

Восточно-Казахстанский технический университет им. Д. Серикбаева

УДК 004:550.34

На правах рукописи

КАРМЕНОВА МАРХАБА АХМЕТОЛЛИНОВНА

**Разработка методов оценки сейсмостойкости городских объектов на
основе применения технологии анализа данных**

6D070300 – Информационные системы (по отраслям)

Диссертация на соискание степени
доктора философии (PhD)

Отечественные научные консультанты
кандидат физико-математических наук,
ассоц. профессор
Г.В. Попова

доктор PhD,
ассоц. профессор
А.С. Глебалдинова

Зарубежный научный консультант
Prof. dr hab., inż.
W. Wójcik

Республика Казахстан
Усть-Каменогорск, 2021

СОДЕРЖАНИЕ

НОРМАТИВНЫЕ ССЫЛКИ.....	4
ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ.....	5
ВВЕДЕНИЕ.....	6
1 ОБЗОР СОВРЕМЕННОГО СОСТОЯНИЯ ИССЛЕДОВАНИЙ ПО ОЦЕНКЕ СЕЙСМИЧЕСКОГО РИСКА.....	12
1.1 Современное состояние исследований и актуальные проблемы обеспечения сейсмической безопасности.....	12
1.2 Анализ современных методов и моделей в оценке сейсмических рисков на основе технологии анализа данных и методов машинного обучения.....	17
1.3 Существующие информационные технологии и инструментальные средства для обработки и аналитики сейсмических данных.....	31
Выводы по первому разделу.....	36
2 РАЗРАБОТКА МЕТОДОВ ОЦЕНКИ СЕЙСМОСТОЙКОСТИ ГОРОДСКИХ ОБЪЕКТОВ НА ОСНОВЕ ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ И МАШИННОГО ОБУЧЕНИЯ.....	38
2.1 Применение технологии анализа данных и методов машинного обучения для оценки сейсмостойкости городских объектов.....	38
2.2 Разработка функциональной модели оценки сейсмостойкости городских объектов.....	42
2.2.1 Оценка сейсмостойкости городских объектов методами кластерного анализа.....	44
2.2.2 Оценка сейсмостойкости городских объектов методом Decision Tree.....	55
2.2.3 Выявление информативных признаков оценки сейсмостойкости городских объектов Ensemble-методом: Random Forest.....	60
Выводы по второму разделу.....	67
3 МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА ПРОСТРАНСТВЕННЫХ ДАННЫХ НА ПРИМЕРЕ ИССЛЕДОВАНИЯ СЕЙСМОАКТИВНОСТИ РЕСПУБЛИКИ КАЗАХСТАН.....	69
3.1 Статистический анализ каталога землетрясений, зарегистрированных на территории Республики Казахстан.....	69
3.2 Оценка сейсмической опасности алгоритмом пространственной кластеризации на основе плотности с присутствием шума DBSCAN.....	79
3.3 Результаты экспериментальных исследований применения алгоритма DBSCAN для кластерного анализа набора сейсмических данных.....	81
Выводы по третьему разделу.....	85
4 АРХИТЕКТУРА И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ИНТЕЛЛЕКТУАЛЬНОЙ ИНФОРМАЦИОННО- АНАЛИТИЧЕСКОЙ СИСТЕМЫ (ИИАС) ОЦЕНКИ СЕЙСМОСТОЙКОСТИ ГОРОДСКИХ ОБЪЕКТОВ С УЧЕТОМ СЕЙСМИЧЕСКИХ ДАННЫХ.....	87

4.1	Общее описание и концептуальная основа ИИАС.....	87
4.2	Инструментарий и технологии анализа данных в архитектуре ИИАС	89
4.3	Структура и описание данных.....	92
4.4	Проектирование архитектуры и программный интерфейс ИИАС.....	96
	Выводы по четвертому разделу.....	106
	ЗАКЛЮЧЕНИЕ	107
	СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	108
	ПРИЛОЖЕНИЕ А – Свидетельство о государственной регистрации.....	120
	ПРИЛОЖЕНИЕ Б – Акт внедрения результатов.....	122

НОРМАТИВНЫЕ ССЫЛКИ

В настоящей диссертации использованы ссылки на следующие стандарты:

Инструкция по оформлению диссертации и автореферата. МОН РК, Внешний аттестационный комитет. – Алматы, 2004.

Закон Республики Казахстан. О науке: принят 18 февраля 2011 года, №407-IV.

ГОСТ 7.1-2003. Библиографическая запись. Библиографическое описание. Общие требования и правила составления.

СП РК 1.01-101-2014. Строительная терминология.

ГОСТ Р 22.0.11-99 Безопасность в чрезвычайных ситуациях. Предупреждение природных чрезвычайных ситуаций. Термины и определения.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

РК	– Республика Казахстан
BIM	– Building Information Modeling или Building Information Model, (информационное моделирование здания или информационная модель здания)
ФГБОУ ВО	– Федеральное государственное бюджетное образовательное учреждение высшего образования
ЭВМ	– Электронно-вычислительная машина
НАО	– Некоммерческое акционерное общество
СОП	– Совместная образовательная программа
МОН РК	– Министерство образования и науки Республики Казахстан
DBSCAN	– Density-based spatial clustering of applications with noise (пространственная кластеризация на основе плотности с присутствием шума)
ИИАС	– Интеллектуальная информационно-аналитическая система
IT	– Information technology (информационные технологии)
ГИС	– Географическая информационная система
ЭС	– Экспертная система
ДЗ	– Дистанционное зондирование
RFID	– Radio Frequency Identification (радиочастотная идентификация)
IDEFO	– function modeling (методология функционального моделирования)
DT	– Decision Tree (дерево решений)
RF	– Random Forest (случайный лес)
СОМЭ	– Сейсмологическая опытно-методическая экспедиция Комитета науки Министерства образования и науки Республики Казахстан
MPVA	– региональная магнитуда по объемным волнам, определяемая по Р-волне, зарегистрированной короткопериодными приборами
ISC-GEM	– International Seismological Centre-Global Earthquake Model (Международный сейсмологический центр-Глобальное инструментальное землетрясение)
PCУБД	– Реляционная система управления базами данных
API	– Application-programming interface (программный интерфейс приложения, интерфейс прикладного программирования)
HDFS	– Hadoop Distributed File System, распределенная файловая система Hadoop для хранения файлов больших размеров
НД	– Набор данных
СУБД	– Система управления базами данных

ВВЕДЕНИЕ

Актуальность работы. Благополучное процветание и динамичное развитие РК по сегодняшний день, связано с качественной разработкой, системным внедрением и эффективной реализацией на различных уровнях и сроках государственных программ и стратегических планов, охватывающие практически все сферы человеческой жизнедеятельности. Мировые тенденции, научные достижения и технологические инновационные решения способствуют росту и процветанию основополагающих отраслей нашей страны, а также позволяют быть конкурентноспособной страной в мире.

В долгосрочной перспективе (Стратегия «Казахстан-2050») сценариев развития нашей страны [1] и в намеченном курсе по вхождению в число 30 наиболее развитых государств мира, особое внимание уделяется индустриализации, внедрению современных технологий в строительстве и коммунальном секторе, развитию концепции умного города [2]. В целом, как отмечено в государственной программе жилищного строительства «Нұрлы жер», строительная отрасль РК на сегодняшний день является одной из стремительно развивающейся отраслью нашей страны, тем самым оказывая значимое воздействие на социально-экономическое развитие страны и ее регионов [3]. Кроме того, в фокусе внимания также остаются прикладные научные исследования, направленные на обеспечение сейсмической безопасности страны [2]. Катализатором внедрения инновационных решений и практически значимых результатов научных исследований, в первую очередь, являются разрабатываемые информационные технологии, платформы, и их архитектуры, а также интеллектуальные системы поддержки принятия решений и экспертные системы.

Современные научные исследования, связанные с оценкой сейсмической уязвимости и сейсмических рисков, берут свое начало из истории развития инженерной теории сейсмостойкости. Среди основоположников инженерной теории сейсмостойкости, важно отметить Ф. Омори (статическая теория), Н. Мононобе, К.С. Завриев (динамическая теория для одномерных систем), М.А. Био, Д. Хадсон, Дж. Алфорд, Г. Хаузнер, И.Л. Корчинский и др. (спектральная теория для многомерных систем), В.В. Болотин, И.И. Гольденблат, Ш.Г. Напетваридзе, Н.А. Николаенко, С.В. Поляков, А.Ф. Смирнов, Я.М. Айзенберг, А.В. Перельмуттер, В.И. Сливкер, Э.Е. Хачиян, Ю.И. Немчинов, Ю.П. Назаров и многие др. (вероятностная и нелинейная теории), Н. Ньюмарк, В. Ли, Х.Т. Шибата, М. Трифунак, Э.Е. Хачиян, А.П. Синицын, Н.А. Николаенко, Ю.П. Назаров и многие др. (волновая теория).

Большой вклад в развитие научных основ по вопросам сейсмостойкого строительства, оценки и снижения сейсмического риска внесли также отечественные ученые, как Т.Ж. Жунусов [4], Ю.Г. Шахнович [5], М.С. Абаканов [6], Т.Д. Тулеев [7], В.А. Лапин [8], С.Е. Ержанов [9], А.К. Курскеев [10], Т.Д. Абаканов [11], Ж.С. Нугужинов [12] и многие др.

Проблемы оценки сейсмической уязвимости объектов городской среды, а также оценки сейсмических рисков и прогнозирования землетрясений

методами машинного обучения и технологиями интеллектуального анализа данных, были рассмотрены в научных трудах зарубежных ученых, таких как I. Riedel, Ph. Guéguen, Z. Zhang, C.S. Chen, G. Brando, E. Irwansyah, P. Shi, L. Seydoux, S. Cesca, Z. Fan и др.

На сегодняшний день цифровые и информационные технологии меняют строительство. Активно внедряются технологии информационного моделирования (BIM-технологии) при строительстве зданий и сооружений, адаптируются и разрабатываются новые нормы, тем самым происходит цифровая трансформация строительной отрасли. Такие глобальные изменения, также поднимают актуальные вопросы, как современное строительство в условиях сейсмического риска, прогнозирование сейсмических рисков, обновление и цифровизация сейсмических карт местностей, а также разработка информационных технологий, моделей и методов, позволяющие «быстрой» оценке сейсмобезопасности, сейсмостойкости объектов в городской среде и анализе сейсмических данных.

Диссертационная работа посвящена решению актуальной задачи по разработке способа оценки сейсмостойкости городских объектов с учетом сейсмических данных на основе технологии анализа данных с применением методов машинного обучения, тем самым привнося свой научный вклад в развитие одного («Создание инновационной экосистемы») из пяти ключевых направлений государственной программы «Цифровой Казахстан» [13]. Совершенствование и интеграция технологических возможностей институциональных сред с новыми информационными технологиями и методами искусственного интеллекта, оказывают существенное влияние на спектр решаемых задач в области обеспечения безопасности страны, что в свою очередь определяет актуальность исследования в целом.

Кроме этого, диссертационное исследование выполнено в рамках госзадания №748715Ф.99.1.ББ97АА00002 ФГБОУ ВО Алтайского государственного университета по теме «Тюркский мир «Большого Алтая»: единство и многообразие в истории и современности на 2020-2021 годы при финансировании Министерства науки и высшего образования Российской Федерации.

Объектом исследования является процесс оценки сейсмостойкости городских объектов и анализа сейсмических данных.

Предметом исследования являются методы и алгоритмы машинного обучения для оценки сейсмостойкости городских объектов и анализа сейсмических данных.

Целью диссертационной работы является разработка способа оценки сейсмостойкости городских объектов с учетом сейсмических данных на основе технологии анализа данных и методов машинного обучения для повышения эффективности «быстрой» оценки данных.

Задачи диссертационной работы. Для достижения поставленной цели диссертационного исследования требуется решение следующих задач:

– проанализировать и исследовать существующие подходы и методы машинного обучения, применяемые в решении задачи оценки сейсмостойкости

городских объектов и анализа сейсмических данных;

- изучить существующие информационные технологии и системы оценки сейсмической уязвимости объектов и обработки сейсмических данных;
- разработать способ оценки сейсмостойкости городских объектов с учетом сейсмических данных на основе применения технологии анализа данных и методов машинного обучения;
- спроектировать архитектуру интеллектуальной информационно-аналитической системы оценки сейсмостойкости городских объектов с учетом сейсмических данных.
- разработать интеллектуальную информационно-аналитическую систему оценки сейсмостойкости городских объектов с учетом сейсмических данных.

Научная новизна диссертационного исследования заключается в том, что впервые предлагается способ оценки сейсмостойкости городских объектов, включающий алгоритмы интеллектуального анализа данных, позволяющие выделять группы городских объектов по их сейсмической уязвимости.

Основные научные положения, выносимые на защиту:

- разработанный способ оценки сейсмостойкости городских объектов, основанный на применении интеллектуального анализа данных, который позволяет осуществить «быструю» оценку сейсмостойкости городских типовых объектов, что значительно сокращает время и затраты для проведения оценки сейсмической уязвимости зданий;
- алгоритм пространственной кластеризации, обеспечивающий эффективный анализ и обработку пространственных сейсмических наборов данных для идентификации плотных участков землетрясений;
- архитектура интеллектуальной информационно-аналитической системы оценки сейсмостойкости городских объектов с учетом сейсмических данных.

Практическая ценность результатов исследования.

Разработанный способ оценки сейсмостойкости городских объектов с учетом сейсмических данных на основе технологии анализа данных и методов машинного обучения послужила основой для разработки архитектуры и информационной технологии интеллектуальной информационно-аналитической системы для реализации оценки сейсмической уязвимости объектов из городской среды с учетом сейсмических событий на примере представленных наборов данных.

Разработанное в диссертационной работе алгоритмическое и программное средство представляет возможность осуществления обработки и анализа наборов данных (городских объектов, сейсмических событий), которые встречаются в структуре систем оценки сейсмических рисков.

Результаты исследований используются в учебном процессе кафедры теоретической кибернетики и прикладной математики ФГБОУ ВО «Алтайский государственный университет» (Барнаул).

Методы исследования. В работе для решения поставленных задач

используется аппарат теории управления системами, методы статистического анализа, искусственного интеллекта, технологии проектирования информационных систем.

Реализация результатов работы. Основные научные и практические результаты диссертационной работы были представлены в виде программных модулей для оценки сейсмостойкости городских объектов с учетом сейсмических данных с получением 2-х свидетельств интеллектуальной собственности под названиями: 1. Web-приложение «Визуализация сейсмических данных (программа для ЭВМ)» №14910 от «5» февраля 2021 года (Приложение А). 2. Интеллектуальная информационно-аналитическая система «Оценка сейсмостойкости городских объектов на основе технологии анализа данных (программа для ЭВМ)» №15298 от «18» февраля 2021 года (Приложение Б).

Внедрение результатов диссертационного исследования подтверждается актом (Приложение В) о внедрении ФГБОУ ВО Алтайского государственного университета - АлтГУ (Барнаул). Применение результатов диссертационного исследования послужило основой для разработки и реализации СОП, включающей утвержденные учебные программы по направлению – 09.04.03 «Прикладная информатика» (магистратура), профиль «Цифровые технологии анализа данных для устойчивого развития регионов Северной и Центральной Азии». Разработка и реализации программы магистратуры выполнена совместно НАО «Восточно-Казахстанский университет имени Сарсена Аманжолова» и ФГБОУ ВО «Алтайский государственный университет» в рамках госзадания Алтайского государственного университета по теме «Тюркский мир «Большого Алтая»: единство и многообразие в истории и современности (№748715Ф.99.1.ББ97АА00002). По разработанной СОП выполнен набор обучающихся на 2021-2022 учебный год и запущен учебный процесс. Разработанное программное обеспечение будет использовано в рамках учебного курса «Методы и модели интеллектуального анализа».

Апробация работы. Основные результаты диссертационной работы доложены и обсуждены на следующих международных конференциях: «World science: Problems and Innovations» (Пенза, 2018) [14]; V Международная научно-техническая конференция студентов, магистрантов и молодых ученых «Творчество молодых – инновационному развитию Казахстана» (Усть-Каменогорск, 2019) [15]; 6th International Conference on Computer and Technology Applications - ICCTA 2020 (Анталья, 2020) [16]; VI Международная научно-техническая конференция студентов, магистрантов и молодых ученых «Творчество молодых – инновационному развитию Казахстана» (Усть-Каменогорск, 2020) [17]; «Computational and Information Technologies in Science, Engineering and Education (CITech-2020)» (Алматы, 2020) [18].

Публикации. По теме диссертационной работы опубликовано 9 работ, в том числе 1 статья в журнале, рецензируемый в базе данных Scopus (показатель процентиля по CiteScore равный 34%), 4 статьи в изданиях, рекомендованные Комитетом по контролю в сфере образования и науки МОН РК, 4 работы в сборниках международных конференций (1 из которых рецензируется в базе

данных Scopus).

Структура и объем диссертации. Диссертация состоит из введения, 4 разделов, заключения, списка использованных источников из 170 наименований, изложенных на 119 страницах компьютерного текста, включает 55 рисунков, 21 таблиц и 3 приложения.

Первый раздел диссертационной работы представляет собой обзор и анализ существующих подходов и методологий, основанных на технологии анализа данных и методов машинного обучения для оценки сейсмической уязвимости объектов городской среды и анализа сейсмических событий с целью прогнозирования и снижения рисков от них. Также, представлены результаты исследования существующих информационных технологий и систем для оценки сейсмической уязвимости объектов городской среды и анализа сейсмических данных. В результате исследования существующих информационных технологий были изучены архитектурные модели оценки сейсмической уязвимости объектов городской среды и анализа сейсмических данных, а также модели управления информационными системами. Обоснована актуальность необходимости проектирования интеллектуальных информационно-аналитических систем анализа данных (данные городских объектов, данные землетрясений) и принятия решений для оценки сейсмостойкости городских объектов и оценки сейсмических рисков.

Во втором разделе диссертационной работы исследован новый подход к решению задачи оценки сейсмостойкости городских объектов и разработан способ оценки сейсмостойкости городских объектов, который позволяет осуществить «быструю» оценку сейсмостойкости городских типовых объектов, что значительно сокращает время и затраты для проведения оценки сейсмической уязвимости зданий. В разработанном способе используется комплексный подход, сочетающий в себе кластерный анализ и классификацию. Привлечение алгоритмов машинного обучения позволяют повысить устойчивость результатов кластерного анализа в случае неопределенности в структуре набора данных. В данной части исследования показано, что целесообразность применения такого подхода была подтверждена экспериментальными результатами, свидетельствующими о том, что использование кластерного анализа и классификации существенно повышают качество решений. Кроме этого, показаны экспериментальные исследования с предложенными алгоритмами и оценки качества полученных результатов, а также сравнение полученных результатов с результатами экспертных оценок.

В третьем разделе диссертационной работы рассмотрено исследование сейсмического набора данных, проведенный в два подэтапа на основе технологии анализа данных и алгоритма пространственной кластеризации на основе плотности с присутствием шума DBSCAN. В первом подэтапе исследования каталога землетрясений проводился статистический анализ набора данных. Результаты статистического анализа данных показали количественные показатели по различным временным атрибутам, выявили классы магнитуд и их распределение на интерактивной карте. Во втором

подэтапе исследования набора данных землетрясений использовался алгоритм DBSCAN для обнаружения участков с повышенной пространственной плотностью сейсмических событий.

В четвертом разделе диссертационной работы представлена разработанная масштабируемая архитектура интеллектуальной информационно-аналитической системы (ИИАС) с веб-интерфейсом для оценки сейсмостойкости городских объектов с учетом сейсмических данных алгоритмами машинного обучения. Обоснована и приведена концепция разработки архитектуры ИИАС. Рассмотрены подмодули оценки сейсмостойкости городских объектов и анализа сейсмических данных. Описаны процедуры ввода и предварительной обработки данных и управления системой. Функционирование первого подмодуля ИИАС позволяет исследователю осуществлять разбиение определенного набора объектов на кластеры, согласно алгоритму кластеризации, и в том числе реализовывать отдельные алгоритмы классификации и получать классификационные модели оценки сейсмостойкости городских объектов. Во втором подмодуле ИИАС реализована обработка и анализ сейсмических данных, визуализация результатов на интерактивной карте и пространственная кластеризация.

В заключительном разделе диссертационной работы приводится перечень основных результатов и выводов диссертационного исследования, исходя из которых положения, выносимые на защиту, были подтверждены, дается оценка научной новизны и практической значимости исследования.

Приложения. Свидетельство о государственной регистрации программ для ЭВМ.

1 ОБЗОР СОВРЕМЕННОГО СОСТОЯНИЯ ИССЛЕДОВАНИЙ ПО ОЦЕНКЕ СЕЙСМИЧЕСКОГО РИСКА

1.1 Современное состояние исследований и актуальные проблемы обеспечения сейсмической безопасности

В истории и на памяти всего человечества запомнились множества случаев, порожденные различными типами катастроф, особенно природными, которые также влекут за собой катаклизмы техногенного характера. Одним из наиболее не предсказуемых и опаснейших природных катастроф на земле являются землетрясения [19, 20]. Ведь как всем известно, землетрясения приносят в различной степени разрушения и в первую очередь потери от человеческих жертв. Ущерб от них приравнивается к суммарному ущербу от всех других природных катастроф [21-23].

На сегодняшний день из-за своей непредсказуемости специалисты и ученые, занимающиеся исследованием сейсмических опасностей все еще не могут с точной вероятностью предсказать землетрясения, хотя в данном направлении по сей день ведутся тщательные и всесторонние исследования. Исследования вопросов сейсмической опасности и рисков остаются все еще актуальными даже в долгосрочной перспективе. Сам процесс землетрясения является очень сложным физическим явлением и в большинстве случаев, оказывается трудно поддающимся моделированию, и в свою очередь требует необходимости междисциплинарного подхода в его изучении.

