений // Математическое моделирование и численные методы. – 2024. – № 2. – С. 46-67.

95. Elmagarmid, A. K., Ipeirotis P. G., Verykios V. S. Duplicate Record Detection: A Survey // IEEE Transactions on Knowledge and Data Engineering. – 2007. – Vol. 19, № 1. – P. 1-16. – doi: 10.1109/TKDE.2007.250581.

96. Заде, Л. Основы нового подхода к анализу сложных систем и процессов принятия решений (пер. с англ. В. А. Горелик, С. А. Орловский, Н. И. Ринго) // Математика сегодня. – М. : Знание, 1974. – с. 5-49.

97. Журавлев, Ю. И., Рязанов, В. В., Сенько, О. В. «Распознавание». Математические методы. Программная система. Практические применения. – Москва: ФАЗИС, 2006 (М. : Типография "Наука" РАН). – 176 с.

98. Горенбургов, М. А., Гончаров, В. В. Выработка правил и форм изложения бизнес-информации в сети интернет как средство противодействия мошенническим схемам // Региональная информатика (РИ-2020): XVII Санкт-Петербургская международная конференция. Материалы конференции. – 2020. – Т. 1. – С. 296–297.

99. Епрынцева, Н. А., Соколова, А. В., Руднева, А. А. Искусственный интеллект в сфере недвижимости // Информационные технологии в

строительных, социальных и экономических системах. – 2018. – № 4(14). – С. 47–50.

100. An Intelligent Recommendation System for Real Estate Commodity / Ts. Y. Ou, G. Yu. Lin, H. P. Fu. [и др.] // Computer Systems Science and Engineering. – 2022. – Vol. 42, № 3. – P. 881–897. – doi: 10.32604/csse.2022.022637.

101. Sinyak, N. G., Tajinder, S., Madhu, K. Ja., Kozlovskiy, V. V. Predicting real estate market trends and value using pre-processing and sentiment text mining analysis // Real Estate: Economics, Management. – 2021. – № 1. – P. 35–43.

102. Automatic Online Fake News Detection Combining Content and Social Signals / M. L. D. Vedova, E. Tacchin, S. Moret [и др.] // Conference of Open Innovations Association, FRUCT. – 2018. – № 22. – P. 272–279.

103. Detection of Fake News Using Machine Learning and Natural Language Processing Algorithms / N. N. Prachi, Md. Habibullah, Md. E. H. Rafi [и др.] // Journal of Advances in Information Technology. – 2022. – Vol. 13, № 6. – P. 652–661. – doi: 10.12720/jait.13.6.652-661.

104. Mohapatra, A., Thota, N., Prakasam, P. Fake news detection and classification using hybrid BiLSTM and self-attention model // Multimedia Tools and Applications. – 2022. – Vol. 81, № 13. – P. 18503–18519. – doi: 10.1007/s11042-022-12764-9.

105. Anderson, J. R. Machine Learning: An Artificial Intelligence Approach. – Elsevier Science, 1983. – 572 p.

106. Gasparetto, A., Marcuzzo, M., Zangari, A., Albarelli, A. A Survey on Text Classification Algorithms: From Text to Predictions [Электронный ресурс] // Information. – 2022. – Vol. 13, № 2. – 83. – doi: 10.3390/info13020083. – Режим доступа : <https://doi.org/10.3390/info13020083> (дата обращения 16.05.2026).

107. Wadud, Md. A. H., Mridha, M. F., Rahman, M. M. Word Embedding Methods for Word Representation in Deep Learning for Natural Language Processing // Iraqi Journal of Science. – 2022. – Vol. 63, No. 3. – P. 1349–1361. – doi: 10.24996/ij.s.2022.63.3.37

108. Rakshit, P., Sarkar, A., A supervised deep learning-based sentiment analysis by the implementation of Word2Vec and GloVe Embedding techniques // *Multimedia Tools and Applications*. – 2025. – Vol. 84. – P. 979–1012 – doi: 10.1007/s11042-024-19045-7.