Анализ статистики природных стихий и катастроф, в частности сильных землетрясений, показывает их частоту и рост. В связи с этим, в целях обеспечения безопасности для всех людей, а также устранения существенных технических, экономических и любых других значимых ущербов от землетрясений, в различных отраслях жизнедеятельности и в научно-исследовательских кругах ведутся активные научные работы. Необходимость развития и улучшения методов и способов обеспечения надежности городских объектов в сейсмически опасных районах и снижение затрат, наносимых в результате сейсмических воздействий на здания, является важнейшим вопросом и требует научно-практических решений [24]. Результаты исследовательских работ, требуют разработки максимально эффективных методов моделирования, прогноза и мониторинга сейсмических активностей, которые в свою очередь основаны на достижениях современных наук, таких как физика, геофизика и ряд других смежных областей, в числе которых, в том числе и есть ИТ. В развитых странах, таких как Япония, США, Греция и Италия, сотни лабораторий и научных центров работают над исследованием данных вопросов и проблем, тем самым разрабатывая различные междисциплинарные подходы и ИТ решения. В большинстве случаев к наиболее высокому риску от сейсмических воздействий, подвержены плотно заселенные пункты и территории с различной городской инфраструктурой. В связи с этим, изучение проблем по оценке сейсмостойкости объектов городской среды все больше и больше приобретает научно-прикладной характер.

Оценка сейсмической уязвимости или сейсмостойкости объектов

городской среды дает возможность в целом оценить положение уже имеющейся застройки городской территории, опередить дополнительные меры по усилению безопасности объектов, а также в будущем планировать строительные мероприятия на определенных участках этой местности. В целом, конечно же, в теории сейсмостойкости, существуют свои научно доказанные фундаментальные, традиционные методы и технологии, а также программно-расчетные комплексы по оценке зданий и сооружений от сейсмических воздействий, которые в свою очередь включают в себя множество различного рода специфических, сложных процессов. В данном направлении, как и в любой другой научной сфере применяются современные информационные технологии, которые заключаются в применении специализированных программ и комплексов, основанных на технологиях моделирования и методах численного расчета.

Из научных источников, связанные с теорией сейсмостойкости можно привести точное определение понятия сейсмостойкости и краткое описание возникновения самой теории сейсмостойкости. Сейсмостойкость (СП РК 1.01-101-2014 «Строительная терминология») – это способность зданий или сооружений противостоять сейсмическим воздействиям без потери эксплуатационных качеств [14, с. 111]. Началом зарождения теории сейсмостойкости сооружений как научной дисциплины следует считать период конца XIX и начала XX в., следующий за разрушительным японским землетрясением 1891 г. (Мино-Овари). При обследовании последствий этой крупнейшей катастрофы японскими учеными был произведен анализ смещения и опрокидывания памятников, массивов и т.п.; на основе этого были получены первые приближенные данные о максимальных сейсмических ускорениях грунта, позволившие определить задачу количественного исследования сейсмического эффекта. С целью анализа сейсмических сил в сооружениях Ф. Омори в 1900 г. провел специальные опыты с кирпичными столбиками [14, с. 112; 25]. Также, помимо основополагающей статической теории Ф. Омори существуют ряд других теорий сейсмостойкости зданий и сооружений, таких как: динамическая теория Н. Монообэ, теория К.С. Завриева, ударно-волновая концепция С.Б. Смирнова [14, с. 112].

Опыт разрушительных землетрясений начала XX в. (Сан-Франциско, 1906 г.; Мессина, 1908 г.; Канто, 1923 г.) выявил недостатки статической теории и несовершенство ее основных положений [14, с. 112; 26]. В СССР интерес к вопросам сейсмостойкости повысился после землетрясений в Ленинкане (1926) и Крыму (1927). Спустя время стало очевидно, что при землетрясении поведение здания преимущественно зависит от динамических свойств здания, и это свидетельствует о том, что решение задач сейсмостойкости осуществимо только в пределах динамической теории [14, с. 112; 26, с. 32]. В результате к началу 40-х годов была разработана новая форма динамического метода расчета на сейсмостойкость, именуемая обычно спектральным методом или методом расчета по спектральным кривым [14, с. 112; 26, с. 33].

Большинство известных ученых продолжают исследования в данной

области и указывают на сложность, неопределенность и неоднозначность в решении проблем, по оценке сейсмической уязвимости объектов. Во многих случаях это зависит от ряда факторов, среди которых можно отметить неточность карт сейсмического районирования территорий. По данному фактору, а именно вопроса карт сейсмического районирования, в целях обеспечения сейсмостойкого строительства важную роль имеет необходимость в проведении сейсмического микрорайонирования для уточнения сейсмического воздействия. Для более глубинного понимания понятия сейсмостойкого строительства зданий и сооружений, важно рассмотреть основные принципы конструирования сейсмостойких зданий и сооружений, методы обеспечения сейсмостойкости и способы сейсмоусиления строительных объектов, что в своем роде не является основной целью данной диссертационной работы. Так как, подобного рода исследования, считаются фундаментальными, основополагающими, наукоемкими и требуют специальных, профильных знаний. Здесь же, интерес представляется в предложении и рассмотрении нового подхода, который имеет научно-прикладной характер и который опирается на технологии анализа данных и алгоритмы машинного обучения, что позволяют получить новые, скрытые знания и закономерности с сырых наборов данных, связанные с объектами городской среды и землетрясениями.

В активно развивающемся цифровом мире и в период высокопроизводительных вычислительных технологий, будучи исследователями, специалистами или инженерами, невозможно не использовать возможности информационных технологий для решения или улучшения проблем по сейсмической безопасности. В настоящее время существует множество программных комплексов [14, с. 113], предназначенных для расчета конструкций на сейсмостойкость. Эти программы позволяют получить результаты, хорошо согласующиеся с работой конструкции в реальных условиях. В работе с такими программами необходимы специальные знания, навыки и высокопроизводительные вычислительные ресурсы, без которых трудно реализовать оценку сейсмостойкости зданий или сейсмического риска. В целом, математическая модель сейсмического риска слишком абстрактна и недостаточно информативна, поскольку в ней не представлены типы зданий, виды негативных последствий для зданий и для людей, находящихся в них [27]. В связи с этим, в диссертационной работе, предлагается способ оценки сейсмостойкости городских объектов, основанный на применении интеллектуального анализа данных, который позволяет осуществить «быструю» оценку сейсмостойкости городских типовых объектов. Разработанный способ учитывает характеристики зданий и позволяет провести кластерный анализ данных, где для уточнения структуры полученных кластеров строится модель дерева решений, а также выявляются наиболее важные параметры для зданий, влияющие на их сейсмостойкость. Также, технологии анализа данных и алгоритмы машинного обучения в совокупности, дают возможность эффективной реализации анализа и обработки пространственных сейсмических данных для идентификации плотных участков

(кластеров) и места возможных очагов землетрясений.

Анализ и оценка сейсмостойкости уже имеющихся и вновь возводимых зданий, как уже было отмечено ранее, процесс не простой и требует проведения ряда мероприятий, а также включает в себя большой объем ресурсов и затрат по времени. Обследование и оценка технического состояния зданий и сооружений регламентируются сводом правил (для каждой страны существуют свои нормативные документы), который разработан в соответствии с требованиями нормативных правовых актов в строительстве. На сегодняшний день в Казахстане, ведутся активные обсуждения по разработке новых строительных нормативно-правовых актов на основе Еврокодов. Данный переход ведется в три этапа, где первый этап был начат еще с 2015 года и завершающий этап, т.е., осуществление третьего этапа приходилось на 2020 год. Переход Республики Казахстан на Еврокоды обуславливается необходимостью интеграции строительной отрасли с европейской и мировой системами путем гармонизации с реформой устаревшей нормативно-технической базы.

В целом, в нашей республике особое значение уделяется исследованиям в области фундаментальных и прикладных исследований, среди которых важную позицию составляют вопросы по обеспечению сейсмической безопасности в РК. Среди важных научно-исследовательских и опытных работ по актуальным проблемам и вопросам сейсмостойкого строительства в РК, можно отметить инженеров-строителей и ученых, таких как Шахнович Ю.Г., Тулеев Т.Д., Есенберлина Д.И., Лапин В.А., Абаканов М.С., Ержанов С.Е., Ицков И.Е., Тастанбеков А.Т. и т.д., которые привнесли наиболее значимый вклад и весомые результаты в развитие данного направления.



Рисунок 1 – Казахстанские научные институты и организации, охватывающие проблемы по обеспечению сейсмической безопасности

На сегодняшний день на разных уровнях деятельности страны были созданы и существуют различные научные институты и центры, занимающиеся актуальными проблемами сейсмической безопасности всей страны, а также активными внедрениями информационных разработок в свою деятельность. Выше в соответствии с рисунком 1 представлены основные институты и организации, которые занимаются исследованиями и решениями вопросов, охватывающие проблемы по обеспечению сейсмической безопасности в нашей стране.

В настоящее время данные институты и организации, занимаются фундаментальными и научно-прикладными исследованиями по проблемам обеспечения сейсмической безопасности и вопросами сейсмостойкого строительства в Республике Казахстан. Фундаментальные исследования, проводимые в рамках научно-исследовательских институтов, являются необходимыми и основополагающими исследованиями. Кроме научных институтов и организаций, приведенные в соответствии с рисунком 1, также существуют и другие объекты, которые осуществляют непрерывные инструментальные наблюдения за сейсмическим режимом и их регистрацией, тем самым накапливая данные о произошедших землетрясениях и передавая их уже в специализированные научные институты для глубоких и комплексных исследований.

Во многих случаях современные исследователи из данных областей науки при проведении расчетного эксперимента и моделирования процессов сталкиваются со сложностью выбора наилучшей структуры и параметров модели. Такие трудности появляются из-за некоторых факторов, таких как наличие широкого спектра решений, направленные на решение одних и тех же задач, которые отличаются друг от друга точностью расчета, своей производительностью, функциональными ограничениями в применении и т.д. Также современный уровень развития вычислительной техники предоставляет исследователю доступ к ресурсам высокой производительности, однако этот доступ чаще всего требует от исследователя значительной технической подготовки. В дополнение ко всему, процесс моделирования информационных систем, требует интеграции в рамках единых решений для моделирования множества отдельных элементов, зачастую имеющих различную природу или даже принадлежащих к различным предметным областям. В решении таких задач, возникает вопрос автоматизации и интеллектуальной поддержки операций моделирования, традиционно возлагаемых на специалиста исследователя, но в виду указанных факторов, приобретающих крайне высокую сложность во многих современных задачах.

Одна из важных предпосылок, определяющая актуальность и общую концепцию диссертационного исследования, является то, что в условиях развития технологий Data mining, появляется доступность к большим объемам данных для анализа, что становится неопределимым источником информации о моделях и их применимости. Современный уровень работы со знаниями, открытость источников данных и знаний позволяет использовать обширную базу источников и технологических решений для логического вывода при

настройке моделей [28]. Существующие технологии интеллектуального анализа данных и методы машинного обучения, представляют возможность быстрыми способами выявлять необходимые знания. В конечном результате, такой подход в исследовании, может характеризоваться междисциплинарностью и опираться на современные разработки, которые позволяют существенно упростить, а иногда и автоматизировать процессы работы с моделями в различных предметных областях науки и техники [29].

В связи с этим, наличие основных наборов данных по характеристикам объектов городской среды, а также наборов данных о сейсмических событиях, представляют отличную возможность применения технологий интеллектуального анализа данных и машинного обучения. Учитывая современные подходы и методы [17, с. 108] анализа больших данных, исследования в области по оценке сейсмической уязвимости, сейсмостойкости объектов городской среды и обработки данных о сейсмических событиях, все еще набирают свою популярность и требуют дополнительных исследований.

В целом, состояние исследований подобного рода, опирающиеся на инновационные возможности технологии анализа данных и методов машинного обучения, в нашей стране все еще находится на начальном пути своего развития.

1.2 Анализ современных методов и моделей в оценке сейсмических рисков на основе технологии анализа данных и методов машинного обучения

Быстрое развитие науки информатики обеспечивает техническую поддержку обработки больших данных, особенно методами машинного обучения [30].

Машинное обучение обеспечивает классификацию и прогнозирование на основе известных данных и может обеспечить высокую точность и надежность, что повышает вероятность принятия правильного решения [31].

Изначально машинное обучение ((ML), McCulloch and Pitts, 1943), было предложено в 1940-х годах в теории вычислительных моделей нейронных сетей [30, p. 8]. Машинное обучение - это компьютерный алгоритм, который может автоматически улучшаться на основе опыта применения (Mitchell, 1997), и также может преобразовывать цепочку данных к решению [30, p. 8]. Машинное обучение может использовать математические и статистические методы для определения внутренней регуляции на основе различных данных [30, p. 8].

На сегодняшний день он эволюционировал в сторону более продвинутых типов обучения, которые ближе к человеческому мозгу, таких как глубокое обучение (Hinton and Salakhutdinov, 2006), трансферное обучение (Pan and Yang, 2010; Mnih et al., 2015). В то же время машинное обучение широко используется во многих областях, особенно в сейсмологии, например, для идентификации и прогнозирования землетрясений (Asim et al., 2016; DeVries et al., 2018; Lubbers et al., 2018; Rajguru et al., 2018; Corbi et al., 2019) и классификации изображений сейсмического дистанционного зондирования (Afonso et al., 2016; Bialas et al., 2016; Frank et al., 2017) и т.д.

Эффективность применения технологии анализа данных и методов машинного обучения во многих сферах человеческой жизни [29, р. 129] все больше и больше доказывает их применимость, на основе них развиваются различного рода аналитические платформы, тем самым расширяются границы научно-исследовательской деятельности в целом.

К примеру, в работе [32], процесс оценки сейсмической уязвимости зданий в городских условиях рассматривается как дорогостоящим и трудоемким. При моделировании сейсмической уязвимости объектов или проведении ее оценки на сейсмостойкость, используются различные вычислительные модели, а также 3D модели [15, с. 100] и технологии [33], которые считаются затратными по времени и по экономическим показателям, а также требуют очень сложных и многомерных характеристик структурной конфигурации зданий и свойств материалов, из которых они построены. В связи с этим за последние десятилетия, развитие технологий больших данных, в свою очередь стало набирать все большую актуальность, тем самым показало применение методов «быстрой» оценки сейсмической уязвимости. Данные методы опираются на сравнительно небольшие, но надежные наборы основных характеристик зданий, доступ к которым возможен даже на региональном уровне [34]. Соответственно, возросло число исследований, в которых методы интеллектуального анализа данных использовались для поиска зависимости между уязвимостью зданий и диапазоном их характеристик, таких как длина здания, количество этажей, общая площадь земли и т.д. [35-37]. Результаты, которые были получены в данных исследовательских работах практически подтвердили способность методов интеллектуального анализа данных эффективно оценивать сейсмическую уязвимость зданий в широком географическом масштабе.

В работе [32, р. 1114], в качестве ограниченного набора характеристик зданий использовались следующие данные: год постройки, ранжированный по интервалам, количество этажей, ранжированных по категории, форма крыши (плоская или с уклоном), строительный материал, качественное описание неровностей плана и высоты, а также положение здания (угловое, между зданиями, отдельное стоящее и т.д.). В дополнение к этой информации эксперты связывали тип здания в соответствии с типологией EMS98 с наиболее вероятным классом уязвимости. Классов уязвимости зданий по типологии EMS98 – 6, от А (наиболее уязвимый) до F (наименее уязвимый). Таким образом, была составлена база данных об уязвимостях зданий Гренобля, в которой 3860 зданий были описаны набором ранее приведенных характеристик и классами уязвимости. Затем полученная база данных делилась на две части: обучающую и тестовую. Отдельно исследовался вопрос о влиянии размера обучающей выборки и влиянии количества признаков на точность классификации. Классификация зданий производилась двумя способами: с помощью метода опорных векторов (Support vector machine) и с помощью метода поиска ассоциаций. Полученные модели оценивались на исторических данных, т.е. на данных повреждений, полученных зданиями в прошлом, после имевших место землетрясений. Авторы доказали, что если бы эти исторические

землетрясения произошли сегодня, то из-за эволюции уязвимости в городах число серьезно поврежденных зданий могло бы удвоиться или даже утроиться.

Предложенная в [32, р. 1116] методика также нашла применение в работе [38]. Авторы этой работы применили методику для оценки сейсмической устойчивости зданий в г. Урумчи Китайской Народной Республики. На основе этих двух методов оценки уязвимости и базы данных города Урумчи были построены карты уязвимости города с помощью ГИС-технологий.

В работе [35, р. 4103] исследуется проблема эффективной оценки сейсмостойкости школьных зданий на Тайване. Тайваньский национальный центр исследований в области землетрясений использует ускорение грунта (AR) в качестве показателя для оценки соответствия школьной структуры требованиям сейсмостойкости. Однако вычислительные процессы, связанные с расчетом ускорения грунта, занимают много времени и требуют участия многих экспертов. Для решения этой проблемы предлагается система логического вывода, в которой интегрированы два метода интеллектуального анализа, а именно метод опорных векторов (SVM) и быстрый неупорядоченный генетический алгоритм (fmGA). Основываясь на результатах обучения, разработанная система может прогнозировать ускорение грунта школьного здания за значительно более короткое время, что значительно повышает эффективность оценки. SVM используется для получения отображения входных переменных в выходные, а fmGA используется для поиска оптимального параметра, требуемого SVM, который может повысить точность прогнозирования SVM. Входные переменные, используемые системой, включают в себя 13 факторов, влияющих на сейсмостойкость школьного здания, а выходная переменная – прогнозируемое значение ускорения грунта. Полученные результаты также подтвердили перспективность применения методов интеллектуального анализа для прогнозирования сейсмической уязвимости зданий.

В следующем исследовании [36, р. 1108] разработана система генетического программирования для построения модели зависимости между сейсмостойкостью школьных зданий и их базовыми параметрами проектирования. Модель основана на информации из базы данных школьных зданий, а также информации, касающейся сейсмостойкости школьных зданий, проанализированной с использованием полных нелинейных методов.

В [37, р. 10258] рассматриваются 9 предиктивных (предсказательных) моделей для определения взаимосвязи между основными проектными параметрами школьных зданий и их сейсмической уязвимостью. Три модели основаны на обобщенной линейной регрессии (GLM), простой регрессии и нейронной сети (ANN). Смешанная модель K-средних и двухступенчатая кластеризация породили еще три модели каждая. Все модели используют только базовые геометрические данные о школьных зданиях для быстрой оценки их сейсмической уязвимости. Лучшие результаты показала модель на основе GLM и K-means.

В работе [39] оценка сейсмической уязвимости проводится в целях защиты объектов народной архитектуры, расположенных в районах,

подверженных землетрясениям. Предлагается новый метод оценки сейсмической уязвимости, адаптированный под особенности местной архитектуры, который учитывает традиционную нехватку ресурсов, выделяемых на изучение и сохранение народного наследия. По мнению авторов работы, такой метод должен быть легким в использовании и опираться на качественные данные, которые можно быстро получить с помощью простого визуального инспектирования. Более детальные и сложные подходы, которые требуют глубокого изучения конструкций, включая, например, исторические исследования, расширенный структурный анализ, как правило, не применимы для отдельных монументальных зданий. Во-вторых, учитывая большую неоднородность народной архитектуры с точки зрения геометрии, материалов или методов строительства, новый метод также должен позволять индивидуальную оценку зданий. Для разработки метода, который авторы назвали SAVVAS, использовались регрессионные модели. По словам авторов, использование аналитического процесса вместо эмпирического для разработки ускоренного метода оценки сейсмической уязвимости местных сооружений является шагом вперед во вкладе в научные знания.

Наряду с работой [39, р. 2927] по сохранению исторического наследия следует отметить работу [40]. В ней представлена предиктивная модель для оценки сейсмической уязвимости малых исторических центров. Модель, разработанная в рамках других аналогичных методов, предложенных в прошлом, также нуждается в ограниченном количестве параметров и основана на информации, собранной после землетрясения в Аквиле в 2009 году.

Также в продолжении примера применения методов интеллектуального анализа для оценки сейсмической уязвимости объектов городской среды, можно привести работу [41], которая посвящена анализу данных, связанных с повреждениями, наблюдаемыми на зданиях после землетрясения в Аквиле в 2009 году. Для выявления возможных видов множественных повреждений в данной работе используется подход, основанный на классификации повреждений, рассматриваемых как статистическая переменная. Также целью работы является выявление кластеров переменных повреждений каменных зданий для различных структурных элементов (вертикальная структура, горизонтальные структуры, лестница, крыша и перегородка). Тем самым, обнаруживаются различные повреждения, которые одинаково присутствуют на разных зданиях. Чтобы выбрать повреждения для конкретного кластера исследуются различные повреждения с помощью гетерогенности Джини. Кроме этого, была проанализирована устойчивость различных кластеров, и анализ устойчивости был выполнен *bootstrap*. Также, чтобы проанализировать чувствительность результатов, в данной работе был рассмотрен другой алгоритм кластеризации по различным переменным. А именно алгоритм K-средних по числу кластеров, выявленных в предыдущем анализе, для сравнения полученных результатов, которые оказались аналогами. В итоге полученных исследований авторы утверждают, что полученные их результаты можно считать согласованными с инженерными ожиданиями.

Не менее интересна работа [42], в которой ярко выражено применение

методов интеллектуального анализа для кластеризации пространственных данных. Исследования проводились с целью разработки кластера пространственных данных и анализа характеристик каждого кластера данных для разработки пространственной зональности опасности повреждения зданий, вызванного землетрясением в городе Банда-Ачех (Индонезия). В качестве данных в данном исследовании используются данные пикового ускорения земли (PGA), данные о литологии и топографической зоне. Исследование проводилось в четыре этапа: 1) нормализация данных методом min-max; 2) определение оптимального числа кластеров (с индексами связности, Данна и силуэта); 3) кластеризация данных с использованием алгоритма k-средних; 4) интерполяция данных с использованием алгоритма кригинга для построения пространственных моделей построения зоны опасности повреждения. Согласно результатам анализа, были получены 2 кластера. В 1-ом кластере 170 данных, которые были в основном сосредоточены на расстоянии менее 0,1-0,5 от центра кластера, а во 2-ом 55 других данных, которые сосредоточены на расстоянии 0,3-0,5 от центра кластера. В качестве инструмента кластеризации данных использовалась система R. По итогам исследования Банда-Ачех и прилегающие районы были пространственно разделены на два класса потенциальных повреждений зданий, вызванных землетрясением. В завершении этой работы авторы представили пространственную картину опасности повреждения зданий в городе Банда-Ачех.

В следующей исследовательской работе [43], также представлен анализ сейсмической уязвимости в городском масштабе (г. Константин, Алжир), основанный на методе интеллектуального анализа данных ARL (association rule learning), т.е. обучение правилам ассоциации. Использование метода ARL заключалось в установлении связей между атрибутами здания (количеством этажей или возрастом здания) и классами уязвимости Европейской Макросейсмической шкалы EMS98. Как отмечается в работе, применение данного метода помогает извлечь «скрытые» связи между элементарными признаками зданий и сейсмической уязвимостью. На следующих этапах исследования был создан прокси-сервер для определения уязвимости для всех комбинаций атрибутов здания. Прокси был получен на этапе обучения, и затем применен к остальной части банка данных зданий города Константина. Точность метода ARL оценивалась по сравнению с методом Risk-UE, который в свою очередь требует большого количества информации о характеристиках здания. Авторы успешно смогли доказать, что применение методологии ARL обеспечивает ту же оценку ущерба, что и Risk-UE, даже если используются только два атрибута. Вместе с тем, метод ARL позволил дать общую оценку сейсмического риска в городских районах. По утверждениям авторов данный подход позволяет избежать дорогостоящего процесса составления кадастра характеристик зданий в полевых условиях, что часто препятствует оценке сейсмических инициатив в слабых и умеренных сейсмоопасных регионах. К дополнению ко всему прочему, применение ARL в данном исследовании имеет отношение к сейсмической уязвимости в работах [32, p. 1120; 44].

В работах [45, 46] предложены способы прогнозирования уровня

повреждения зданий, одним из которых является модель байесовской сети. Оценка уровня разрушения основана на байесовской сети и позволяет точно установить причинно-следственную связь между переменными, а также отражает взаимосвязь между состояниями. В данных работах представлен новый метод оценки уровня повреждения здания. Набор данных основан на единице здания. Данные получены из Падангского регионального агентства по борьбе со стихийными бедствиями и Индонезийского метеорологического, климатологического и геофизического агентства. В работе [46, р. 722] использованы 61344 данных о зданиях. Данные исследования состояли из трех независимых переменных и четырех зависимых переменных (конструкция, риск оползня, PGA и ущерб).

Наряду со всеми работами, посвященные оценке сейсмической уязвимости, также можно привести в пример работу [47]. В данной работе представлен подход глубокого обучения, который основан на рекуррентной нейронной сети с длительной кратковременной памятью (LSTM), для моделирования отклика и прогнозирования структурной сейсмической реакции. В отличие от классических численных методов, предложенная модель глубокого обучения дает возможность предсказывать как упругую, так и неупругую реакцию строительных конструкций, на основе обученных данных. Данный подход прогнозирования нелинейных структурных откликов является актуальным в области анализа сейсмической хрупкости зданий для оценки надежности.

Ущерб, наносимый землетрясениями всегда огромен. Данное природное явление имеет свою периодичность возникновения. В связи с этим возникает потребность в решениях задач для минимизации потерь от землетрясения путем разработки различных методов и технологий, а также инструментов прогнозирования и анализа сейсмических данных. Данные решения позволяют грамотно проектировать и переконструировать сейсмоустойчивые сооружения и здания, а также использовать их в географических базах знаний.

Многие исследования, связанные с прогнозом и оценкой сейсмической уязвимости жилых массивов, с сейсмическим районированием, с мониторингом и обновлением данных о сейсмических областях страны или граничащих с ней территориях, в основном опираются на характеристики, которые определяют силу землетрясений. Такими характеристиками измерения силы землетрясений являются глубина, энергетический класс и магнитуда. Для оценки силы и воздействий землетрясений используются шкала магнитуд (например, шкала Рихтера) и различные шкалы интенсивности. Также при исследовании такого рода задач используются исторические данные о землетрясениях с географическими координатами и зафиксированным временем.

В [48] на основе метода data mining, линейной регрессии исследуется связь между смертностью и следующими показателями: 1) магнитудой землетрясения; 2) энергией, выделенной землетрясением; 3) повреждением здания от землетрясения; 4) временем возникновения землетрясения. В качестве набора данных для анализа использованы исторические данные о землетрясениях, произошедшие в Турции в период с 1900 года по 2012 год.

Таким образом, были исследованы закономерности возникновения землетрясений на основе результатов корреляции атрибутов землетрясений и показателей смертности, а также показателей повреждения зданий. В анализе данных использовались величины магнитуд землетрясений ≥ 5 . В результате пространственно-временного анализа землетрясений исследователи получили кластера, которые улучшают понимание пространственного распределения сильных землетрясений и наносимых ущербов от них.

В другой работе [49] также представлена методология анализа данных о землетрясениях методами интеллектуального анализа данных. Анализ данных землетрясений основывался на географических координатах события, таких как широта и долгота, а также на данных как глубина эпицентра. Также, как и в [48, p. 1950], в данном исследовании были учтены землетрясения магнитудой выше 5. Исследователи использовали показатель магнитуда для разделения диапазона значений на 3 класса землетрясений: средний, серьезный и очень серьезный. Для анализа данных о землетрясениях авторы применили expectation maximization clustering method. Кластерный анализ показал результаты, что эпицентры землетрясения тесно сгруппированы по широте и долготе, но имеют большой вертикальный диапазон. В продолжении исследования авторы для описания классов магнитуд по расположению эпицентров землетрясения применили деревья решений. Данное исследование было проведено на сравнительно малых количествах данных. По мнению авторов, предлагаемая методология считается надежной и определяет уровни рисков для каждого класса землетрясений с магнитудой 5, 6 или 7.

Наряду с вышеуказанными работами, где также рассмотрены методы кластерного подхода для анализа сейсмических наборов данных о землетрясениях, можно отметить и [50]. В данной работе представлен кластерный анализ на основе плотности, который позволяет находить сейсмическую зону с высококоррелированными пространственными данными в кластере. Аналогичное применение алгоритма кластеризации на основе плотности к данным о точках землетрясения можно увидеть и в [51]. В исследовании [52], предлагается новая система, которая анализирует причинно-следственные связи при распределении данных о землетрясениях в Индонезии. Данная система способствует построению карт риска в регионах в Индонезии. В исследовании использована автоматическая кластеризация пространственно-временных данных, а также алгоритм поиска ассоциативных правил для выявления факта связи или связи между распространением землетрясения. В следующей работе [53], рассмотрен один из методов data mining, нечеткая C-Means кластеризация (Fuzzy C-Means Clustering), которая была применена к анализу сейсмических данных. По утверждениям авторов данный метод показал хороший результат в анализе данных. Как и в большинстве исследований в [54], также используется пространственный анализ распределения магнитуды для определения оптимального числа кластеров на основе сейсмических данных.

В рассмотренной работе [55], предложено исследование на основе использования нечеткой логики для получения зависимостей затухания,

включающая такие параметры, как эпицентральное расстояние, магнитуда и интенсивность землетрясения. В качестве базы данных использованы данные о землетрясениях на Тайване и в США, состоящие из 1082 записей. В результате исследования, были определены два нечетких набора, среди которых один для магнитуд землетрясений, которые классифицируются как сильные, умеренные, легкие. Также определен третий набор данных для эпицентральных расстояний, которые в свою очередь определены как ближние, промежуточные и дальние. Авторы работы получили результаты и доказали, что нечеткий подход в целом дает более высокий коэффициент эффективности в применении, и что эпицентральное расстояние и тип почвы имеют большое влияние на характеристики затухания. Эффективное применение подхода нечеткой логики, также рассмотрены в [56], где система нечеткого логического вывода основана на правилах, включающая правило Такаги-Сугено для оценки вероятности возникновения землетрясений и сейсмического зонирования Загроса. В работе используется обновленный каталог данных Ирана о землетрясениях (интервалы 10 и 50 лет) для прогнозирования максимальной магнитуды будущих землетрясений. Применение нечеткого предиктора используется для того, чтобы правильно предсказывать магнитуду землетрясения и время его возникновения. По утверждениям авторов из-за неполноты данных, нечеткая модель хорошо предсказала и предоставила точные данные в процессе проверки. В оценке точности используемого нечеткого предиктора используется диаграмма Молчана. Полученные результаты из диаграммы Молчана, предоставили возможность утверждать, что модель прогнозирования опасности землетрясений является надежной моделью для прогнозирования максимальных магнитуд будущих землетрясений в поясе Загроса. Нечеткая карта сейсмического прогнозирования Загроса на основе правил показала пять различных зон. Кроме вышеназванных работ об эффективном применении нечеткой логики для анализа и прогнозировании землетрясений, а также для сейсмического зонирования определенных территорий, есть и [57], где предлагается исследование, и экспертная система на основе нечетких правил, используемые в качестве системы классификации сейсмических сигналов. Данная экспертная система была применена к локальной сейсмической сети Агадира. Также данную систему можно настроить и для других сейсмических сетей. Полученные результаты классификации показали, что нечеткая экспертная система обеспечивает высокую производительность при низкой сложности, и ее можно использовать как онлайн-классификатор. В другом исследовании [58], также представлена экспертная система, которая основана на разработанной методологии BRBES (Belief Rule Based Expert System) для прогнозирования землетрясения в условиях неопределенности. Разработанная авторами экспертная система рассматривает различные категории неопределенностей, связанные с переменными поведения животных и экологическими изменениями.

Применение методов интеллектуального анализа данных для прогнозирования природной опасности как землетрясения, в который раз уже доказывают свое эффективное применение в различных исследованиях. Так,

наряду с примерами исследований и применения методов интеллектуального анализа данных, можно привести работу [59], где цель исследования состоит в прогнозировании следующего землетрясения на основе данных предыдущего землетрясения. Такое исследование достигается путем применения интеллектуального анализа ассоциативных правил к данным о землетрясениях с 1979 по 2012 годы. Процесс прогнозирования землетрясения выполняется с помощью экспертной системы, которая использует только текущие атрибуты землетрясения для прогнозирования следующего землетрясения. Правила, созданные для прогнозирования землетрясений проверены на реальных данных. Результаты данного исследования показывают, что предлагаемая авторами экспертная система, основана на правилах и способна обнаруживать 100% землетрясений, которые произошли в течение 15 часов в пределах определенного диапазона, глубины и местоположения.

В исследовательской работе [60] были применены искусственные нейронные сети, машины опорных векторов и случайные леса для выполнения временных исследований каталога землетрясений в регионе Кипра и рассчитаны 60 сейсмических индикаторов для краткосрочного прогнозирования землетрясений. В [61] применялись различные алгоритмы машинного обучения, а именно: метод опорных векторов (SVM), алгоритмы К-ближайшего соседа (KNN), случайного леса (RF) и наивный байесовский классификатор (NB) на языке программирования R для прогнозирования землетрясений с использованием набора сейсмических данных Индии. В [62] изучались тепловые аномалии, которые произошли до землетрясения в Имфале, Индия, в 2016 году, а также исследовались многочисленные сейсмические факты с помощью спутниковых данных с использованием алгоритмов машинного обучения для землетрясения. В следующей работе [63] собраны записи афтершоков землетрясения в Керманшахе (Иран) и применены различные алгоритмы машинного обучения (ML), включая наивный байесовский классификатор, k-ближайших соседей, машину опорных векторов и случайные леса для прогнозирования будущих землетрясений путем наблюдения за образцами афтершоков. В [64] использовались нейронные сети для обнаружения сигналов землетрясений. В [65] приводится подробное описание методов мониторинга, используемых для прогнозирования землетрясений. В источниках [66, 67] представлен всесторонний обзор методов машинного обучения, используемых для прогнозирования землетрясений. В [68] были сделаны прогнозы сейсмических опасностей с использованием двух различных методов, основанных на машинном обучении, для пространственного и пространственно-временного прогнозирования сильных землетрясений. В [69] определено значение триггеров оползней на мелководье при прогнозировании землетрясений с использованием методов машинного обучения. В [70] усовершенствован традиционный метод корреляции формы волны и представлен новый метод обнаружения сейсмических сигналов для контроля ложных тревог с использованием машинного обучения. В [71] были определены, классифицированы и рассмотрены известные модели машинного/глубокого обучения, используемые в энергетических системах. В

[72] обсуждается несколько моделей искусственного интеллекта, использовавшихся для прогнозирования гидрологических моделей за последнее десятилетие. В [73] освещены возможности и проблемы, связанные с большими данными для принятия обоснованных решений.

Кроме вышеприведенных подходов, также в исследованиях применяются и другие методы машинного обучения и разрабатываются различные экспертные системы. К примеру, в работе [74] разработана нечеткая экспертная система Neuro, которая использовалась во многих исследованиях для анализа различного рода данных с целью прогноза землетрясений. В работе [75], исследуется проблема прогнозирования будущих землетрясений с магнитудой 5,5 и выше, где применяется подход с использованием адаптивной системы нейро-нечеткого вывода (ANFIS). В качестве данных, были использованы данные о землетрясениях с 1950 по 2013 гг., которые были зарегистрированы в Иране. При разработке моделей адаптивной системы нейро-нечеткого вывода (ANFIS), были использованы три алгоритма: разделение сетки (GP), кластеризации (SC) и нечеткие C-means (FCM). Экспериментальные результаты работы системы, по утверждениям авторов показали, что система системы нейро-нечеткого вывода ANFIS-FCM с высокой точностью предсказывает магнитуду землетрясения. Работа [76], представляет разработку персональной портативной системы прогнозирования погоды и мониторинга землетрясений с модулем GSM, данная система отправляет SMS-уведомления пользователю и на компьютерный интерфейс. Цель разработки такой системы заключается в обеспечении мониторинга интенсивности землетрясений, а также в ежедневном мониторинге погоды каждые 6 часов. В целом большинство исследований применили подход искусственной нейронной сети с методами машинного обучения на основе системы нечеткого вывода [76, p. 204; 77-79].

В целом применение методов машинного обучения во многих исследованиях, основываются на различных подходах и алгоритмах для определения сейсмичности территорий и их зонирования, изучения влияния геологических и географических структур земли, исследование форшоков и афтершоков, а также сейсмических волн, распространяющиеся в толще горных породах, вычисление гипоцентров землетрясений и их времени возникновения и т.д.

В продолжении обзора научных работ и применяемых подходов по прогнозированию землетрясений, также можно выделить ряд следующих работ. В работах [80, 81], применяется неконтролируемое машинное обучение. В первом случае применяется неконтролируемое машинное обучение к трехлетним непрерывным сейсмическим данным, чтобы понять эволюцию свойств поля сейсмических волн в период землетрясения. Во втором случае для обнаружения и кластеризации сейсмических сигналов в непрерывных сейсмических записях. Данный подход объединяет сеть глубокого рассеяния и модель смеси Гаусса для кластеризации сегментов сейсмического сигнала и обнаружения новых структур. В работах [82, 83] подходы прогноза землетрясений основываются на применении методов глубокого обучения как нейронные сети. Применение модели нейронной сети, также представлено в

работе [84], где рассмотрен один из подходов к оценке поля напряжений земной коры, который основан на очаговых механизмах землетрясений. Эффективность применения алгоритма неконтролируемого машинного обучения нечеткой кластеризацией представлена в работе [85], где данный вид кластерного анализа помогает распознать формы сейсмических волн в микросейсмических данных или в данных о землетрясениях. В работе [86], предложена система прогнозирования землетрясений, в которой объединены сейсмические индикаторы с ансамблевым методом AdaBoost на основе генетического программирования (GP). Система прогнозирования показывает землетрясения с магнитудой 5,0 и выше за пятнадцать дней до землетрясения.

Как видно, из выше приведенного списка [80, р. 3; 81, р. 4; 82, р. 1383; 83, р. 10; 84, р. 1659; 85, р. 94; 86, р. 3] исследовательских работ, существуют различные методологии и технологии по анализу сейсмических каталогов разного объема данных и атрибутов землетрясений. В приведенных работах [80, р. 3; 81, р. 4; 82, р. 1383; 83, р. 10; 84, р. 1659; 85, р. 94; 86, р. 3], в основном применены методики неконтролируемого машинного обучения, нечеткой кластеризации, а также методы глубоко обучения и классификационные модели данных. В следующих же исследованиях по анализу сейсмических данных, применяются другие методы кластерного анализа, как K-Means и DBSCAN. В работе [87], представлено описание приложения под названием seiscLOUD, предназначенное для кластеризации и визуализации локальных сейсмических каталогов. В структуре seiscLOUD процедура кластерного анализа сейсмических данных проводится с помощью алгоритма кластеризации DBSCAN, основываясь на таких характеристиках как эпицентральное положение и время происхождения. В работе [88], были рассмотрены три метода кластерного анализа K-Means, GMM (Модель Гауссовой смеси) и SOM (Самоорганизующиеся карты) для разделения сейсмичности рудника Юншаба (Yongshaba, Китай). В исследуемом наборе использованы 1516 сейсмических событий ($M > -1.5$), полученные на руднике. В проведенном исследовании при сравнительном анализе данных методов утверждается, что метод кластерного анализа K-Means (при $k=11$, Евклидово расстояние) показал наиболее точные результаты кластеризации и обеспечивает эффективный способ разделения шахтной сейсмичности, интерпретации геологической структуры и оценки сейсмической опасности. В работе [89], для кластерного анализа сейсмических данных были использованы 23 413 записей, каждая из которых включает 21 атрибут. Однако из всех атрибутов для анализа данных были отобраны такие атрибуты, как дата, время, широта, долгота, глубина и магнитуда. В кластерном анализе применены алгоритмы кластеризации K-средних и DBSCAN, которые сравниваются между собой. Применение данных алгоритмов обусловлены тем, что исследуются распространения сейсмических данных, полученные результаты аппроксимируются и визуализируют сейсмические пояса. На основе полученных результатов землетрясения группируются по глубине очага и сейсмической магнитуде. Сравнение данных алгоритмов при кластеризации сейсмических данных показали, что для изучения подгонки сейсмических поясов алгоритм DBSCAN, основанный на плотности, намного превосходит

алгоритм К-средних.

Таким образом, существуют различные подходы и методологии по анализу сейсмических каталогов с различным объемом данных и атрибутов землетрясений. Отметим, что методы машинного обучения, в отличие от детерминистских подходов обработки сейсмических данных [90-92], открыли новые перспективы в научных исследованиях для быстрого анализа ежедневно накапливаемых и растущих объемов сейсмических данных. Прогнозирование землетрясений – очень сложная задача, и многие исследователи использовали разные подходы для составления прогнозов. Среди этих различных подходов методы и алгоритмы, которые основаны на различных ЭС, показали многообещающие результаты в этой области [93].

Всесторонний обзор и глубокий анализ литературы в диссертационной работе, показал актуальность исследований в области применения технологии анализа данных и методов машинного обучения для решения задач оценки сейсмической уязвимости и рисков, связанные с ними. В целом обзор и анализ источников проводился по нескольким ключевым шагам, таких как цель исследования, располагаемые данные и их атрибуты, применяемые методы машинного обучения, критерий качества и результат. В будущем это даст возможность говорить о применимости тех или иных методов машинного обучения. В итоге, опираясь на вышеперечисленные особенности научно-исследовательских работ и учитывая проделанный аналитический обзор, обобщенно можно представить в таблице 1, существующие подходы и технологии оценки сейсмических рисков и сейсмических уязвимостей объектов, основанные на подходах применения технологии анализа данных и методов машинного обучения.

Таблица 1 – Существующие подходы и технологии оценки сейсмических рисков и сейсмических уязвимостей объектов

Методы и технологии оценки сейсмических рисков (событий) и сейсмических уязвимостей объектов (зданий, конструкций и т.д.)			
методы и технологии	область применения	используемые характеристики для анализа данных	результат применения
1	2	3	4
По характеристикам зданий и сооружений			
Метод опорных векторов (Support vector machine), метод поиска ассоциаций	Классификация уязвимости зданий	Год постройки, количество этажей, форма крыши, строительный материал, описание неровностей плана и высоты, положение здания т.д.)	Разработан метод
Метод опорных векторов (Support vector machine), быстрый неупорядоченный генетический алгоритм (fmGA)	Оценка сейсмостойкости школьных зданий	Входные переменные - 13 факторов, влияющих на сейсмостойкость школьного здания; выходная	Разработана система прогнозирования ускорения грунта школьного здания

Продолжение таблицы 1

1	2	3	4
		переменная – прогнозируемое значение ускорение грунта	
Обобщенная линейная регрессия (GLM), простая регрессия и нейронные сети (ANN). Смешанная модель K-means и двухступенчатая кластеризация.	Оценка сейсмической уязвимости зданий	Базовые геометрические данные о школьных зданиях	Сравнение моделей (лучшие результаты показала модель на основе GLM и K-means)
Регрессионные модели	Оценка сейсмической уязвимости местных сооружений	Параметры сооружений	Разработан метод SAVVAS
Обучение правилам ассоциации ARL (association rule learning)	Оценка сейсмической уязвимости в городских районах	Атрибуты здания (количество этажей или возраст здания) и классы уязвимости Европейской Макросейсмической шкалы EMS98	Создан прокси-сервер для определения уязвимости для всех комбинаций атрибутов здания
Байесовская сеть	Оценка уровня повреждения здания	Конструкция, риск оползня, PGA и ущерб	Новый метод оценки уровня повреждения здания
Классификация, K-средних, bootstrap	Анализ данных, связанных с повреждениями, наблюдаемыми на зданиях после землетрясения	Вертикальная структура, горизонтальные структуры, лестница, крыша и перегородка	Выявлены кластеры переменных повреждений каменных зданий для различных структурных элементов
По атрибутам сейсмических событий (землетрясений)			
Нейронные сети, машины опорных векторов и случайные леса	Исследование каталога землетрясений	Данные землетрясений	Рассчитаны 60 сейсмических индикаторов для краткосрочного прогнозирования землетрясений
Кластерный анализ (Maximization clustering method), деревья решений	Методология анализа данных о землетрясениях	Географические координаты события (широта, долгота), глубина эпицентра	Определение классов магнитуд
Нечеткая C-Means кластеризация (Fuzzy C-Means Clustering)	Анализ сейсмических данных	Магнитуда	Пространственный анализ распределения магнитуды и определение оптимального

Продолжение таблицы 1

1	2	3	4
			числа кластеров
Нечеткая логика	Получение зависимостей затухания	Эпицентральное расстояние, магнитуда и интенсивность землетрясения	Определены два нечетких набора (для магнитуд землетрясений) и третий набор данных для эпицентральных расстояний
Разделение сетки (GP), кластеризация (SC) и нечеткая кластеризация C-means (FCM)	Прогнозирование будущих землетрясений с магнитудой 5,5 и выше	Данные о землетрясениях	Адаптивная система нейро-нечеткого вывода ANFIS
Линейная регрессия	Пространственно-временной анализ землетрясений	Показатели смертности и землетрясений (магнитуда, энергия, время возникновения землетрясения, повреждения зданий)	Пространственно-временные кластеры землетрясений

Приведенные исследовательские работы и текущий аналитический обзор по данным исследованиям, показывают высокую производительность и эффективность применения технологии анализа данных и методов машинного обучения в решении подобных задач. Также заметно, что методы машинного обучения и технологии анализа данных могут использоваться для прогнозирования природных сейсмических опасностей, таких как землетрясения. Тем самым, в который раз доказывая, что технология интеллектуального анализа данных и методы искусственного интеллекта состоят из развивающегося комплекса методов и алгоритмов, которые можно использовать для извлечения ценной и уникальной информации и знаний из больших объемов данных.