109. Паршакова, Л. И. Сравнительный подход к оценке недвижимости // *Студенческая наука - взгляд в будущее : Материалы XVI Всероссийской студенческой научной конференции*, Красноярск, 26 марта 2021 года. Том Часть 2. – Красноярск : Красноярский государственный аграрный университет, 2021. – С. 222-224.

110. Эпова, Д. Р., Каниковская, А. Р. Применение сравнительного подхода к оценке недвижимости // *Неделя науки СПбПУ : Материалы научной конференции с международным участием*. Институт промышленного менеджмента, экономики и торговли. В 3-х частях, Санкт-Петербург, 18–23 ноября 2019 года. Том Часть 1. – Санкт-Петербург: Федеральное государственное автономное образовательное учреждение высшего образования "Санкт-Петербургский политехнический университет Петра Великого", 2019. – С. 403-405.

111. Aggarwal, C. C., Reddy, C. K. *Data Clustering : Algorithms and Applications*. – Chapman & Hall, 2013. – 652 p. – ISBN 978-1-315-37351-5

112. Cornwell, B. *Social Sequence Analysis, Methods and Applications*. – Cambridge University Press, 2015. – 315 p.

113. Upton, G., Cook, I. *Understanding Statistics*. – Oxford University Press, 1996. – 657 p.

114. Единая классификация многоквартирных жилых новостроек [Электронный ресурс] // Российская гильдия риэлторов. – Режим доступа : https://www.fondrgs.ru/files/docs/Edinaya_klassifikaciya1.pdf (дата обращения 16.05.2026).

115. Стерник, Г. М. Классификация офисной недвижимости [Электронный ресурс] // Рынок недвижимости России. – Режим доступа : http://realtymarket.org/docs/lib_58.htm (дата обращения 16.05.2026).

116. Зеленский, И. С. Расчёт рейтинга объектов недвижимости на основе нормативов и пользовательских предпочтений / И. С. Зеленский, Д. С. Парыгин, Т. В. Смирнова // Моделирование, оптимизация и информационные технологии. – 2019. – Т. 7, № 1(24). – С. 221-233. – doi: 10.26102/2310-6018/2019.24.1.008.

117. Apartments for Rent - Your Trusted Apartment Finder Tool [Электронный ресурс] // Apartmentguide. – Режим доступа : <https://www.apartmentguide.com/> (дата обращения 16.05.2026).

118. Apartment Ratings, Prices and Reviews [Электронный ресурс]. – Режим доступа : <https://www.apartmentratings.com/> (дата обрац. 06.05.2026).

119. Азгальдов, Г. Г., Азгальдова, Л. А. Количественная оценка качества (Квалиметрия). Библиография. – М. : Издательство стандартов, 1971. — 176с.

120. Елизарова, М. И., Хрусталёв, О. Е., Кураева, О. А. Методологический подход к оценке состояния и качества содержания объектов недвижимости // Вопросы экономики. – 2014. – № 21 (207). – С. 14-23.

121. Мищенко, В. В., Мищенко, Л. А. Мищенко, В. В. Вопросы применения сравнительного подхода при оценке недвижимости // Вестник Кузбасского государственного технического университета. – 2014. – № 4(104). – С. 158-161.

122. Иванова, Е. Н. Оценка стоимости недвижимости. Учебное пособие / Е. Н. Иванова; под редакцией М. А. Федотовой. – М. : КНОРУС, 2007 — 344 с.

123. European Standards for Statistical Valuation Methods. 2nd Edition [Электронный ресурс] // European AVM Alliance – Режим доступа : <https://www.europeanavmalliance.org/avm-standards.html> (дата обращения 16.05.2026).

124. Robson, G., Downie M. L. Automated Valuation Models: an international perspective. – CML Research, 2007. – 78 p.

125. Петровский А. Б. Теория принятия решений : учебник для студ. высш. учеб. заведений. – М. : Издательский центр «Академия», 2009. – 400 с. – (Университетский учебник. Сер. Прикладная математика и информатика).

126. Зеленский, И. С., Парыгин, Д. С., Савина, О. В. Особенности проектирования рекомендательных систем комплексного развития городских территорий // XIV Всероссийское совещание по проблемам управления : сборник научных трудов, Москва, 17–20 июня 2024 года. – Москва: Институт проблем управления им. В.А. Трапезникова РАН, 2024. – С. 2641-2645.

127. Effective Implementation of Integrated Area Development Based on Consumer Attractiveness Assessment / I. Zelenskiy, D. S. Parygin, O. Savina [и др.] // Sustainability. – 2022. – Vol. 14, № 23. – P. 16239. – doi: 10.3390/su142316239.

128. Подиновский, В. В. Введение в теорию важности критериев в многокритериальных задачах принятия решений. – М. : ФИЗМАТЛИТ, 2007. – 64 с. – ISBN 978-5-9221-0743-3.

129. Lotfi, F. H., Fallahnejad, R. Imprecise Shannon's entropy and multi attribute decision making // Entropy. – 2010. – Vol. 12, № 1. – P. 53–62.

130. Машунин, К. Ю., Машунин, Ю. К. Векторная оптимизация с равнозначными и приоритетными критериями // Известия Российской академии наук. Теория и системы управления. – 2017. – № 6. – С. 80–99.

131. Wang, Y. M., Luo, Y. Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making // Mathematical and Computer Modelling. – 2010. – Vol. 51, № 1–2. – P. 1–12.

132. Саати, Т. Л. Относительное измерение и его обобщение в принятии решений. Почему парные сравнения являются ключевыми в математике для измерения неосязаемых факторов // Cloud of Science. – 2016. – Т. 3, № 2. – С. 171-262.

133. Попов, Е. В., Кац, И. С., Веретенникова, А. Ю. Доступность социальной инфраструктуры городских территорий // Региональная экономика: теория и практика. – 2016. – № 2(425). – С. 54-67.

134. Иванова, Н. А. Классификация объектов региональной инфраструктуры // Современные наукоемкие технологии. Региональное приложение. – 2014. – № 1. – С. 38-43.

135. Сычева, И. В., Сычева, Н. А. Исследование содержания категории "Социальная инфраструктура" // Известия Тульского государственного университета. Экономические и юридические науки. – 2012. – № 2-1. – С. 230-238.
136. ImageHash [Электронный ресурс] // PyPI – the Python Package Index. – Режим доступа : <https://pypi.org/project/ImageHash/> (дата обращения 16.05.2026).
137. opencv-python [Электронный ресурс] // PyPI – the Python Package Index. – Режим доступа : <https://pypi.org/project/opencv-python/> (дата обращения 16.05.2026).
138. Жамбю, М. Иерархический кластер-анализ и соответствия. – М. : Финансы и статистика, 1988. — 345 с.
139. Halkidi, M., Batistakis, Y., Vazirgiannis, M. On Clustering Validation Techniques // Journal of Intelligent Information Systems. – 2001. – Vol. 17, No. 2-3. – P. 107-145.
140. van der Maaten L., Hinton G. Visualizing Data Using t-SNE // Journal of Machine Learning Research. – 2008. – Vol. 9. – P. 2579-2605.
141. Vukelic, B., Baca, M. Comparison of RADIAL variance based and Maar-Hildreth operator perceptual image hash functions on biometric templates // Central European Conference on Information and Intelligent Systems, Varazdin, 2014 - Varazdin: Faculty of Organization and Informatics, 2014. – P. 286-292.
142. Зеленский, И. С. Разработка системы рейтингования жилой недвижимости // Смотр-конкурс научных, конструкторских и технологических работ студентов Волгоградского государственного технического университета (г. Волгоград, 13-17 мая 2019 г.) : тез. докл. / редкол.: С. В. Кузьмин (отв. ред.) [и др.] ; Волгоградский гос. техн. ун-т, Совет СНТО. – Волгоград, 2019. – С. 167.
143. «Земельный кодекс Российской Федерации» от 25.10.2001 N 136-ФЗ (ред. от 04.05.2026) [Электронный ресурс]. – Режим доступа :

https://www.consultant.ru/document/cons_doc_LAW_33773/ (дата обращения 16.05.2026).