Следуя из анализа источников можно утверждать, что наиболее распространенными методами являлись: метод поиска ассоциаций, линейная регрессия, нейронные сети, машины опорных векторов, определенные методы классификации и кластеризации, а также наивный Байесовский классификатор. В качестве признаков обычно выступали некоторые показатели конструкций жилых объектов, а также данные о землетрясениях. В настоящем исследовании учтены результаты анализируемых работ, а также были взяты другие методы машинного обучения, расширено множество признаков в соответствии с имеющимися данными.

1.3 Существующие информационные технологии и инструментальные средства для обработки и аналитики сейсмических данных

Информационные технологии до недавнего времени использовались для относительно рутинных вычислений, тем самым очень редко использовались, чтобы помочь обучить и визуализировать фундаментальные концепции или изучить альтернативные решения для различных систем.

В первой половине 60-х гг. прошлого столетия связано с созданием сети сейсмических станций и также в это время начали проводиться работы, связанные с процессами сбора сейсмологической информации, её обработкой и хранением. С началом с 70-х гг. проведение дальнейших работ связано с работой по внедрению электронно-вычислительных машин в исследования, связанные с сейсмологией, тем самым это привело к развитию систем по сбору, обработке, хранению сейсмологической информации [93, с. 43]. По истечению времени, в конце 80-х гг. массово вводятся персональные компьютеры, которые привели к активному процессу по автоматизации сейсмологических наблюдений [93, с. 43]. В 1996 г. уже начались работы, связанные с регистрацией сейсмических сигналов в цифровом формате [93, с. 43]. Таким образом, начиная с данного периода можно сказать начались процессы обработки сейсмологических данных в реальном режиме времени [93, с. 43]. Однако компьютерные и информационные технологии уже давно делают стремительные шаги вперед, и постоянно развивают и обогащают инновационными решениями не только IT-направление, но и также умело интегрируются в области сейсмологических наук и в сейсмостойкое строительство.

Проблемы интеграции информационных технологий в гражданское строительство, в географические науки и др., известны как сложные и динамичные. Решения этих задач требуют использования передовых компьютерных и информационных технологий для сложных математических моделирований, вычислений, обмена данными, манипулирования и хранения данных [94]. Среди зарубежных информационных технологий по исследованию и анализу надежности объектов городской среды на сейсмостойкость, а также по исследованию сейсмических событий, существуют ряд платформ, представляющие такие возможности. Основная идея таких платформ, заключается в том, чтобы сместить акцент инженерных исследований землетрясений с их нынешней зависимости от физических испытаний на комплексные эксперименты, вычисления, теорию, базы данных и моделирование. Так, к примеру, платформа под названием OpenSees [95], является фреймворком с открытым исходным кодом, который сообщество инженеров-исследователей землетрясений использует для разработки приложений для обеспечения надежности методом конечных элементов (FE) [95]. OpenSees является объектно-ориентированным по своей конструкции и, помимо достижения вычислительной эффективности, спроектирован так, чтобы быть гибким, расширяемым и переносимым. Он написан в основном на C++, но использует значительный устаревший код Fortran, многие оболочки C++

предоставляются для с внешними числовыми библиотеками, написанными на С или Fortran, такими как Lapack, ScaLapack, Petsc, Mumps, SuperLU и Arpack [96]. OpenSees имеет модульную структуру. В отличие от традиционных коммерческих и исследовательских приложений, эта модульность позволяет разработчикам расширять код в части программного обеспечения для анализа, надежности и вывода спецификаций. Это особенно важно при моделировании землетрясений, где исследователи изучают новые алгоритмы и схемы интеграции, например, для анализа обрушения и гибридных испытаний. Платформа OpenSees содержит несколько классов, для взаимодействия с процессами, выполняемыми удаленно через Интернет [96, р. 64].

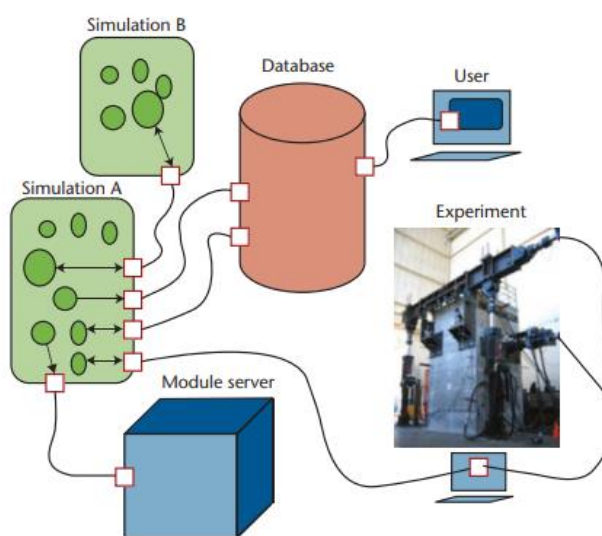


Рисунок 2 – Распределенные приложения OpenSees

Примечание – Рисунок адаптирован из работы [96, р. 64]

Распространенные распределенные интерфейсы обеспечивают доступ к приложениям удаленной базы данных и веб-сервера с использованием клиентской базы данных и библиотек URL-адресов, а также протоколов связи низкого уровня, таких как TCP/IP. Как показано в соответствии с рисунком 2, разработчики OpenSees обычно используют эти средства несколькими способами [96, р. 64]. Как уже ранее было отмечено, OpenSees предназначен для разработки приложений для моделирования работы структурных и геотехнических систем, подверженных землетрясениям. OpenSees имеет ряд приложений. Одной из таких приложений OpenSees, является экспертная система Build-X. Build-X – экспертная система для надежной предварительной и последующей обработки зданий [95]. Build-X был разработан исключительно для операционных систем Microsoft Windows. Его исходный код написан на VB.NET и разработан в среде Microsoft Visual Studio [97]. В сотрудничестве с OpenSees он облегчает статический анализ гравитационных нагрузок, анализ собственных значений, анализ спектра модального отклика и нелинейный статический анализ (Pushover) трехмерных строительных конструкций [97, р. 24]. Фактически, Build-X работает в режиме реального времени как

преобразователь выбора пользователя в сценарий Tool Command Language (Tcl), который будет использоваться в качестве входных данных для платформы OpenSees [97, p. 24]. После того, как OpenSees сгенерирует результат анализа, Build-X вызывается для обработки и представляет его в наглядной форме. Внутренняя архитектура программного обеспечения отличается от архитектуры обычной программы с графическим пользовательским интерфейсом, управляемой событиями [97, p. 24]. Build-X структурирован в соответствии с принципами системы, основанной на знаниях [98], состоящей из семи отдельных, но взаимодействующих компонентов, как показано в соответствии с рисунком 3 [97, p. 25].

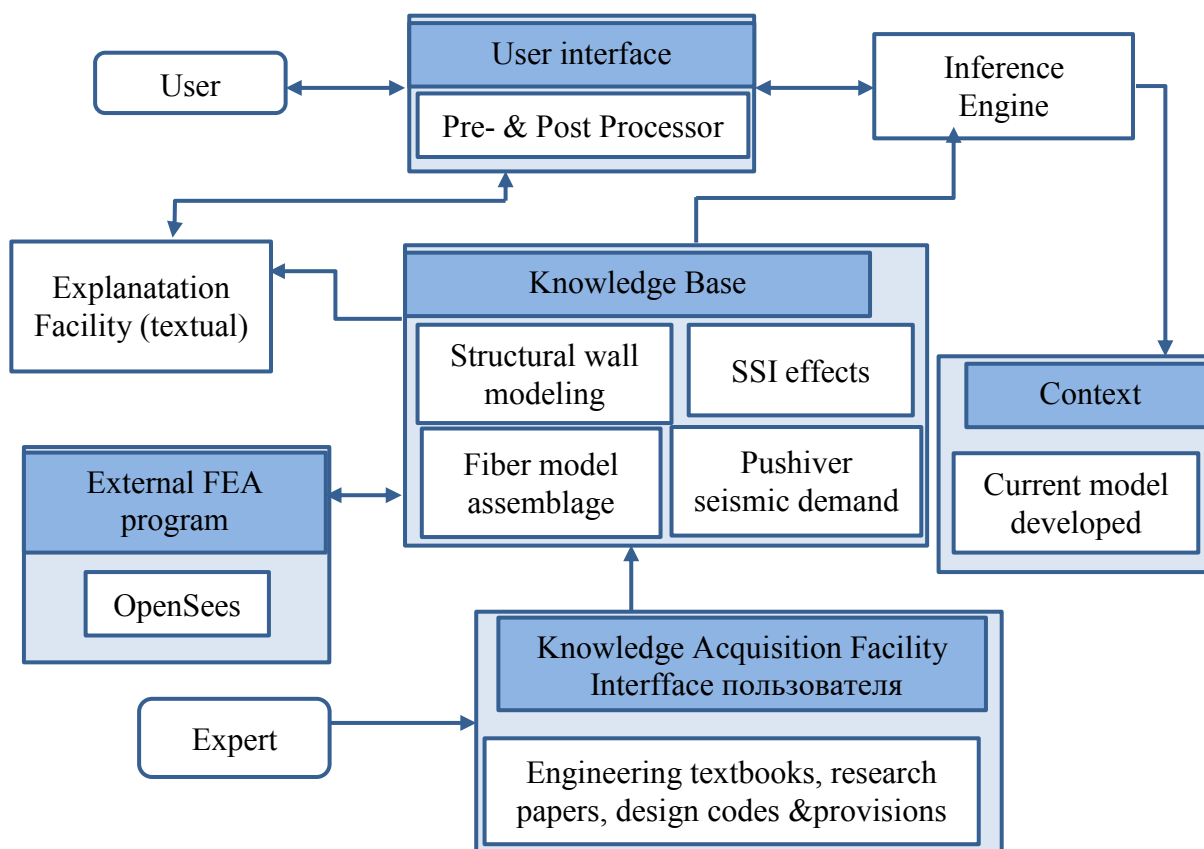


Рисунок 3 – Архитектура экспертной системы Build-X

Примечание – Рисунок адаптирован из работы [97, p. 25]

Беспрецедентная эволюция в области информатики и информационных технологий за последние десятилетия привела к разработке множества программных приложений с конечными элементами, ориентированных на численное решение структурных и геотехнических задач [97, p. 23]. Такие инструменты могут предлагать широкий спектр вариантов моделирования за счет обширных библиотек материалов и элементов, обеспечивать алгоритмическую стабильность и точность решения, а некоторые также способны реализовывать предписанные кодом процедуры проектирования [97, p. 23]. Кроме вышеприведенных программных платформ и систем, также существуют ряд других программных комплексов и систем, часть которых

представлена в таблице 2.

Таблица 2 – Зарубежные программные комплексы и платформы, предназначенные для оценки сейсмических рисков

Название программных комплексов и платформ	Назначение и функции	Особенности
1	2	3
OpenQuake Engine	Бесплатное программное обеспечение с открытым исходным кодом для оценки опасности и риска землетрясений	Работает в Linux, macOS и Windows, на ноутбуках, рабочих станциях, автономных серверах и многоузловых кластерах
EqHaz	Программный пакет с открытым исходным кодом для выполнения вероятностного анализа сейсмической опасности (PSHA) методом моделирования Монте-Карло	Разработан для оптимизации скорости, гибкости и простоты применения для типичных проблем PSHA в средах с умеренной сейсмичностью, таких как восточная часть Северной Америки (ENA)
IRMA	ИТ-платформа позволяет обмениваться данными, методами и моделями, предназначенными для оценки сейсмического риска жилых домов в Италии, чтобы соответствовать требованиям «Sendai Framework for Disaster Risk Reduction 2015-2030» (Sendai Framework)	IRMA использует OpenQuake, вычислительную машину, разработанную как часть модели глобального землетрясения, для оценки потерь от землетрясений
OPENSIGNAL	Бесплатная программная платформа с закрытым исходным кодом для обработки и отбора сейсмических записей, обработки сигналов, анализа спектров отклика, анализа спектров почвы и многого другого	Компьютерная платформа, основанная на обработке сигналов, была разработана с использованием модульного подхода к программированию, чтобы обеспечить возможность выбора и обработки записей движения грунта землетрясений
RadExPro	Программа обработки сейсмических данных под Windows. Она прекрасно подходит для контроля качества и экспресс обработки данных наземной и морской 3D/2D сейсморазведки, углубленной обработки высоко разрешающей морской сейсморазведки, обработки наземной инженерной сейсморазведки (MOB, МПВ, томография, MASW), базовой обработки ВСП	Программа развивается с 1992 года. Сотни российских геолого-геофизических компаний, проектных и научно-исследовательских институтов, университетов уже используют RadExPro. За рубежом программой пользуются геофизики более 50-и стран мира

Продолжение таблицы 2

1	2	3
SMA (Strong Motion Analyst Processing Software)	Предназначен для интерактивной обработки акселерограмм сильных движений, включая коррекцию инструмента, редактирование данных, фильтрацию, интеграцию движения земли, вычисления Фурье и спектров отклика, а также вывод файлов в формате V1, V2, V3	Коммерческий инструмент на базе Windows. Разработан компанией Kinemetrics
EQ-TOOLS (последняя версия 3.0)	Бесплатное программное обеспечение с закрытым исходным кодом для инженерного образования в области сейсмостойкости, которое позволяет пользователю выбирать, анализировать, масштабировать и изменять движения грунта	Разработан Сетью Джорджа Э. Брауна-младшего по инженерному моделированию землетрясений (NEES)
PRISM (Processing and Review Interface for Strong Motion Data [105])	Бесплатное программное обеспечение с открытым исходным кодом, используемое для обработки записей о сильных движениях	Можно установить и запустить как автономную систему в распространенных операционных системах, таких как Linux, Mac и Windows, и он является гибким и расширяемым для включения новых методов обработки сильных движений
SEISMOSIGNAL	Коммерческое программное обеспечение для обработки данных о сильных движениях	Может рассчитывать упругую, постоянную пластичность, амплитуду Фурье и спектры мощности
USDP (Utility Software for Data Processing)	Компьютерная программа, которая может использоваться для обработки данных о сильном движении грунта с помощью различных методов фильтрации и корректировки базовой линии, а также спектральных расчетов	Разработан командой METU-Earthquake Engineering Research Center и использует общедоступные исходные коды Fortran
TSPP (Time Series Processing Programs)	Разработан для обработки и управления акселерограммами сильных движений с точки зрения смещения, скорости и ускорения, временных диаграмм, отклика, спектров Фурье и фильтрации	Набор программ FORTRAN
VIEWWAVE (v2.2.0)	Бесплатное программное обеспечение с закрытым	Может читать большое количество файлов во многих

Продолжение таблицы 2

1	2	3
	исходным кодом для обработки и просмотра записей сильного движения.	форматах и может вычислять различные формы сигналов, спектр Фурье и мощности, а также спектры реакции ускорения, скорости и смещения
OpenSeismoMatlab	Программное обеспечение с открытым исходным кодом для обработки данных о сильном движении грунта, написанное в MATLAB	Может рассчитывать различные временные диаграммы и соответствующие пиковые значения, интенсивность Agias и ее временную историю
Примечание – Составлено по источникам [99-108]		

Использование последних достижений в области управления сейсмологическими пространственными данными и технологиями сейсмостойкого проектирования в борьбе с природными бедствиями (землетрясениями), включает в себя важность использования информационных технологий. Разработка современных информационных технологий и внедрение новых IT-решений (к примеру, географические информационные системы (ГИС), интеллектуальные информационно-аналитические системы (ИИАС), технологии дистанционного зондирования (ДЗ) и др.) в значительной степени могут облегчить аналитику данных и поддержку принятия решений.

Выводы по первому разделу:

1. В данном разделе проведено исследование современного состояния по вопросам сейсмической ситуации в целом и в нашей стране. В процессе исследования данного аспекта выявлены актуальные вопросы, такие как: реализация прогнозирования землетрясений с применением новых информационных технологий; оценка сейсмической опасности и уязвимости объектов населенных пунктов на территории РК научными и интеллектуальными методами; обновление и цифровизация карт сейсмического микрорайонирования территории Республики; разработка интеллектуальных, информационно-аналитических и экспертных систем для поддержки принятия решений в условиях сейсмической опасности.

2. Проведен обзор и анализ существующих отечественных и зарубежных подходов оценки сейсмического риска с возможностью применения методов машинного обучения и интеллектуального анализа сейсмических данных. Данного рода исследования показали высокую эффективность применения методов машинного обучения, которые позволяют обнаружить влияние и зависимость между определенными характеристиками объектов, подверженные сейсмическим воздействиям.

3. Изучены и описаны архитектурные модели и модели управления информационными системами по оценке сейсмической уязвимости объектов, а

также мониторинга и анализа сейсмических рисков.

4. Актуальность исследования и разработки интеллектуальных информационно-аналитических систем оценки сейсмической уязвимости объектов, мониторинга и анализа сейсмических рисков, была раскрыта обзором существующих методов и моделей оценки сейсмостойкости городских объектов и анализа сейсмических данных. Полученные результаты исследований доложены и опубликованы в сборниках международных научно-практических конференций [14, с. 111; 15, с. 100; 17, с. 108] и в издании [33, с. 322], рекомендованного Комитетом по контролю в сфере образования и науки МОН РК.

2 РАЗРАБОТКА МЕТОДОВ ОЦЕНКИ СЕЙСМОСТОЙКОСТИ ГОРОДСКИХ ОБЪЕКТОВ НА ОСНОВЕ ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ И МАШИННОГО ОБУЧЕНИЯ

2.1 Применение технологии анализа данных и методов машинного обучения для оценки сейсмостойкости городских объектов

В современных условиях жизни внедрение информатизации во все сферы человеческой деятельности, с целью развития всего общества и технологических процессов, неразрывно связано с глобальным использованием и внедрением технологий анализа данных для принятия правильных управленческих решений [109]. В соответствии с терминологией аналитической компании IDC [109, с. 8; 110] в настоящее время мы находимся на третьей технологической платформе в соответствии с рисунком 4 [111], где ключевыми индикаторами и важнейшими направлениями развития IT-сферы, являются:

- мобильные вычисления;
- облачные решения;
- Big Data/Аналитика;
- интернет вещи (IoT).



Рисунок 4 – Третья технологическая платформа IDC

Примечание – Рисунок адаптирован из работы [111]

В настоящее время стремительное развитие технологий анализа данных насколько отражают насколько они быстро прогрессируют и показывают большую перспективу применения этого направления в будущем [112]. Основные позиции в данном направлении занимают подходы, которые основываются на машинном обучении и его современных методах, и технологиях, к примеру, как Big data, Deep learning и т.д. [112, с. 113]. В целом, при решении задач для получения эффективного решения не обязательно применять самые современные интеллектуальные технологии [112, с. 113]. Известные «классические» методы машинного обучения при их правильном применении могут решать задачи, выдавая хорошие результаты, которые не уступают результатам применения более новых разработок [112, с. 113].

С точки зрения научного содержания анализ данных по сути представляет

собой несколько направлений современной математики, информатики и информационных технологий ИТ, которые занимаются разработкой различных математических моделей, алгоритмов для обьсчета/поиска решений, а также развитием технических средств для их реализации (разработка процессоров, вычислительных архитектур, низкоуровневых вычислительных алгоритмов и т.п.) [113].

В настоящее время есть большое количество разнородных методик анализа данных [114], которые представлены в соответствии с рисунком 5, на котором отражены наиболее востребованные из них [114].

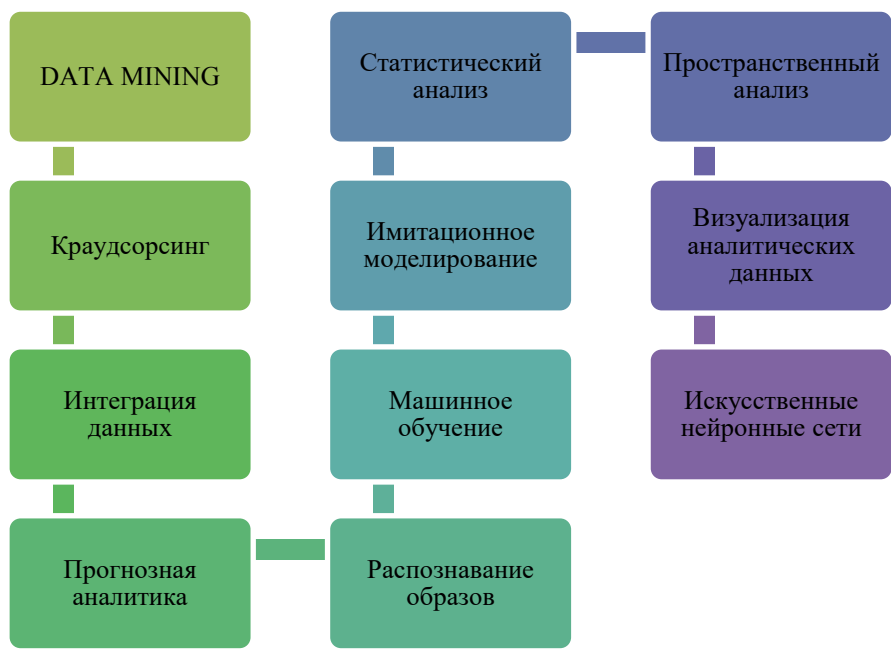


Рисунок 5 – Анализ массивов данных

Примечание – Рисунок адаптирован из работы [114]

В этой связи необходимо понимать, что все еще продолжают исследования над разработками новых методик анализа данных и над улучшением существующих способов анализа [114]. Также важно отметить, что некоторые из приведенных методик могут быть применимы как к большим данным, так и могут быть применимы к меньшим по объему массивам данных [114]. Конечно же, чем больше будут объемы данных и применятся к ним анализ, тем более релевантными на выходе получатся результаты по анализу данных [114].