144. Прокофьева, Л. М., Корчагина, И. И. Демографическая структура семей и домохозяйств в России, её динамика по данным переписей населения // Демографическое обозрение. – 2023. – Т. 10, № 2. – С. 4-17. – doi: 10.17323/demreview.v10i2.17763.

Приложение А – Схема базы данных разработанной системы

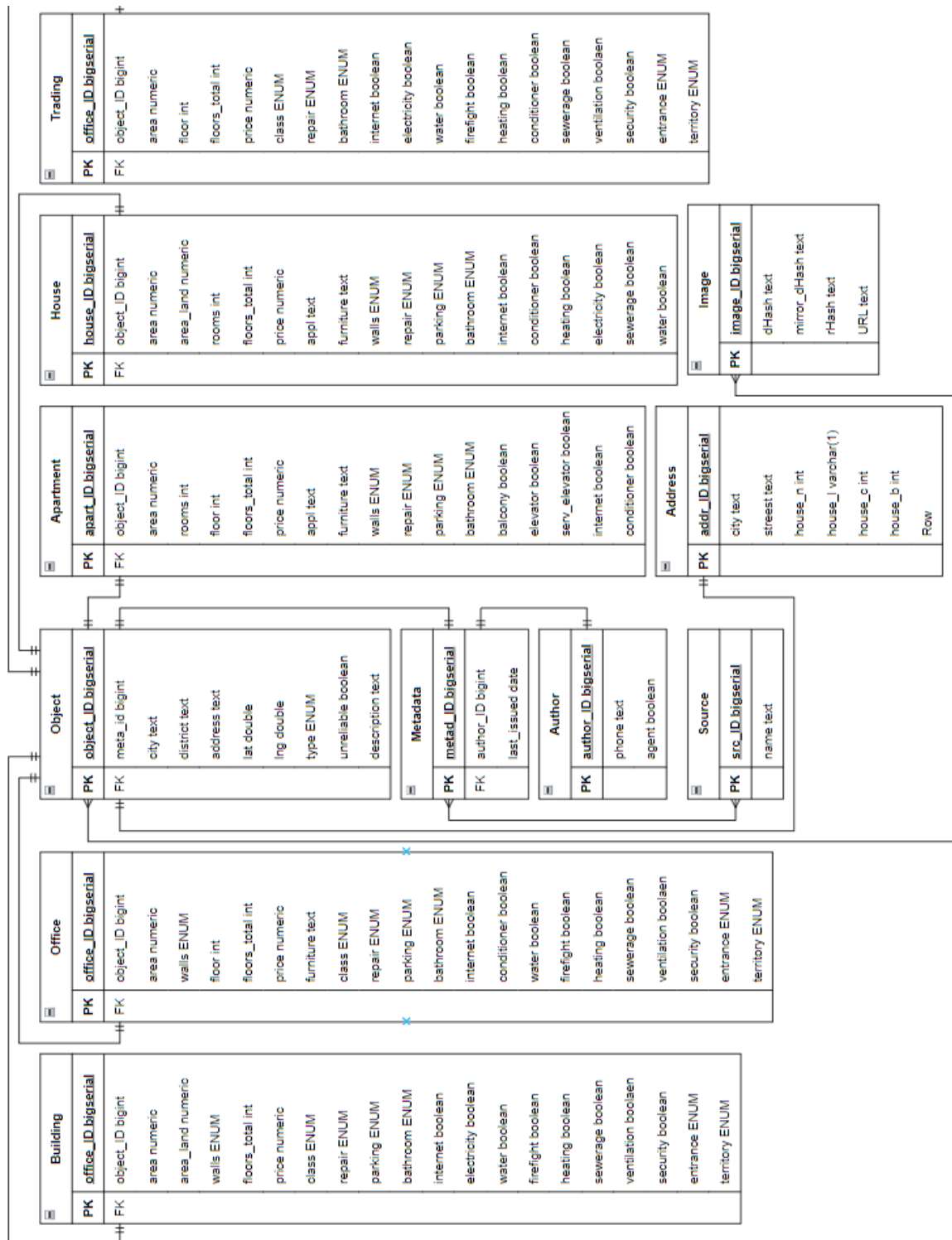


Рисунок А.1 – Схема базы данных прототипа системы оценки состояния городской среды

Приложение Б – Перечень характеристик в структурированных описаниях объектов городской среды

Таблица Б.1 – Перечень характеристик в структурированных описаниях объектов городской среды

Характеристика	Возможные значения /Единицы измерения	Тип данных	Квартира	Дом	Офис	Торговая площадь	Отдельное здание
Город	название города	string	+	+	+	+	+
Район	название района	string	+	+	+	+	+
Адрес	улица, дом	string	+	+	+	+	+
Материал стен	кирпич, монолит и др.	enum	+	+	-	-	+
Класс здания	А, А+, В и т.п.	enum	-	-	+	+	+
Цена	сумма в рублях	numeric	+	+	+	+	+
Этаж	целое число	int	+	-	+	+	-
Этажность здания	целое число	int	+	+	+	+	+
Число комнат	целое число	int	+	+	-	-	-
Площадь	кв.м.	numeric	+	+	+	+	+
Площадь участка	кв.м.	numeric	-	+	-	-	+
Санузел	совмещенный/раздельный/отсутствует	enum	+	+	+	+	+
Балкон/лоджия	наличие/отсутствие	boolean	+	-	-	-	-
Пассажирский лифт	наличие/отсутствие	boolean	+	-	-	-	-
Грузовой лифт	наличие/отсутствие	boolean	+	-	-	-	-
Парковка	организованная/стихийная	enum	+	-	+	-	+
Парковочное место	наличие/отсутствие	boolean	-	+	-	-	-
Ремонт	евроремонт, типовой и т.п.	enum	+	+	+	+	+
Мебель	список	string	+	+	+	-	-
Бытовая техника	список	string	+	+	-	-	-

Продолжение таблицы Б.1

Характеристика	Возможные значения /Единицы измерения	Тип данных	Квартира	Дом	Офис	Торговая площадь	Отдельное здание
Интернет	наличие/отсутствие	boolean	+	+	+	+	+
Кондиционер	наличие/отсутствие	boolean	+	+	+	-	-
Отопление	наличие/отсутствие	boolean	-	+	+	+	+
Электро-снабжение	наличие/отсутствие	boolean	-	+	-	+	+
Система пожаро-тушения	наличие/отсутствие	boolean	-	-	+	+	+
Канализация	наличие/отсутствие	boolean	-	+	+	+	+
Водоснабжение	наличие/отсутствие	boolean	-	+	+	+	+
Вентиляция	наличие/отсутствие	boolean	-	-	+	+	+
Охрана	наличие/отсутствие	boolean	-	-	+	+	+
Вход	свободный/ по пропускам	enum	-	-	+	+	+
Планировка	открытая/ кабинетная	enum	-	-	+	-	-
Территория	зарытая/открытая	enum	-	-	+	+	+
Телефон автора	В формате 8xxxxxxxxxxx	string	+	+	+	+	+