Значимые достижения в области технологий анализа данных, привели к появлению новых платформ для их обработки. Как принято, массив данных хранится в различных цифровых расширениях, что представляет возможность для разработки новых методов интеллектуального и автоматического анализа данных, таких как кластеризация, классификация и др. С ростом объема данных, также прибавляются разнородные данные и доступы к ним. Недорогие цифровые устройства сделали доступными огромные наборы данных. Широкая

распространенность RFID-меток (транспондеров) из-за своей не большой стоимости и размеров способствовала к развертыванию огромного количества датчиков, которые способны передавать в режиме реального времени потоки данных. Различные электронные данные, геоинформационные данные и большое количество Web-страниц ежедневно создают терабайты данных. Большинство потоков данных являются неструктурированными и это приводит к определенным трудностям в их анализе, соответственно требуя необходимости предварительной обработки и очистки данных.

За последние десятилетия были в значительной мере увеличены мощности и скорости обработки данных. Данный факт позволил науке от ручных вычислений перейти к более легкой, быстрой, автоматизированной обработке и анализу данных. В связи с этим, у большинства специалистов базы данных возникал главный вопрос, суть которого состоит в том, что необходимо найти способы для эффективной обработки данных, а также способы для хранения и извлечения данных, тогда как у специалистов по машинному обучению, основной проблемой была разработка методов извлечения знаний из самих данных. Таким образом, чем сложнее и обширнее оказывались накапливаемые наборы данных, тем и больше представлялась возможности для извлечения необходимых сведений. На основе таких глобальных вопросов и проблем широкое применение получили технологии Data Mining [115].

Исследования в области, связанные с обработкой и анализом накапливаемых наборов цифровых данных, для исследователей открывают различные статистические шаблоны, тенденции и скрытую в данных информацию, способные помочь при построении вычислительных моделей и прогностических шаблонов. Так, подобного рода исследования показывают, что интеллектуальный анализ данных и связанные с этим технологии, интуитивно намного лучше ясны исследователям, по сравнению с традиционными методами анализа данных.

В процессе и в применении технологии анализа данных, исследователи могут найти любые аналогичные данные из набора данных, которые могут привести к любым изменениям в исследовании. Результаты от полученных исследований с помощью технологии анализа данных дают возможность визуализации данных, тем самым интеллектуальный анализ позволяет нам четко понимать представление данных. Области применения методов интеллектуального анализа довольно широки и в основе них также применяются методы, которые часто являются общеизвестными математическими методами или алгоритмами. Несмотря на это, новаторство данной технологии заключается именно в применении таких методов для решения различных задач. Данный факт стал возможным из-за увеличения возможности доступа к данным, а также из-за не дорогого хранения и обработки. Также, следует отметить, что использование различного рода графических интерфейсов, способствовало тому, что средства анализа данных стали более понятными и доступными в применении.

В целом, Data Mining технологии имеют большой потенциал при

изучении сейсмического режима и оценки сейсмической опасности территории Республики Казахстан (общее, детальное и микросейсмораионирование), мониторинга и прогноза землетрясений, оценки сейсмического риска населенных пунктов и особо ответственных объектов, прогноза ущерба от землетрясений, а также при оценке сейсмостойкости уже имеющихся объектов населенных пунктов, а именно городских объектов, зданий и сооружений. Таковыми примерами применения методов машинного обучения и интеллектуального анализа данных для оценки сейсмической уязвимости объектов городской среды, являются научные труды и изыскания различных исследователей, описанные в работах [116-119], которые доказали свою эффективность применения в решении подобных задач, и кроме всего, также более подробно уже были представлены в Разделе 1.

В диссертационной работе задача по оценке сейсмостойкости городских объектов, как уже было отмечено ранее, основывается на применении технологии анализа данных и методов машинного обучения. В основу методов машинного обучения относятся различные методы кластеризации, классификации, моделирования и прогнозирования. Ранее изученные научные труды и исследования предлагают свои подходы для решения задач, связанные с оценкой сейсмостойкости городских объектов и анализа сейсмических данных. Однако решения подобных задач все еще не предлагают более точных и логически завершенных способов оценки сейсмической уязвимости объектов с учетом сейсмических данных для прогнозирования сейсмических рисков и землетрясений. В связи с этим, данная предметная область все еще требует более тщательного исследования и выдвижения более эффективных способов.



Рисунок 6 – Этапы применения технологии анализа данных и методов машинного обучения для оценки сейсмостойкости городских объектов

В связи с этим, в данной работе предлагается способ оценки сейсмостойкости городских объектов с учетом сейсмических данных на основе применения технологии анализа данных и методов машинного обучения. Применение технологии анализа данных к данной задаче исследования можно представить в следующих этапах, которая представлена в соответствии с рисунком 6. В диссертационной работе также особое внимание уделяется применению методов машинного обучения, как основополагающим методам для решения задачи оценки сейсмостойкости городских объектов с учетом сейсмических данных.

2.2 Разработка функциональной модели оценки сейсмостойкости городских объектов

Одной из основополагающей задач диссертационной работы является задача оценки сейсмической уязвимости городских объектов, т.е. определение жилого объекта городской среды на предмет сейсмостойкости. В связи с этим, разработанный способ оценки сейсмостойкости городских объектов использует комплексное применение методов машинного обучения, включающее кластерный анализ методом K-Means и алгоритмы классификации, такие как Decision Tree (деревья решений) и Random Forest (Случайный лес).

На начальной стадии исследования при реализации разработанного способа оценки сейсмостойкости городских объектов с целью осуществления интеллектуального анализа набора данных о жилых объектах, использовался метод кластерного анализа K-Means. Применение метода K-Means, позволило выявить группы схожих объектов и получить кластеры типовых городских объектов с их соответствующими характеристиками. Следующим шагом в исследовании для каждого кластера строятся собственные деревья решений, которые позволяют уточнить границы полученных решений. Для улучшения работы предложенного алгоритма используется контролируемый алгоритм обучения - Random Forest. Использование метода Random Forest позволило выявить главные характеристики, принимающие участие в прогнозе сейсмостойкости объекта, а также уменьшить переобучение путем усреднения результата. Применение методов классификации позволило точно увидеть разбиение набора данных по характеристикам и проанализировать переменные, по которым сделан прогноз основных характеристик, которые определяют сейсмостойкость объектов. В качестве набора данных для исследования были использованы данные, предоставленные в отчетной документации АО «КАЗНИИСА» (Казахский научно-исследовательский и проектный институт строительства и архитектуры) о городских объектах с 19 характеристиками, которые принадлежат к определенным микрорайонам города (Алматы). Для интеллектуального анализа данных были использованы следующие характеристики данных, которые представлены в таблице 3.

Таблица 3 – Названия характеристик из набора данных о городских объектах

Названия характеристик объектов, в наборе данных	Названия характеристик объектов (на английском языке), используемые в анализе данных
Наименование объекта	object
Местоположение	location
Год разработки проекта	pr_develop_year
Год постройки здания	year_constr
Вид проекта	type_pr
Категория грунтов по сейсмическим свойствам	seism_cat_soils
Сейсмичность района	space-plan_sol
Этажность	floors
Наличие антисейсмических мероприятий	antiseism_activ
Общая характеристика	gen_char
Фундаменты	foundations
Конструкции перекрытий	bearing_struct
Несущие конструкции	floor_struct
Стеновое ограждение	wall_fence
Перегородки	partitions
Высота (H), м	height
Общая площадь, м ²	total_area
Строительный объем, м ³	construct_vol
Оценка сейсмической уязвимости (сейсмостойкости)	assessment

Полная структура этапов реализации разработанного способа оценки сейсмостойкости городских объектов представлена в соответствии с рисунками 7, 8 в виде функциональных моделей в нотации IDEF0.



Рисунок 7 – Функциональная модель оценки сейсмостойкости городских объектов в нотации IDEF0

В предлагаемой структуре функциональной модели оценки сейсмостойкости городских объектов реализуются методы интеллектуального анализа данных и машинного обучения. Согласно функциональной модели в соответствии с рисунком 8, после этапа предварительной обработки набора данных, на следующем втором этапе реализация разработанного способа состоит из кластерного анализа. В следующем этапе проводится построение классификационной модели дерева решений и в заключительном этапе с помощью усиленного классификатора RF, выявлены и определены важнейшие характеристики, участвующие в прогнозной оценке сейсмостойкости городских объектов.

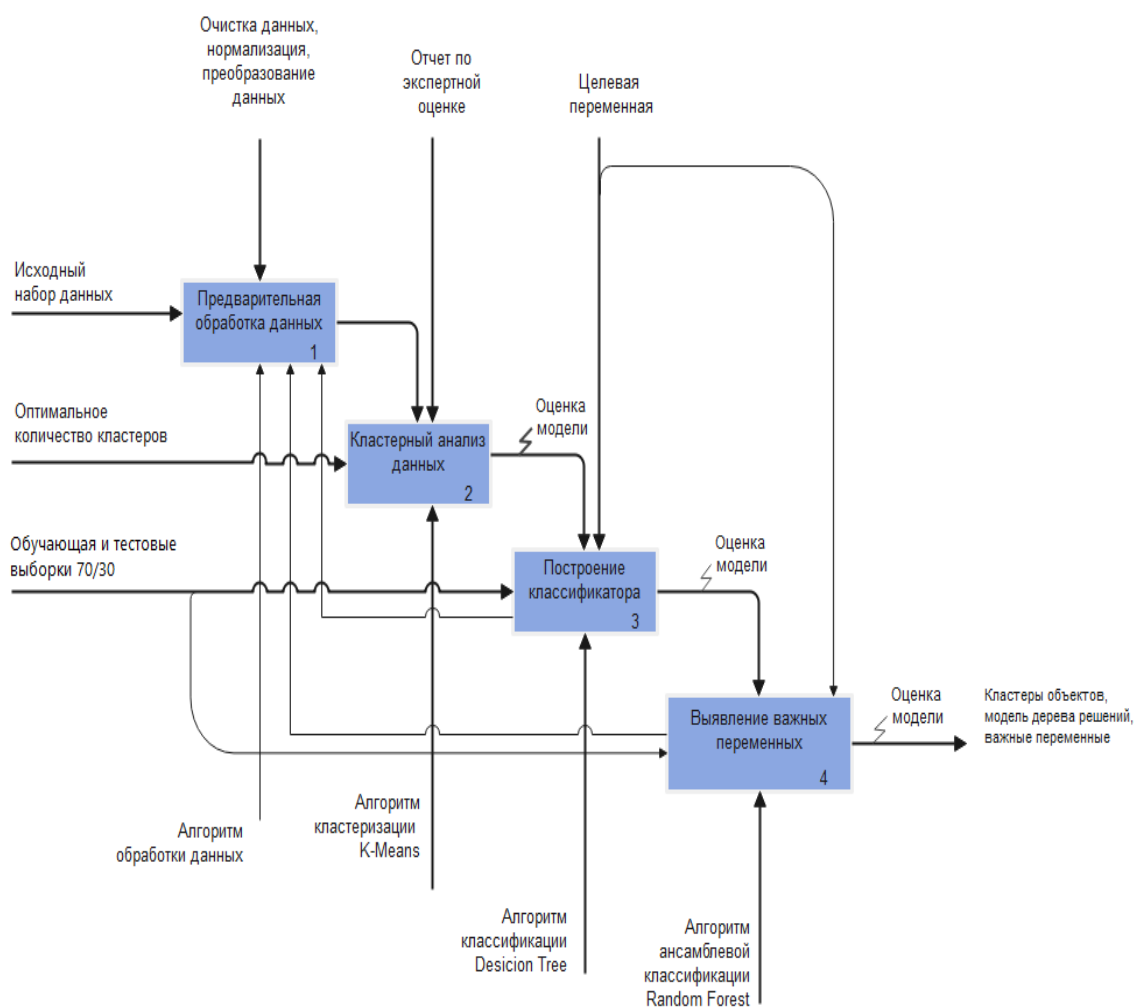


Рисунок 8 – Декомпозиция функциональной модели оценки сейсмостойкости городских объектов в нотации IDEF0

2.2.1 Оценка сейсмостойкости городских объектов методами кластерного анализа

Согласно задачам диссертационной работы и поставленной в ней цели, реализация задачи по разработке способа оценки сейсмостойкости городских объектов, была достигнута с помощью технологии анализа данных и методов машинного обучения. Этапы исследования данной задачи и процесс реализации

комплексного подхода оценки сейсмостойкости городских объектов, был представлен в структуре функциональной модели, отображенные в соответствии с рисунками 7, 8. В соответствии со структурой функциональной модели оценки сейсмостойкости городских объектов и вышеуказанного ее описания, на первоначальном этапе для обработки набора данных о городских объектах, а также в целях получения группы объектов со схожими признаками, т.е. кластеров, использовался кластерный анализ методом K-Means. Следует отметить, что среди методов кластерного анализа наиболее популярным, является метод кластеризации K-Means. Метод кластерного анализа данных является одним из важных методов машинного обучения для обнаружения знаний в многомерных наборах данных.

Рассматриваемый алгоритм является хорошо масштабируемым для применения к большим объемам данных, поэтому может быть использован во многих областях [120]. Алгоритм вполне доступен в реализации, и достаточно хорошо программируется. В основном к его использованию прибегают в том случае, если перед аналитиком предстает задача поиска групп в неразмеченном наборе данных. Также алгоритм K-Means является исторически одним из самых важных алгоритмов, применяемый в интеллектуальном анализе данных. Данный метод был открыт несколькими исследователями различных дисциплин, в первую очередь Ллойдом (1957, 1982), Форджи (1965), Фридманом и Рубином (1967), и МакКуином (1967) [121].

Алгоритм K-means кластеризует данные путем распределения элементов исходной выборки по n группам равных дисперсий, при этом минимизируя параметр внутрикластерной суммы квадратов. Каждая из полученных таким способом групп именуется кластером и характеризуется единственным параметром – центроидом кластера. Центроиды могут и не относиться к элементам исходной выборки, хотя могут существовать в том же пространстве, что и все остальные элементы. Таким образом, кластерный анализ стремится разделить данный набор данных на группы на основе заданных характеристик, чтобы точки данных в группе были более похожи друг на друга, чем точки в разных группах [121, р. 247]. Основная идея метода заключается в определении k центроидов, по одному для каждого кластера. Алгоритм направлен на минимизацию целевой функции формула (1), в общем случае квадрат функции ошибок [122]. Целевая функция [111].

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2 \quad (1)$$

где $\|x_i^{(j)} - c_j\|^2$, выбранная мера расстояния между точкой данных $x_i^{(j)}$ и центром кластера c_j является показателем расстояния n точек данных от их соответствующих центров кластера [122, р. 262; 123]. В основе алгоритма для определения расстояния между точками данных и центроидами, в качестве

меры близости используется Евклидово расстояние [124]. Евклидово расстояние между двумя многомерными точками данных $X = (x_1, x_2, x_3, \dots, x_m)$ и $Y = (y_1, y_2, y_3, \dots, y_m)$ будет описываться следующей формулой (2):

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_m - y_m)^2} \quad (2)$$

Исходя из основной интерпретации и определения сути метода кластеризации К-Means, можно построить следующий алгоритм реализации этого метода. Алгоритм данного метода кластерного анализа представлен в соответствии с рисунком 9. В кластерном анализе одной из важных задач является подбор оптимального значения k . В данном методе существует проблема – необходимость априори задавать фиксированное число кластеров для разбиения, которое, безусловно, далеко не всегда выбирается оптимальным.

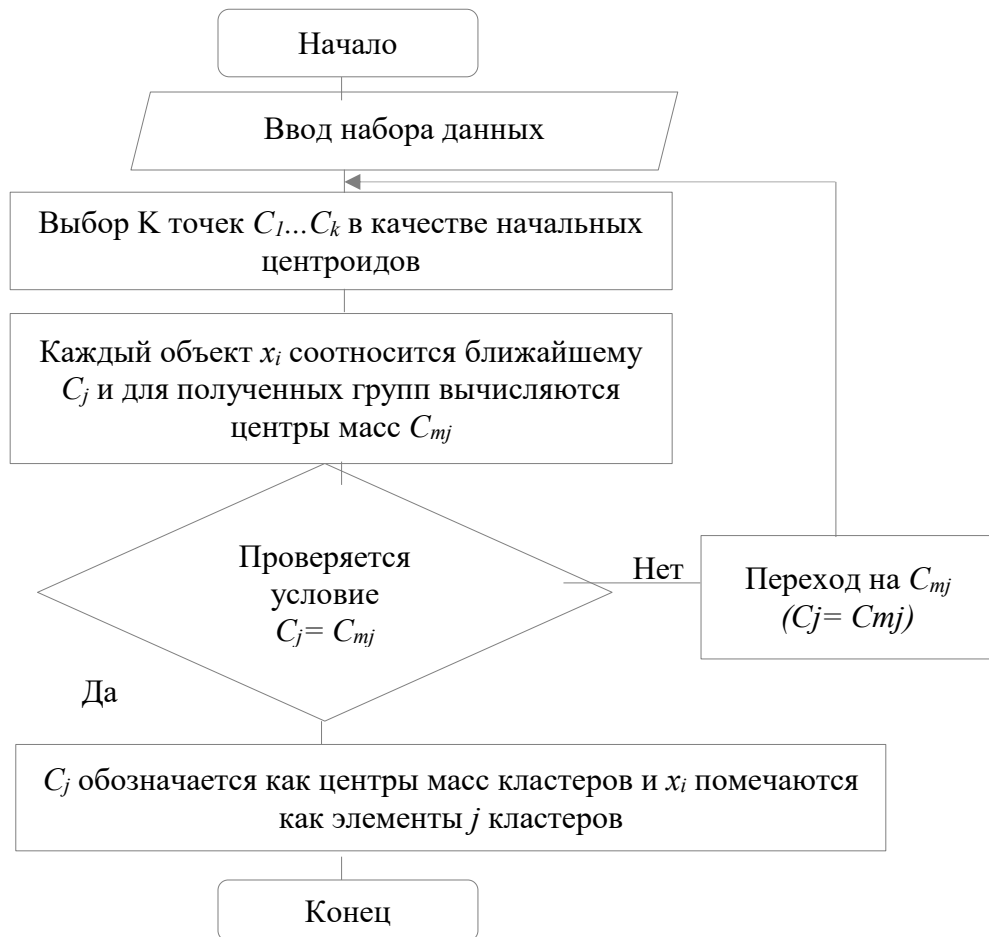


Рисунок 9 – Алгоритм метода К-Means

В исследовании при кластеризации набора данных городских объектов алгоритмом К-Means для определения оптимального количества кластеров, использован метод «локтя» («Elbow method») [16, p. 156; 122, p. 263; 124].

Данный метод рассматривает характер изменения разброса W_{total} с увеличением числа групп k [16, р. 156]. Объединив все n наблюдений в одну группу, имеем наибольшую внутрикластерную дисперсию, которая будет снижаться до 0 при $k \rightarrow n$.

Алгоритм определения оптимального количества кластеров будет выглядеть следующим образом:

1. Вычисляется алгоритм кластеризации методом K-Means для различных значений k . В исследовании значение k изменялось от 1 до 10.

2. В следующем шаге для каждого k вычисляется общая сумма квадратов внутри кластера.

3. Необходимо построить кривую по вычислениям шага 2 в зависимости от количества кластеров k .

4. Находится расположение изгиба на графике и выбирается значение соответствующего количества кластеров.

Как видно, из ниже представленного в соответствии с рисунком 10 [16, р. 156; 122, р. 263], оптимальное значение k для набора данных составляет 4, тем самым метод «локтя» определил $k=4$, как оптимальное количество кластеров.

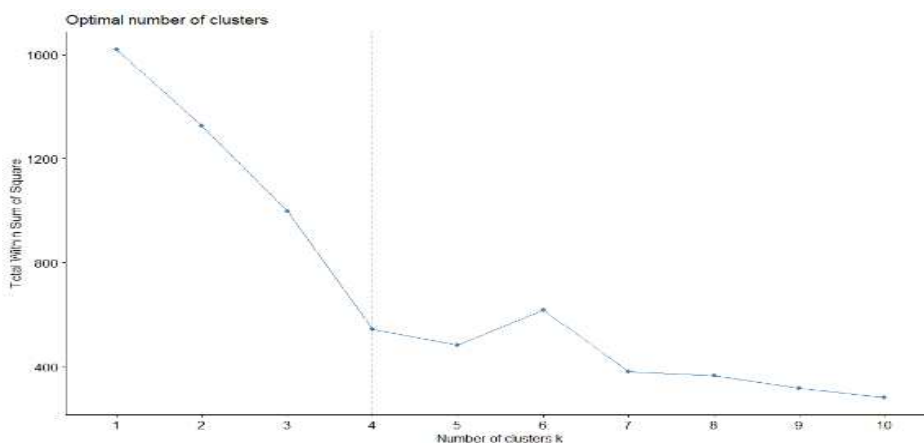
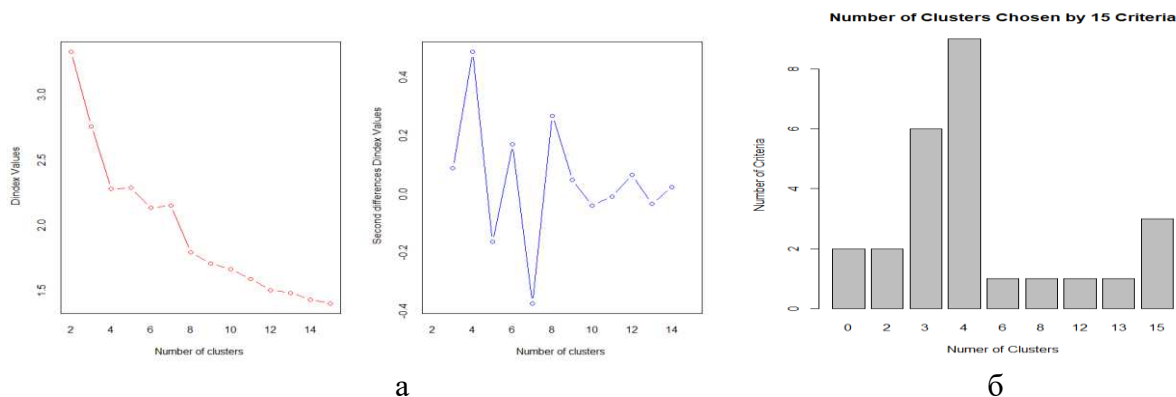


Рисунок 10 – Определение оптимального количества кластеров методом «локтя» («Elbow method»)

В реализации кластерного анализа набора данных о городских объектах, при определении оптимального количества кластеров, кроме метода «локтя», также был применен индекс Данна в соответствии с рисунком 11 (пакет NbClust, данный пакет используется для определения наилучшей схемы кластеризации по разным результатам, полученным путем варьирования всех комбинаций количества кластеров и методов расстояния [124, с. 104], инструментарий R+Rstudio), с той же целью получения наилучшего количества кластеров и для сравнения достоверности полученных результатов оптимального количества кластеров.