Приложение В – Конфигурации оценки потребительских

Таблица В.1 – Конфигурации оценки потребительских свойств

Критерий оценки	Тип шкалы	Способ оценки	Квартиры	Частные жилые дома	Офисы	Торговые площади	Отдельно стоящие здания
Материал стен	Номинальная	Корректировка	Кирпич – 1,00 Монолит – 0,95 Панели – 0,92	Кирпич – 1,00 Пенобетон – 0,88 Дерево – 0,84	Кирпич – 1,00 Бетон – 1,00 Панели – 0,86 Дерево – 0,76		
Район города			Центральный – 1,00; Ворошиловский – 0,98; Дзержинский – 0,97; Советский – 0,95; Тракторозаводский – 0,94; Кировский – 0,93; Краснооктябрьский – 0,92; Красноармейский – 0,93				
Этаж			Первый – 0,93 Последний – 0,94 Другие – 1,00	-	Подвал – 0,71 Цоколь – 0,82 Первый этаж – 1,00 Другие – 0,87	-	
Ремонт	Номинальная	Корректировка	Нет данных – 0 Требуется – 0,80 Сделан – 1,00		Требуется – 0,79 Не требуется – 1,00 Люкс – 1,22		
Площадь	Числовая	Прямая оценка	Достаточный уровень – 72,6 кв. м.		-		
Площадь участка			-	Достаточный уровень – 300 кв. м.	-		
Парковочное место	Номинальная (двоичная)	Прямая оценка	-	Нет данных – 0 Отсутствует – 0 В наличии – 1	-		

Продолжение таблицы В.1.

Критерий оценки	Тип шкалы	Способ оценки	Квартиры	Частные жилые дома	Офисы	Торговые площади	Отдельно стоящие здания
Санузел	Номинальная	Прямая оценка	Нет данных – 0 На улице – 0 Совмещенный – 0,5 Раздельный – 1				
Электричество	Номинальная (двоичная)	Прямая оценка	-		Нет данных – 0 Отсутствует – 0 В наличии – 1		
Водоснабжение							
Отопление							
Канализация							
Балкон/лоджия			Нет данных – 0 Отсутствует – 0 В наличии – 1		-		
Пассажирский лифт							
Грузовой лифт							
Мебель			Нет данных – 0 Отсутствует – 0 В наличии – 1			-	
Бытовая техника							
Интернет			Нет данных – 0 Отсутствует – 0 В наличии – 1				
Кондиционер							
Прод. магазины							
Аптеки рядом							
Вентиляция			-		Нет данных – 0 Отсутствует – 0 В наличии – 1		
Система пожаротушения							

Продолжение таблицы В.1.

Критерий оценки	Тип шкалы	Способ оценки	Квартиры	Частные жилые дома	Офисы	Торговые площади	Отдельно стоящие здания
Охрана	Номинальная (двоичная)	Прямая оценка	-	-	Нет данных – 0 Отсутствует – 0 В наличии – 1		
Класс здания	Номинальная	Корректировка	-	-	В и выше – 1,27 Другие – 1,00		
Вход	Номинальная (двоичная		-	-	Отдельный – 1,00 Общий – 0,86 Нет данных – 0,86		
Территория			Огороженная – 1,00 Свободный проход – 0,94	-	Закрытая – 0,77 Свободный доступ – 1,00 Нет данных – 1,00		
Остановки транспорта			Рядом – 1,00 На удалении – 0,94	Рядом – 1,00 На удалении – 0,93	Рядом – 1,00 На удалении – 0,85		
Парковка	Номинальная (двоичная)		Корректировка	Организованная – 1,05 Стихийная – 1,00 Нет данных – 1,00	-	Организованная – 1,11 Стихийная – 1,00 Нет данных – 0,91	
Объекты, нарушающие экологию		Рядом – 0,84 На удалении – 1,00		-			

Продолжение таблицы В.1.

[illegible]

Приложение Г – Копии свидетельств о регистрации программ для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

САПР
47 нр/18



СВИДЕТЕЛЬСТВО
о государственной регистрации программы для ЭВМ
№ 2018662442

**Программный модуль для сравнительного анализа и
выявления дубликатов изображений**

Правообладатель: **Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Волгоградский государственный технический университет»
(ВолГТУ) (RU)**

Авторы: **Зеленский Илья Сергеевич (RU), Голубев Алексей
Владимирович (RU), Парыгин Данила Сергеевич (RU)**

Заявка № **2018619346**
Дата поступления **03 сентября 2018 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ **08 октября 2018 г.**



Руководитель Федеральной службы
по интеллектуальной собственности

 Г.П. Излиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ



СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2020666421

Программный комплекс для массовой оценки жилых помещений сравнительным подходом

Правообладатели: *Бойко Дмитрий Олегович (RU), Парыгин Данила Сергеевич (RU), Савина Оксана Владимировна (RU), Голубев Алексей Владимирович (RU), Зеленский Илья Сергеевич (RU)*

Авторы: *Бойко Дмитрий Олегович (RU), Парыгин Данила Сергеевич (RU), Савина Оксана Владимировна (RU), Голубев Алексей Владимирович (RU), Зеленский Илья Сергеевич (RU)*



Заявка № **2020665667**

Дата поступления **01 декабря 2020 г.**

Дата государственной регистрации

в Реестре программ для ЭВМ **09 декабря 2020 г.**

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ



СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024685595

**Программа генеративного проектирования
комплексного развития городских территорий на основе
знаний из онтологической модели**

Правообладатель: *Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Волгоградский государственный технический университет»
(RU)*

Авторы: *Соколов Денис Александрович (RU), Рашевский Николай
Михайлович (RU), Чикин Артем Дмитриевич (RU), Зеленский
Илья Сергеевич (RU), Трудов Ярослав Анатольевич (RU)*

Заявка № 2024685147

Дата поступления 30 октября 2024 г.

Дата государственной регистрации

в Реестре программ для ЭВМ 30 октября 2024 г.



Руководитель Федеральной службы
по интеллектуальной собственности

документ подписан электронной подписью
Сертификат 0692e7c10e300b154f2401670bco2026
Владелец: **Зубов Юрий Сергеевич**
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

Приложение Д – Копия акта внедрения результатов работы



СтройБилд
строим для вас

ООО «СтройБилд»

тел./факс 8 (8442) 73-81-20 e-mail: sbild@yandex.ru

400016 г. Волгоград, проспект Волжский, 4к

ИНН 3459010288 КПП 345901001 ОГРН 1143443014802

УТВЕРЖДАЮ

Директор

С.А. Чурилов

«22» мая 2025 г.

АКТ

о внедрении результатов диссертационной работы
на соискание учёной степени кандидата технических наук

Настоящим актом подтверждаем, что разработанные Зеленским Ильёй Сергеевичем методы извлечения из гетерогенных источников, очистки и обогащения данных о состоянии городской среды и алгоритмы оценки состояния городской среды по критериям потребительских свойств жилых и коммерческих объектов, а также обеспеченности социальной инфраструктурой (образованием и здравоохранением) на основе собранных данных, описанные в его диссертационной работе, легли в основу работ по оценке городской среды г. Волгограда, проводимых ООО «СтройБилд».

Разработанные Зеленским И.С. методы и алгоритмы, реализованные в виде программного обеспечения, использовались для решения задач, связанных с оценкой потребительских свойств жилых и коммерческих объектов и оценкой загруженности социальной инфраструктуры (школ, детских садов, поликлиник). Полученные результаты позволяют утверждать, что качество предпроектного анализа территорий было улучшено за счет получения более полной и актуальной информации.

Внедрение системы мониторинга состояния городской среды для поддержки принятия решений по выбору места строительства и конфигурации новых жилых и нежилых зданий, позволяет повысить эффективность и обоснованность принимаемых решений и сократить затраты времени на анализ состояния городской среды.

Директор ООО «СтройБилд»
печать/подпись



С.А. Чурилов