а – по индексу Данна; б – в виде графика barplot

Рисунок 11 – Определение оптимального количества кластеров

Цель применения индекса Данна заключается в том, чтобы идентифицировать наборы кластеров, которые являются хорошо разделенными и имеют небольшую дисперсию между членами кластера, а также средние значения разных кластеров будут достаточно далеко друг от друга по сравнению в пределах кластерной дисперсии. Таким образом, чем выше значение индекса Данна, тем лучше будет кластеризация. Тем самым, количество кластеров, которое максимизирует индекс Данна будет приниматься за оптимальное количество кластеров k . Как видно из графиков в соответствии с рисунком 11а «Dindex Values» относится к индексу Данна [124, с. 105]. Более высокие значения индекса Данна указывают на лучшую кластеризацию и в нашем случае значение $k=4$ [124, с. 105]. Также, оптимальное количество кластеров определяется как количество кластеров, выбранных по наибольшему количеству критериев [124, с. 105], которое представлено в соответствии с рисунком 11б [124, с. 105]. Выявление оптимального количества кластеров при $k=4$, в дальнейшем позволило получить 4 кластера городских объектов, которые были получены с помощью метода кластеризации K-Means. В результате в соответствии с рисунком 12, получены 4 кластера, и в таблице 4 представлены их описания [124, с. 105].

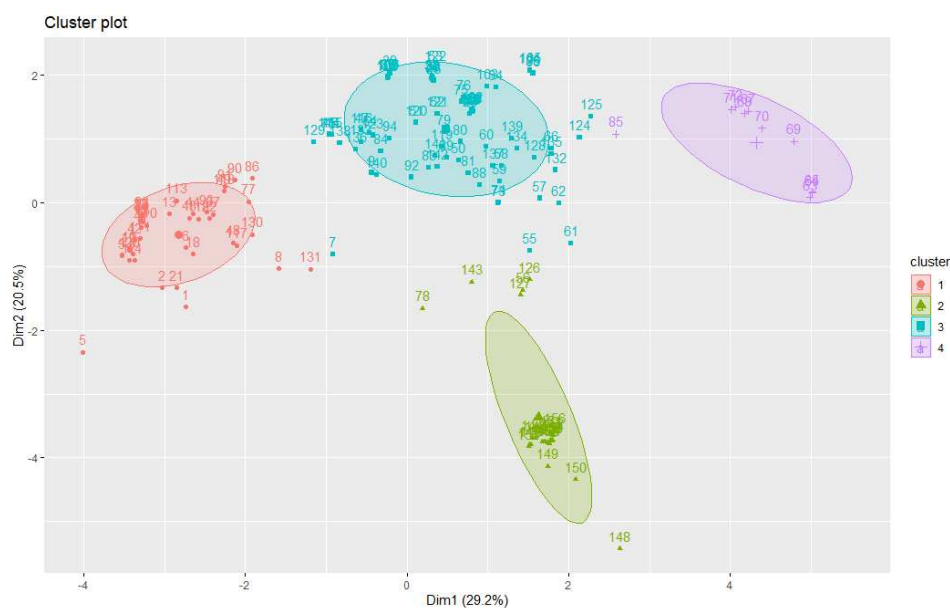


Рисунок 12 – Результат кластерного анализа методом К-Means

Таблица 4 – Описание полученных кластеров

Кластер	Размер кластера	Среднее расстояние между кластерами
1	44	0.47
2	25	0.40
3	84	0.35
4	10	0.64

Прежде чем, как интерпретировать результаты, полученные кластерным анализом с помощью метода К-Means, в исследовательской работе на последующем шаге проведена оценка качества результатов кластеризации с помощью метода оценки Silhouette Index (SI) [125] – индекс силуэта. Метод оценки Silhouette Index (SI) является метрикой, которая позволяет оценить качества кластеризации основываясь лишь на самой выборке. Коэффициент SI рассчитывается для каждого элемента в наборе данных. Рассматриваемая метрика способна указывать на те объекты, которые были хорошо помещены в их кластер, а также на те объекты, которые не соотнесены какому-либо кластеру и находятся между кластерными объектами.

Расчет в данной метрике основывается на разности двух величин, т.е. среднем расстоянии между объектом и всеми остальными объектами того же кластера, а также минимальным расстоянием между объектом и объектами ближайшего к данному кластера. Силуэт всей кластерной структуры (Silhouette Width Criterion – SWC) определяется по следующей формуле (3):

$$SWC = \frac{1}{N_x} \sum_{j=1}^{N_x} S_{xj} \quad (3)$$

как деленная на количество элементов в кластеризуемом множестве N_x сумма Силуэтов каждого отдельного элемента, определяемых по формуле (4):

$$S_{xj} = \frac{b_{pj} - a_{pj}}{\max(a_{pj}, b_{pj})} \quad (4)$$

где a_{pj} – это среднее расстояние от объекта до других объектов своего кластера, а

b_{pj} является средним расстоянием от объекта до других объектов ближайшего другого кластера. Результат вычитания нормируется наибольшим значением, для приведения оценки к фиксированному диапазону. Итоговым значением Silhouette для всей выборки является усредненное значение результатов для каждого из элементов выборки.

Коэффициент Silhouette измеряет, насколько хорошо кластеризовано наблюдение, и оценивает среднее расстояние между кластерами, т.е. среднюю ширину силуэта. В соответствии с рисунком 13 [124, с. 106] по кластеру 2, видно, что наблюдение с отрицательным силуэтом расположено не в той группе. Силуэтный коэффициент ($S(i)$) измеряет, насколько объект i похож на другие объекты в своем кластере по сравнению с объектами в соседнем кластере. В целом значения $S(i)$ находятся в диапазоне от +1 до -1, где +1 – лучший результат, -1 – худший результат. Отрицательные значения указывают на то, что точка назначена неправильному кластеру и значение 0 указывает на перекрывающийся кластер. В рассматриваемом кластерном анализе силуэтный коэффициент $S(i) = 0,41$.

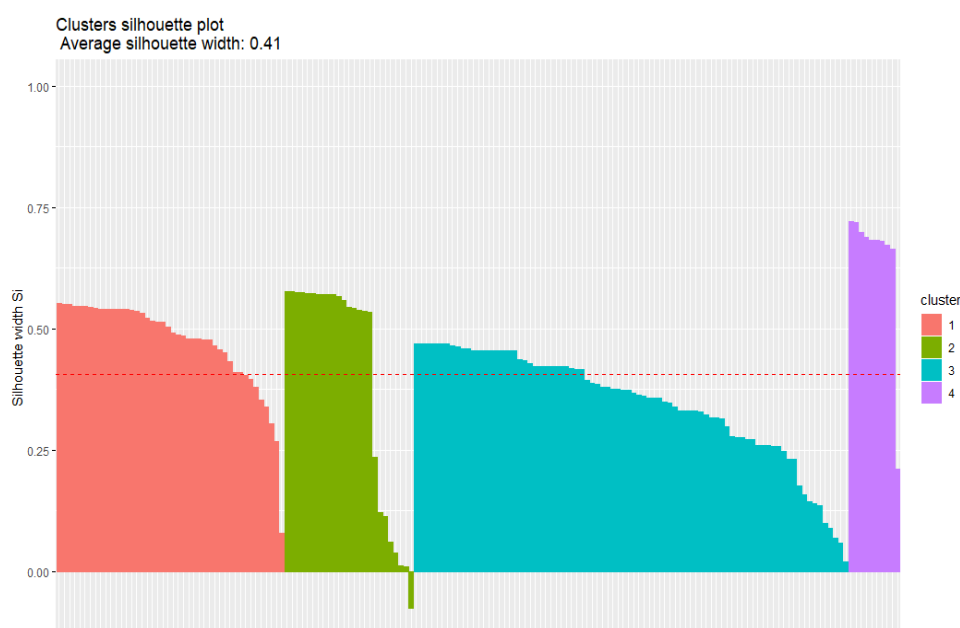


Рисунок 13 – Результат применения метода Silhouette Index для оценки качества кластеризации

Следуя из вышесказанного, наблюдение с отрицательным силуэтом в соответствии с рисунком 13 [124, с. 106], соответствует наблюдению в таблице 5 [124, с. 106].

Таблица 5 – Наблюдение с отрицательным силуэтом

№	Кластер	Neighbor	sil width
56	2	3	-0.07663963

Итак, получив полный кластерный анализ по набору данных, который состоит из 250 наблюдений и 15 характеристик городских объектов, можно привести описание по полученным кластерам и дать оценку сейсмостойкости объектов городской среды [124, с. 106]. По кластеру 2, наблюдение в точке 56, детально было изучено и по сходству элементов относится к кластеру 3 [124, с. 106]. Также наблюдение в точке 85, относится к сейсмоопасному объекту и по мере сходству ближе к кластеру 4 [124, с. 106]. По кластеру 1 (таблица 6), выявлены объекты сейсмостойкие объекты, в основном они 1975, 1978, 1986, 1987, 1993, 1994 годов постройки и имеют общую характеристику крупнопанельного здания [124, с. 106]. В кластере 2 (таблица 7), сгруппированы тоже сейсмостойкие объекты, которые 1973, 1975, 1983 годов конструкции [124, с.106]. Данные объекты имеют общую характеристику как каркасно-панельные здания с буронабивными сваями, где несущие стены кирпичные [124, с.106]. Кластер 3 (таблица 8), сейсмобезопасные объекты 1972, 1985, 1992 годов конструкции [124, с. 107]. Общая характеристика жилых объектов каркасный тип. Кластер 4 (таблица 9), не сейсмостойкие объекты 1932, 1936, 1952, 1954, 1956, 1958 годов конструкции, кирпичные здания [124, с. 107]. Категория грунтов 3 [124, с. 107].

Таблица 6 – Кластер 1 – сейсмостойкий

Объект	Жилое здание	Общая характеристика	Крупнопанельное здание
Год конструкции	1975, 1978, 1986, 1987, 1993, 1994	Фундаменты	железобетонные ленточные
Вид проекта	типовой - серия КЗ- 464ДС; типовой - серия 158	Несущие конструкции	железобетонные панели
Категория грунтов	1	Конструкции перекрытий	железобетонные панели
Объемно-планировочные решения	прямоугольное в плане, без перепадов по высоте, подвал; прямоугольное в плане, с перепадом по высоте, подвал, антисейсмические швы	Стеновое ограждение	железобетонные панели
Этажность	5, 9	Перегородки	железобетонные панели
Наличие антисейсмических мероприятий	есть	Высота	13, 15, 27
Год конструкции	1973, 1975, 1983	Фундаменты	буронабивные сваи

Таблица 7 – Кластер 2 – сейсмостойкий

Объект	Жилое здание	Общая характеристика	Каркасно-панельное здание
Вид проекта	индивидуальный	Несущие конструкции	железобетонный каркас
Категория грунтов	3	Конструкции перекрытий	железобетонные плиты
Объемно-планировочные решения	прямоугольной формы в плане, без перепадов по высоте, с подвалом, антисейсмические швы	Стеновое ограждение	железобетонные навесные панели
Этажность	3	Перегородки	кирпичные, гипсопрокатные
Наличие антисейсмических мероприятий	есть	Высота	9

Таблица 8 – Кластер 3 – сейсmobезопасный

Объект	Жилое здание	Общая характеристика	Каркасное здание
Год конструкции	1972, 1985, 1992	Фундаменты	железобетонные столбчатые
Вид проекта	Индивидуальный; типовой - серия ВП	Несущие конструкции	железобетонный каркас
Категория грунтов	3, 2	Конструкции перекрытий	железобетонные плиты
Объемно-планировочные решения	прямоугольное в плане, с перепадом по высоте, подвал; прямоугольной формы в плане, без подвала	Стеновое ограждение	кирпичные стены
Этажность	3, 5	Перегородки	кирпичные, гипсобетонные
Наличие антисейсмических мероприятий	есть	Высота	8.1, 13.5

Таблица 9 – Кластер 4 – не сейсмостойкий

Объект	Жилое здание	Общая характеристика	Кирпичное здание
1	2	3	4
Год конструкции	1932, 1936, 1952, 1954, 1956, 1958	Фундаменты	бутобетонные ленточные; железобетонные ленточные
Вид проекта	индивидуальный	Несущие конструкции	кирпичные стены
Категория грунтов	3	Конструкции перекрытий	деревянные

Продолжение таблицы 9

1	2	3	4
Объемно-планировочные решения	прямоугольной формы в плане, без подвала; прямоугольной формы в плане, без перепадов по высоте, с подвалом	Стеновое ограждение	кирпичные стены
Этажность	2	Перегородки	кирпичные, деревянные
Наличие антисейсмических мероприятий	Нет	Высота	6.0, 6.6, 6.8, 8.1

Построение кластерной модели оценки сейсмостойкости городских объектов, также была исследована другим методом кластеризации, которая сочетает в себе гибридный подход.

Алгоритм K-Means решает типы задач, имеющие предположения или гипотезы относительно числа кластеров, где они должны быть различны между собой насколько это возможно [122, р. 262]. Результат кластеризации методом K-Means является чувствительным к начальному случайному выбору центров кластеров и может отличаться при каждом вычислении k-средних [122, р. 263]. Для того, чтобы проверить объективность полученных кластеров, в решении задачи оценки сейсмостойкости городских объектов для сравнительного анализа полученных результатов кластеризации, был также использован гибридный подход, а именно гибридная иерархическая кластеризация K-Means (НК-means) [122, р. 263]. В данном подходе комбинируется иерархическая кластеризация и метод K-Means. В иерархической кластеризации основной целью является построение структуры кластеров [122, р. 262]. В данном случае применен объединяющий вид алгоритма [122, р. 262]. Расстояние между кластерами принимают за меру, используемую для определения, какие кластеры должны быть объединены, а какие кластеры требуется разделить [126, 127].

В объединяющем иерархическом алгоритме каждый элемент из множества X наблюдений принимается как отдельный кластер [122, р. 262]. Далее на каждом шаге работы алгоритма в большей степени похожие пары элементов объединяются в один кластер [122, р. 262]. Такое условие объединения кластеров представлено в формуле (5) [122, р. 262].

$$D = \min(\text{dist}(a,b)) \quad (5)$$

где a,b принадлежит X.

В результате работы алгоритма получается структура кластеров, представляющая из себя граф [122, р. 262]. Работа алгоритма завершается при достижении требуемого количества кластеров [127, р. 262].

Применение гибридной иерархической кластеризации K-Means (НК-

means) к исследуемым наборам данных, как видно в соответствии с рисунком 14, также точно показал результат по кластерам, как и в обычном методе кластеризации K-Means.

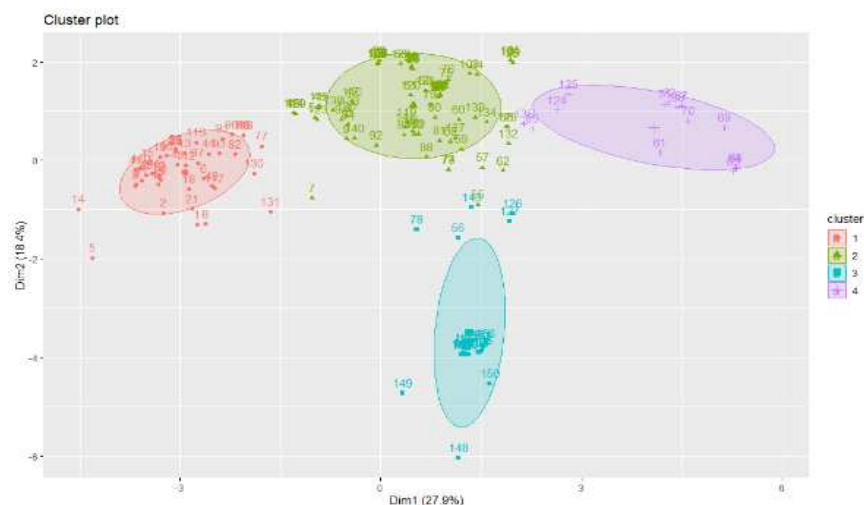


Рисунок 14 – Кластеры, полученные методом НК-means

Сравнительные результаты полученных кластеров, методами обычной кластеризации K-means и гибридного подхода кластерного анализа НК-means, в целом, показали достоверность полученных результатов кластеризации, таблица 10. В таблице 10, кроме метрики Silhouette, для оценки качества кластеров, также были использованы меры гомогенность, полнота и V-мера.

Таблица 10 – Оценка качества результатов кластеризации различными метриками

Методы кластерного анализа	Гомогенность Homogeneity	Полнота Completeness	V-мера V-measure	Силуэт Silhouette
K-means	1,00	1,00	1,00	0,41
НК-means	1,00	1,00	1,00	0,41

В реализации задачи по разработке способа оценки сейсмостойкости городских объектов, применение метода кластерного анализа на начальном этапе исследования, позволило выявить группы схожих объектов, т.е. кластеры, которые требовали дальнейшего анализа структуры полученных объектов с целью построения классификационной модели оценки сейсмостойкости городских объектов, а также выявления информативных признаков, влияющие на решение оценки сейсмостойкости городских объектов. Тем самым, в продолжении исследования набора данных городских объектов, были применены методы классификации DT и RF. Построение деревьев решений позволило наглядно продемонстрировать структуру данных, участвующие в решении оценки сейсмостойкости городских объектов, которое создало работающую модель классификации данных. Дерево решений в каждом кластере уточняет границы решений, изучая подгруппы в кластере.

Использование метода классификации RF, позволило точно увидеть разбиение по характеристикам набора данных и выявить главные переменные, участвующие в построении модели дерева решений.

2.2.2 Оценка сейсмостойкости городских объектов методом Decision Tree

Реализация комплексного подхода в оценке сейсмостойкости городских объектов, основывалось на разработке способа оценки важности влияния факторов на сейсмостойкость городских объектов методами кластерного анализа и классификации [122, р. 261]. На первоначальном этапе был проведен разведочный анализ, где были изучены структуры данных городских объектов методами кластерного анализа K-means и НК-means [122, р. 261]. В последующих шагах для каждого кластера строятся собственные деревья решений, которые позволяют уточнить границы полученных решений, представляя возможность построения классификационной модели прогнозирования сейсмостойкости объекта с помощью метода Decision Tree. На заключительном этапе исследования в разработке способа оценки сейсмостойкости городских объектов, для уточнения построенной модели дерева решений и выявления наиболее важных характеристик, влияющие на прогнозную модель сейсмостойкости объектов [122, р. 261], был применен усиленный классификатор RF.

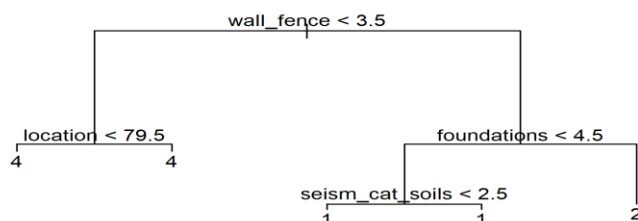
Построение классификационной модели прогнозирования сейсмостойкости объекта методом DT [128, 129], разбивают данные несколько раз в соответствии с определенными значениями отсечения в объектах. В дереве решений критерий разбиения и условие останковки уже заранее будут определены на основе энтропии [122, р. 262]. При разбиении создаются различные подмножества набора данных и каждый экземпляр принадлежит одному подмножеству [122, р. 262]. Конечные подмножества являются конечными узлами, а промежуточные подмножества называются внутренними узлами [122, р. 262]. Для прогнозирования результата в каждом узле листа используется средний результат обучающих данных в этом узле. В каждом узле дерева решений энтропия определяется по формуле (6) [122, р. 262].

$$E = - \sum_{i=1}^c p_i \times \log(p_i) \quad (6)$$

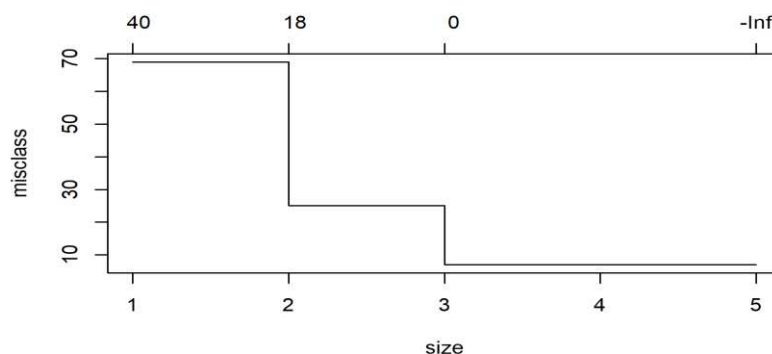
где p_i представляет долю наблюдений с метками класса i , где $i = \{1 \dots c\}$ [122, р. 262].

На текущем этапе исследования, особый интерес представляет выявление важных параметров, как вид проекта и конструкционных решений, которые в зависимости от сейсмической активности района и категории грунтов по сейсмическим свойствам, будут определять городской объект на сейсмостойкость [16, р. 156]. В построении классификационной модели прогнозирования сейсмостойкости объекта методом DT, были использованы

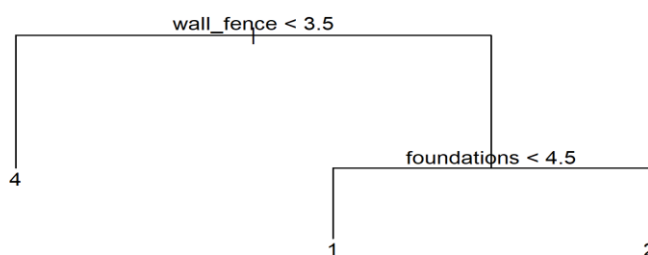
пакеты `tree`, `caret`, `rpart` в среде RStudio. При работе с данными для оценки производительности деревьев решений, используя пакет `caret`, исходный набор данных был разделен на обучающую (70%) и тестовую (30%) выборки. На основе обучающих наборов данных построена модель дерева решений прогнозирования сейсмостойкости городского объекта. В данную модель дерева решений ввели набор тестовых данных, для того чтобы увидеть насколько точно достигнуты ожидаемые результаты в соответствии с рисунком 15а [16, p. 156].



а



б

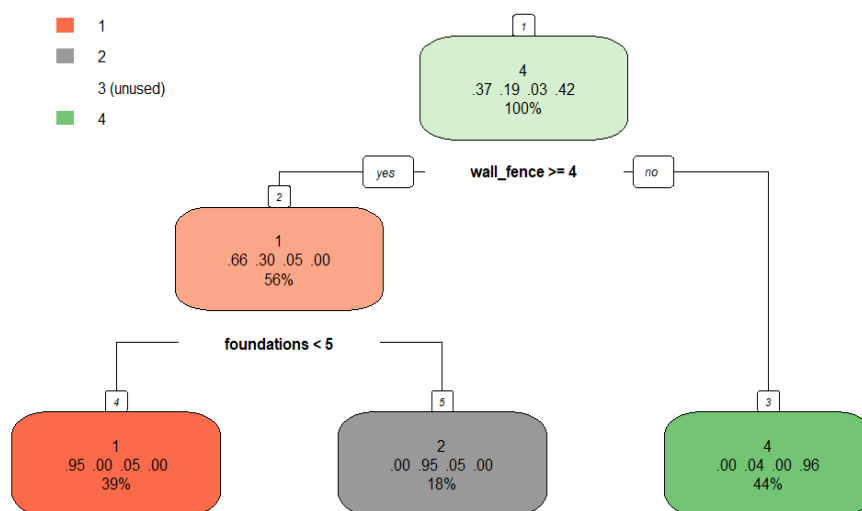


в

а – дерево решений; б – обрезка дерева; в – прогнозная модель дерева решений

Рисунок 15 – Результат применения метода дерева решений с помощью пакета «tree»

С помощью функции `CV.tree` в среде RStudio, используя k-кратный метод перекрестной проверки, обучающая выборка была протестирована несколько раз, и выбраны ветви с наименьшей дисперсией. В соответствии с рисунком 15б [16, p. 156], функция `CV` показывает, что дерево решений с тремя ветвями имеет самую низкую дисперсию, поэтому дерево решений было обрезано. Таким образом, согласно полученной прогнозной модели дерева решений для оценки сейсмостойкости объекта городской среды в соответствии с рисунком 15в, если значение параметра `wall_fence` $< 3,5$ (стенное ограждение), то объект считается сейсмостойким, в противном случае следует проверять фундамент объекта, где если фундаменты $< 4,5$, то объект не сейсмоустойчив, в противном случае считается сейсмотехнически безопасным [16, p. 157; 122, p. 264]. Эффективность предложенных моделей оценивалась по ряду показателей, определенных с использованием матрицы неточностей (`confusion Matrix`). Данная модель показала точность 91,3% [16, p. 157; 122, p. 263]. Также для более наглядного представления построенной классификационной прогнозной модели дерева решений, использовался пакет `rpart` в соответствии с рисунком 16.



1 – не сейсмостойкий; 2 – сейсмотехнически безопасный; 3 – сейсмоопасный; 4 – сейсмостойкий

Рисунок 16 – Построение прогноза методом дерева решений с помощью пакета «`rpart`»

В решении задачи о сейсмостойкости или несейсмостойкости городского объекта, прогнозная модель дерева решений использует характеристики конструкции как `wall_fence` (стенное ограждение) и `foundations` (фундаменты).

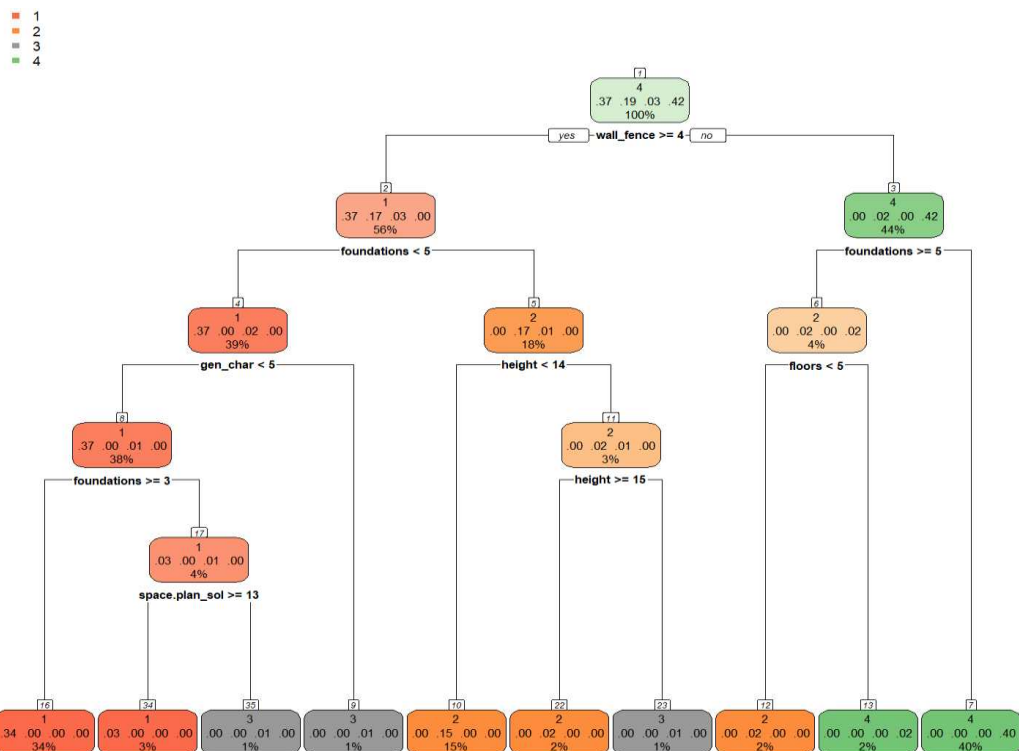


Рисунок 17 – Прогноз с использованием дерева решений оценки сейсмостойкости городских объектов

Эффективность рассмотренной другой модели, построенной в пакете gpart согласно матрице неточностей (confusion Matrix) классифицируют городские объекты на сейсмостойкость с точностью 93,6%. Полная структура прогнозной модели оценки сейсмостойкости городских объектов, представлена в соответствии с рисунком 17. Для оценки качества работы полученного классификатора, также был использован ROC-анализ (Metz, 1978; Fawcett, 2005) [130]. ROC (Receiver Operator Characteristic) – кривая или кривая ошибок – график, позволяющий оценить качество классификации, который отображает соотношение между чувствительностью (TPR, True Positive Rate) алгоритма и долей из объектов отрицательного класса, которые алгоритм предсказал неверно (FPR, False Positive Rate) при варьировании порога решающего правила [131].

В случае бинарной классификации прогностическая характеристика классификатора, как принято измеряется с использованием матрицы неточностей в таблице 11, имеющая четыре значения:

1. TP (True Positives, истинные положительные результаты), количество правильно спрогнозированных положительных примеров.
2. FN (False Negatives, ложно отрицательные), количество неверно предсказанных положительных случаев.
3. FP (False Positives), (ложные срабатывания), количество неверно предсказанных отрицательных случаев.
4. TN (True Negatives, истинно отрицательные) - количество правильно предсказанных отрицательных случаев.

Таблица 11 – Матрица неточностей (Confusion matrix)

Фактический класс	Предиктивный класс		
	ДА		НЕТ
	Да	True Positives (TP)	False Negatives (FN)
Нет	False Positives (FP)	True Negative s(TN)	

На основе применения этих четырех ячеек, можно привести одну из часто используемых мер оценки, т.е. общую формулу точности (7), которая представляет собой степень правильной классификации.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (7)$$

В случае несбалансированности данных, эффективней рассматривать следующие две меры одновременно, где:

1. TPR (True Positive Rate) – это отношение правильно классифицированных положительных экземпляров к истинно положительным экземплярам.

2. FPR (False Positive Rate) – это отношение неправильно классифицированных отрицательных экземпляров к истинно отрицательным экземплярам.

$$TPR = \frac{TP}{TP+FN} \quad (8)$$

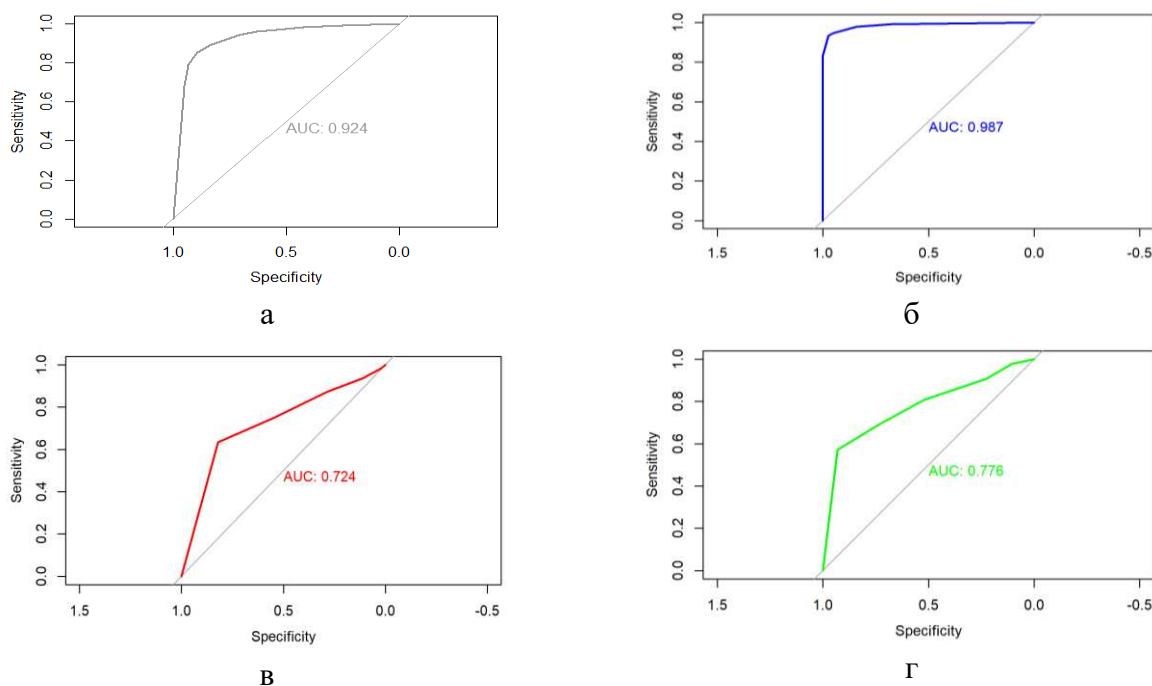
$$FPR = \frac{FP}{FP+TN} \quad (9)$$

Формула (8) подразумевает выгоды, полученные от задачи классификации, а формула (9) указывает затраты. График рабочих характеристик приемника (ROC), является полезным инструментом для визуализации и сравнения производительности классификатора на основе показателей TPR и FPR [130, p. 862].

В целом, ROC был разработан для случая двух классов, этот инструмент стал стандартным в вышеупомянутых сценариях оптимизации, но его применение можно также рассматривать в случае мультикласса [132]. За последнее время в этой области началась некоторая работа, такая как расширение случая трех классов (Mossman, 1999), исследование возможности расширения (Srinivasan, 1999) и исследование некоторых формализаций в (Ferri et al., 2003; Edwards et al., 2004) [132, p. 1747]. Так как, в построенной прогнозной модели дерева решений по оценке сейсмостойкости городского объекта, выявлены 4 класса объектов, таких как сейсмостойкий, несейсмостойкий, сейсmobезопасный и сейсмоопасный, то к анализу качества классификатора использовался подход мультиклассового ROC (AUC) анализа.

Количественную интерпретацию ROC-кривой дает показатель AUC (Area

Under Curve)-ROC [131, с. 110]. Она эквивалентна вероятности того, что классификатор присвоит большее значение случайно выбранному положительному объекту, чем случайно выбранному отрицательному объекту. Сама AUC оценка является агрегированной характеристикой качества классификации, не зависящей от соотношения цен ошибок [133]. В результате применения ROC (AUC) анализа, были получены следующие графики, представленные в соответствии с рисунком 18.



а – ROC (AUC) анализ для 1-го классификатора; б – ROC (AUC) анализ для 2-го классификатора; в – ROC (AUC) анализ для 3-го классификатора; г – ROC (AUC) анализ для 4-го классификатора

Рисунок 18 – ROC-анализ для оценки качества полученного классификатора прогноза оценки сейсмостойкости городских объектов

Исходя из полученных значений ROC (AUC), было рассчитано средневзвешенное значение $ROC (AUC) = (0.924+0.987+0.724+0.776)/4=0.852$. Тем самым, оценка качества классификатора показала 85,2% точности, что соответствует хорошему классификатору. Таким образом, подводя итоги данному этапу исследования, в результате применения метода DT, было получено дерево решений, которое представляет прогнозную классификационную модель для оценки сейсмостойкости городского объекта.

2.2.3 Выявление информативных признаков оценки сейсмостойкости городских объектов Ensemble-методом: Random Forest

На предыдущих этапах исследования в соответствии с разработанным способом (в соответствии с рисунками 7, 8) для решения задачи оценки сейсмостойкости городских объектов, в результате были получены кластеры

городских объектов, а также построена прогнозная классификационная модель дерева решений. Однако построение модели с помощью дерева решений не в полной мере показывает одни из главных характеристик, которые были использованы для построения дерева решений [122, р. 264]. В связи с этим, для устранения данной проблемы и выявления степени влияния различных характеристик объекта на сейсмостойкость, использовался другой метод классификации, как Random Forest (RF, Случайный лес) [122, р. 263]. Модель случайного леса (RF) представляет собой ансамблевый метод классификации и регрессии [134, 135]. RF состоит из множества деревьев, и решение, принятое случайной моделью леса, основано на ансамбле решений, принимаемых множеством деревьев решений [135, р. 71]. Он позволяет большому количеству слабо коррелированных классификаторов формировать сильный классификатор [122, р. 264]. RF – рекомендуемый алгоритм для стабильного выбора важных компонентов [136]. RF-алгоритм сочетает в себе идеи метода бэггинга (bagging, bootstrap aggregating) и метода случайных подпространств (RSM, random subspace method) [137-139]. Описание алгоритма построения случайного леса (Брейман, 2001) можно представить следующим образом. Пусть обучающая выборка состоит из N объектов, размерность пространства признаков равна M , и задан параметр m – количество признаков, из которых происходит выбор признаков для разбиения в узлах деревьев. Все деревья строятся независимо друг от друга по следующему шагу [122, р. 263]:

- генерируется случайная подвыборка с повторением (т.е. некоторые объекты попадут в неё несколько раз) того же размера, что и обучающая выборка (т.е. размерностью N), она также называется бутстреп выборкой;

- строится решающее дерево, классифицирующее объекты данной подвыборки, причём в ходе создания очередного узла дерева признак, на основе которого производится разбиение, выбирается не из всех M признаков, а лишь из m случайно выбранных (выбор наилучшего из этих m признаков может осуществляться с помощью индекса Gini);

- дерево строится до полного использования подвыборки и не подвергается процедуре отсечения [122, р. 263].

В отличие от классических алгоритмов построения деревьев решений [16, р. 157] в методе случайных лесов при построении каждого дерева на стадиях расщепления вершин используется только фиксированное число случайно отбираемых признаков обучающей выборки (второй параметр метода) и строится полное дерево (без усечения), т.е. каждый лист дерева содержит наблюдения только одного класса. Классификация осуществляется с помощью голосования классификаторов, определяемых отдельными деревьями, а оценка регрессии – усреднением оценок регрессии всех деревьев [122, р. 263; 140].

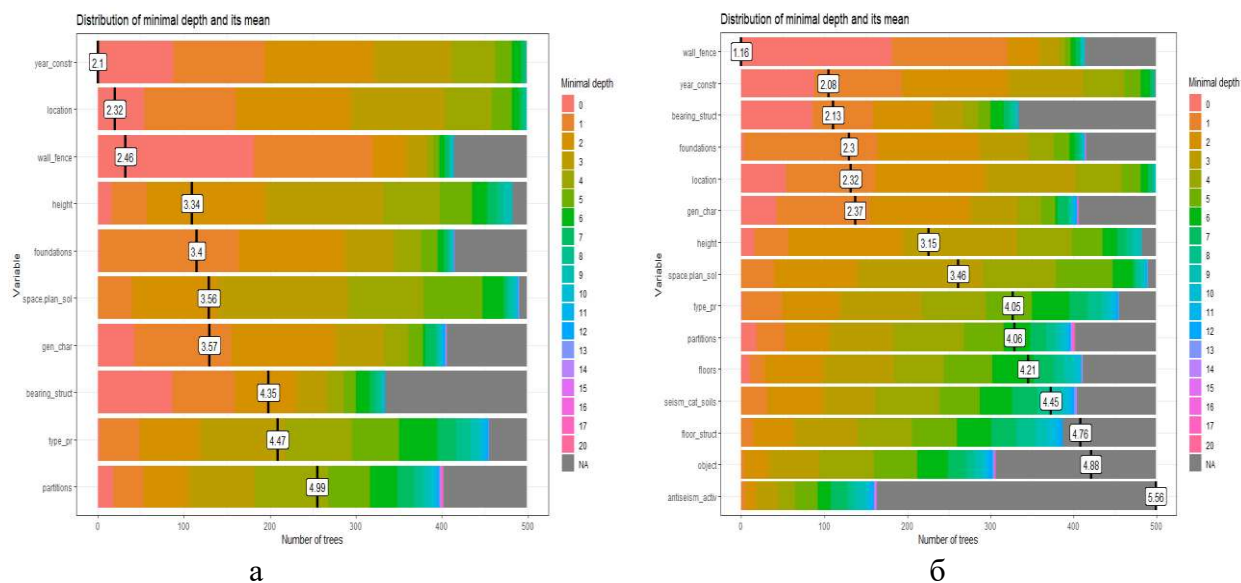
Случайный лес, дает много преимуществ для его применения [122, р. 263]:

1. Он эффективно работает с большими базами данных.
2. Он может обрабатывать большое количество входных переменных без удаления переменных.

3. Он дает оценку важности переменным, участвующие в классификации.
4. Он интерпретирует внутреннюю оценку и объективно отражает ошибки обобщения (oob error).
5. Он вычисляет близость между парами наблюдений, которые можно использовать для поиска выбросов.
6. Он относительно устойчив к выбросам и шуму.
7. Он вычислительно легче, чем другие методы ансамбля деревьев (например, Boosting) [141]. В рассматриваемой задаче тип модели – regression, number of trees – 500, no. of variables tried at each split – 5, mean of squared residuals – 0.08. Значения распределения минимальной глубины представлены в таблице 12 [122, p. 264]. Распределение минимальной глубины для первых десяти переменных в соответствии со средней минимальной глубиной, рассчитанной с использованием верхних деревьев показаны в соответствии с рисунком 19а [122, p. 264].

Таблица 12 – Распределение минимальной глубины

tree	variable	minimal_depth
1	bearing_struct	0
1	floors	2
1	foundations	1
1	gen_char	2
1	height	2
1	location	3
1	partitions	3
1	seism_cat_soils	9
1	space.plan_sol	1
1	type_pr	4



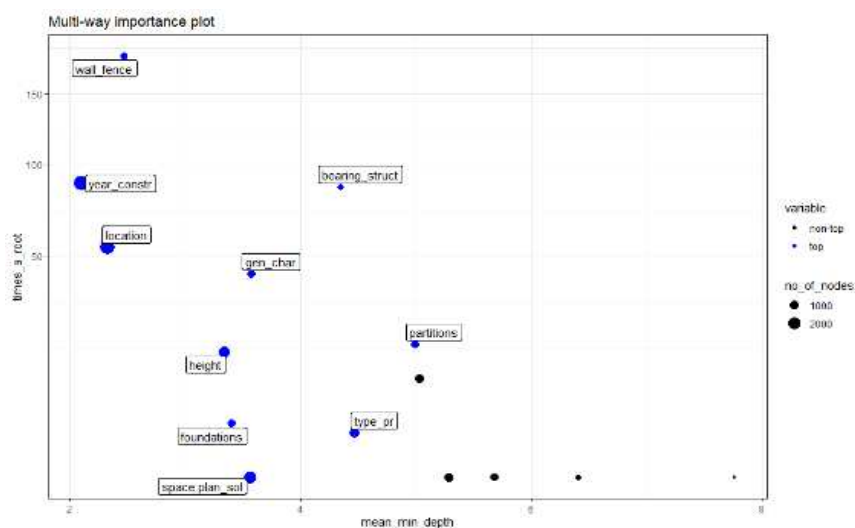
а – top_trees; б – relevant_trees

Рисунок 19 – Вычисление средней минимальной глубины

Для лесов со многими переменными с большим количеством пропущенных наблюдений всегда должны рассмотреть возможность добавления опции `min_no_of_trees`, чтобы для графика учитывались только те переменные, которые использовались для расщепления хотя бы в объявленном количестве деревьев [122, р. 264]. Это позволяет нам избежать выбора переменных, которые были случайно использованы для расщепления (например, только один раз, но в корне, их среднее значение будет равно 0) [122, р. 264]. Однако в нашем случае мы можем просто увеличить параметр `k`, чтобы построить все деревья, в соответствии с рисунком 19б [122, р. 264]. Использование только релевантных деревьев для вычисления среднего не меняет его для переменных, у которых нет пропущенных значений. Кроме того, в этом случае изменение не влияет на порядок переменных [122, р. 264].

В следующем этапе исследования изучили показатели важности переменной, рассчитав меры `accuracy_decrease (classification)`, `gini_decrease (classification)`, `mse_increase (regression)`, `node_purity_increase (regression)`, которые извлекли их из нашего объекта леса [122, р. 264].

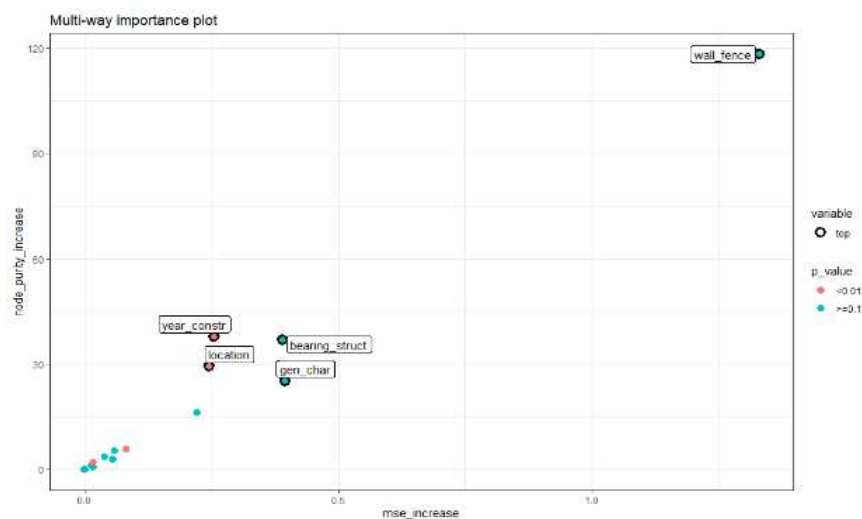
Меры `accuracy_decrease (classification)` и `mse_increase (regression)` основаны на снижении предсказательной точности леса. Меры `gini_decrease (classification)` и `node_purity_increase (regression)` основаны на изменениях в чистоте узла после разделения на переменную. Меры `mean_minimal_depth`, `no_of_trees`, `no_of_nodes`, `times_a_root`, `p_value` основаны на структуре леса [122, р. 264]. Результат извлечения важных переменных, представлен ниже в соответствии с рисунком 20 [122, р. 264].



a

a – multi-way importance plot (10 верхних переменных)

Рисунок 20 – Извлечение важных переменных, лист 1

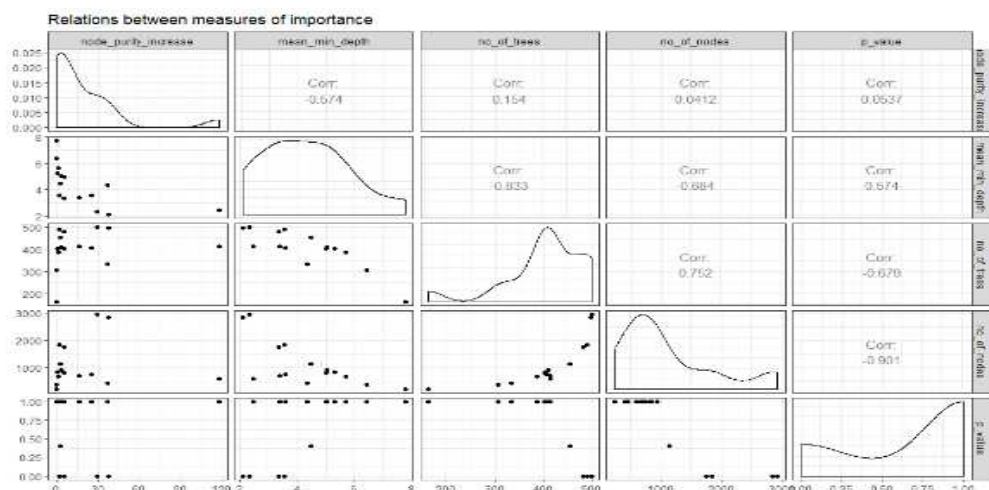


б

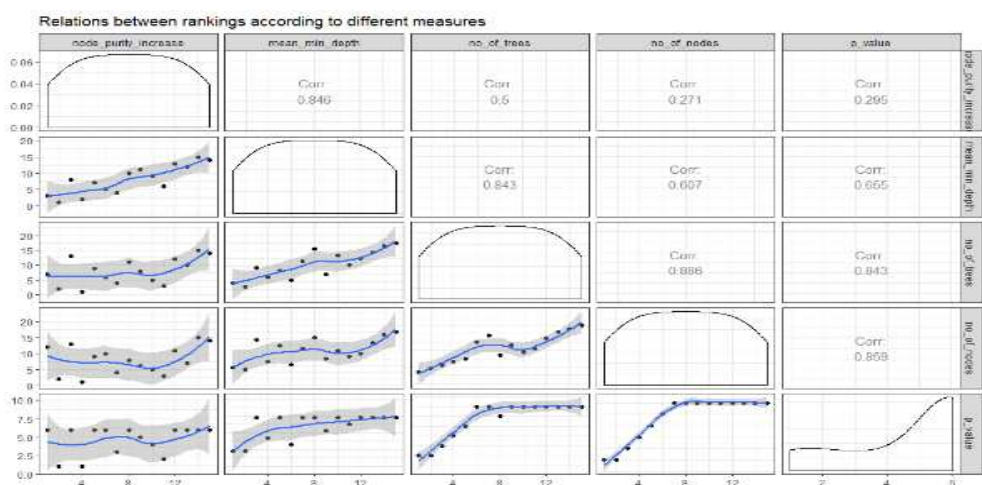
б – multi-way importance plot (5 верхних переменных)

Рисунок 20, лист 2

Как видно в соответствии с рисунком 20а, по умолчанию 10 верхних переменных на графике выделены синим цветом и помечены, они выбираются с помощью использования суммы рангов, основанных на показателях важности. Превосходство wall_fence (стенное ограждение), bearing_struct (конструкции перекрытий), year_constr (год постройки здания), location (местоположение), gen_char (общая характеристика) очевидно во всех трех измерениях [122, р. 264]. Далее представлен график важности характеристик для нескольких направлений другого набора показателей важности: увеличение среднеквадратической ошибки. Здесь, также установили пометку переменных на пять, чтобы только пять верхних переменных были выделены в соответствии с рисунком 20б [122, р. 264]. На обоих графиках wall_fence (стенное ограждение), bearing_struct (конструкции перекрытий) показывают структуру леса на важность переменных [122, р. 264]. Результаты сравнения мер в соответствии с рисунком 21а, 21б [122, р. 264], предлагает множество вариантов подбора переменной важности, поэтому это не дает нам возможности выбора наиболее информативного графика для анализа [122, р. 264]. В связи с этим, произведено изучение отношения между различными мерами важности, а затем выбрано три, которые наименее согласуются друг с другом, и использовали их на графике разносторонней важности для выбора верхних переменных в соответствии с рисунком 21а, 21б [122, р. 264].



а



б

а – Compare measures using ggpairs; б – Compare different rankings

Рисунок 21 – Сравнение мер

Сравнивая ранжирование на приведенном выше графике, видно, что две пары мер почти точно совпадают в своих ранжировках переменных: mean_min_depth против mse_increase и mse_increase против node_purity_increase [122, р. 265]. После выбора набора наиболее важных переменных, исследовали взаимодействия по отношению к ним, то есть расщепления, появляющиеся в максимальных поддеревьях по отношению к одной из выбранных переменных (оценка). Тем самым, извлекли имена 5 наиболее важных переменных в соответствии со средней минимальной глубиной и количеством деревьев: «location» (местоположение), «year_constr» (год постройки здания), «height» (высота), «space.plan_sol» (сейсмичность района), «wall_fence» (стенное ограждение). На следующем этапе построили график в соответствии с рисунком 22 [122, р. 265], содержащий информацию о средней условной минимальной глубине переменных по каждому элементу переменных [122, р. 265].

методов машинного обучения для решения не простой задачи оценки сейсмостойкости городских объектов. Как уже было отмечено, быстрое развитие интеллектуального анализа данных в настоящее время дает возможность ускорить процесс оценки сейсмической уязвимости в доступных условиях. Примерами таких подходов и решений, показывающие зависимости между уязвимостью зданий и диапазоном их характеристик методами интеллектуального анализа данных, были приведены в обзоре диссертационной работы, что в свою очередь еще больше подтверждает актуальность данной работы. В результате проведенного исследования с помощью методов машинного обучения на основе интеллектуального анализа данных были выявлены основные характеристики оценки сейсмостойкости типовых объектов городской среды. В ходе исследования, применение методов кластеризации K-means и НК-means, позволили получить кластеры ($k=4$) со сходными переменными объектов и определить объекты к определенной группе. При кластерном анализе данных также были выявлены элементы, которые не были соотнесены к кластерам, что требовало дополнительных исследований и применения других методов, таких как DT и RF. Таким образом, в результате реализации поставленной цели и задач диссертационной работы разработан способ оценки сейсмостойкости городских объектов, отличие которой заключается в применении кластерного анализа и алгоритмов классификации, обладающие высокой прогнозной точностью и стабильностью (где деревья решений показали 93,6% точности, а наиболее высокую точность показал метод RF – 100%).

Выводы по второму разделу:

1. Разработан способ оценки сейсмостойкости городских объектов, основанный на применении интеллектуального анализа данных, который позволяет осуществить «быструю» оценку сейсмостойкости городских типовых объектов, что значительно сокращает время и затраты для проведения оценки сейсмической уязвимости зданий. Разработанный способ включает в себя кластерный анализ и алгоритмы ансамблевой классификации, обладающие высокой прогнозной точностью и стабильностью.

2. Результат применения кластерного анализа методом K-Means позволил определить 4 кластера городских объектов. При кластерном анализе набора данных городских объектов в качестве метода оценки оптимального числа кластеров использовался «метод локтя», который определил оптимальное количество кластеров $k=4$. Для подтверждения результатов метода K-Means, использовался другой метод кластерного анализа НК-Means. В интерпретации обоих методов для объединения объектов в кластеры при вычислении расстояния между точками (объектами), использовалась мера близости Евклидово расстояние. Сравнительный анализ обоих методов кластерного анализа данных, показал идентичные результаты, что доказало применимость кластерного решения в разбиении на группы объектов городской среды по определенным критериям. Также важно отметить, что результаты полученных

кластеров совпали с результатами экспертной оценки.

3. В исследовании особенностей полученных кластеров и принятия решений о сейсмостойкости объекта с помощью метода Decision Tree была построена классификационная модель дерева решений. Построенные для каждого кластера собственные деревья решений позволили уточнить границы полученных решений. В оценке эффективности классификатора и для обеспечения достоверности полученных результатов были проведены оценки точностей построенных моделей, где построенная модель дерева решений показала высокую точность 93,6% (на основе метрики производительности классификации Confusion Matrix) и 85,2% (метод оценки качества классификации на основе мультиклассового ROC(AUC)-анализа).

4. Для улучшения работы и точности построенной классификационной модели дерева решений использован ансамблевый метод Random Forest (RF). С помощью метода RF проведена оценка важности переменных, которые влияют на окончательный результат в оценке сейсмостойкости городского объекта. Оценивание значимости отдельных признаков в классификационной модели дерева решений, выявило следующие характеристики: месторасположение объекта и год постройки на основании которой определяется типовой проект объекта. Также не менее важны объемно-планировочные решения, которые влияют на общие характеристики объекта. Кроме этого, исходя из общих характеристик объекта, RF также определил переменную wall_fence (стенное ограждение) как наиболее значимую. ROC(AUC)-анализ оценки качества классификатора полученного методом Random Forest показал высокую точность в 100%.

5. Полученные результаты исследований доложены и опубликованы в сборнике международной научно-практической конференции [16, p. 154-159], материалы которых индексируются в базе данных Scopus. Кроме этого, опубликована статья [122, p. 261-266] в международном журнале (в базе данных Scopus показатель процентиля по CiteScore равный 34%) и опубликована статья [124, p. 102-109] в издании, рекомендованного Комитетом по контролю в сфере образования и науки МОН РК.

3 МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА ПРОСТРАНСТВЕННЫХ ДАННЫХ НА ПРИМЕРЕ ИССЛЕДОВАНИЯ СЕЙСМОАКТИВНОСТИ РЕСПУБЛИКИ КАЗАХСТАН

Интеллектуальный анализ данных и разнообразие инструментария обработки данных представляют исследователю множество возможностей проведения качественного анализа различных данных. Построенные на основе таких технологий вычислительные алгоритмы дают возможность извлечения знаний из экспериментальных данных, а также преобразования и моделирования данных с целью извлечения полезной информации и принятия решений. Следует также отметить, что отечественные научные исследования не в достаточной мере используют технологии Data mining и методы машинного обучения для обработки сейсмических данных, которые регистрируются станциями на территории РК ежедневно. Так, с помощью инструментария технологии анализа данных и алгоритма машинного обучения, были проанализированы данные зафиксированных сейсмических событий. Реализация основной задачи оценки сейсмостойкости городских объектов также учитывает и исследование набора сейсмических данных, где анализ данных сейсмических событий рассматривает применение более эффективных подходов для обнаружения и кластеризации сейсмических данных на примере исследования сейсмоактивности Республики Казахстан (РК). Исследование набора данных о землетрясениях в диссертационной работе проводилось в два подэтапа: 1) статистический анализ сейсмических данных; 2) пространственный анализ сейсмических данных алгоритмом кластеризации на основе плотности DBSCAN.

3.1 Статистический анализ каталога землетрясений, зарегистрированных на территории Республики Казахстан

Вопросы современного строительства в условиях сейсмического риска по сей день остаются актуальными. В подобного рода исследованиях исходной и наиболее важной, необходимой информацией для изучения, являются данные о самих землетрясениях. Детальное изучение данных о землетрясениях представляют большой интерес не только для сейсмологов, но и для специалистов в области сейсмостойкого строительства [18, с. 149].

В своих исследованиях специалист по сейсмостойкому строительству должен подходить к изучению землетрясений с иных позиций, чем сейсмолог [142]. Сейсмологи сосредоточивают свое внимание прежде всего на глобальных или макроскопических последствиях землетрясений и поэтому имеют дело с небольшими амплитудами колебаний грунта, не вызывающими ощутимой реакции сооружений [142, с. 152]. Инженеры, наоборот, имеют дело в основном с местными эффектами сильных землетрясений там, где колебания грунта достаточно интенсивны, чтобы вызвать повреждения конструкций [142, с. 152]. Тем не менее, несмотря на различие сферы, в соответствии с рисунком 23 деятельности инженеров в области сейсмостойкого строительства и

сейсмологов, многие вопросы сейсмологии представляют для инженера большой интерес [142, с. 152].

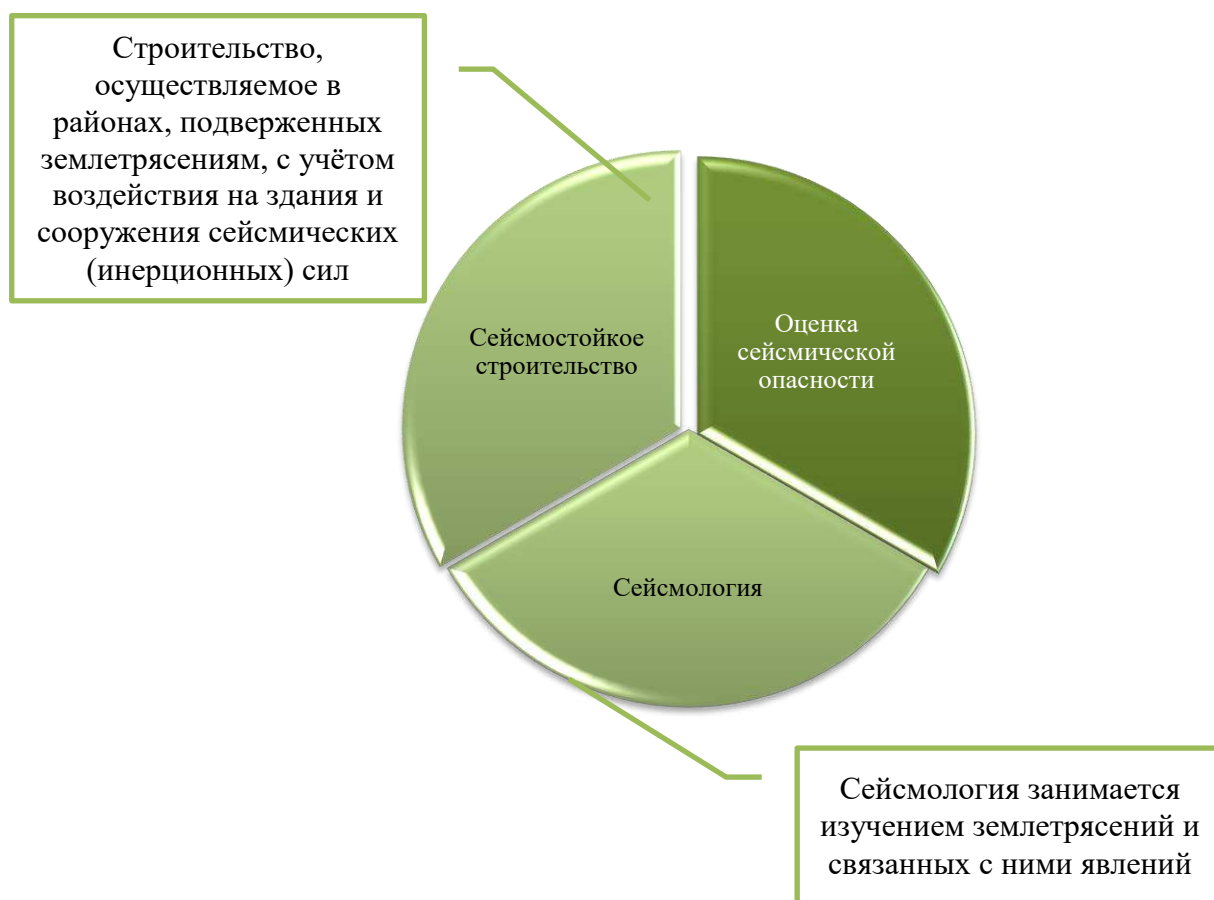


Рисунок 23 – Сферы деятельности науки сейсмологии и направления сейсмостойкого строительства

Примечание – Составлено по источнику [143, 144]

В целом, стратегия сейсмостойкого строительства основывается на информации, связанной с потенциальной сейсмической опасностью. Интенсивный рост городов и высотное строительство в начале XX века предъявляли все большие требования к качеству определения потенциальной сейсмической опасности [144, с. 22]. Оценка сейсмической опасности, которая носит прогнозный характер, зависит от нескольких параметров очага землетрясения. Это географические координаты эпицентра, глубина очага, магнитуда и функция затухания интенсивности сейсмического эффекта с расстоянием [144, с. 53]. Для оценки сейсмической опасности необходимо собрать исходную информацию об исторических сейсмических событиях и за период инструментальных наблюдений в форме каталогов землетрясений [144, с. 60-61].

В анализе сейсмической опасности качество процесса анализа определяется одним из значимых факторов, которым является модель сейсмичности зоны очагов землетрясений, где сейсмичность характеризуется:

– частотой возникновения землетрясений;

- статистическим распределением силы толчков (магнитуды);
- пространственным распределением очагов;
- макросейсмическими наблюдениями сильных сейсмических событий (интенсивность сейсмических воздействий, картина нанесенного ущерба) [144, с. 96].

Определение названных характеристик основывается на сейсмостатистике, в виде каталогов землетрясений сейсмических зон. Каталоги землетрясений являются важной продукцией сейсмологии, и их составление связано с обработкой большого объема геолого-геофизических и сейсмических данных [144, с. 300]. Землетрясение можно охарактеризовать множеством параметров, однако для анализа сейсмической опасности достаточно знать: время возникновения; координаты эпицентров; глубину очага и магнитуду землетрясений [144, с. 98]. Как известно, территория юга и юго-востока Казахстана входит в систему сейсмоактивных поясов Евразии и характеризуется интенсивными геодинамическими процессами, одним из проявлений которых являются сейсмические события [145]. Кроме того, в прошлом были факты о ряде произошедших сильнейших и катастрофических землетрясений [18, с. 149], два из которых (Чиликское – 1889 г. и Кеминское – 1911 г.) имеют магнитуду более 8 [145, с. 7]. Уровень сейсмической активности еще остается высоким и в настоящее время [145, с. 7]. В исследованиях, связанных с сейсмической опасностью в целях предотвращения больших потерь от землетрясений исследуются накопленные исторические данные о них и обновляются сейсмические карты рисков [18, с. 149]. В связи с этим возникает необходимость тщательного исследования и анализа пространственных данных о землетрясениях, зафиксированные в РК [18, с. 149-150].

Статистический анализ сейсмических данных проводился с помощью программной среды R+Rstudio. На данном подэтапе исследования использовался набор данных из 1905 записей (в предыдущих экспериментальных исследованиях [18, с. 150] количество зафиксированных сейсмических событий в наборе данных содержал 1742 записей), который содержит следующую структуру: дата (день/месяц/год) происшествия, точное время по Гринвичу (час/минута/секунда), широта, долгота, глубина, энергетический класс и магнитуда (Data, GreenwichMeanTime, Latitude, Longitude, Depth (km), Energy class and Magnitude) [18, с. 150; 146]. Сейсмические данные получены из источника ТОО «СОМЭ» АО «Национальный центр сейсмологических наблюдений и исследований», станциями которых проводятся сейсмические наблюдения (регистрация землетрясений) на территории РК [147]. Сейсмический набор данных включает события о зафиксированных землетрясениях на территории РК с 2012 года по 2021 год включительно. В таблице 13 [18, с. 151; 146, с. 38] представлен фрагмент набора данных.

На данном подэтапе статистического анализа набора сейсмических данных проведена предварительная обработка данных. Факторные переменные,

такие как, дата и время, были преобразованы в соответствующие форматы date и POSIXct. Также, переменные широта, долгота, глубина, энергетический класс, магнитуда преобразованы в числовые форматы [18, с. 150].

Таблица 13 – Фрагмент сейсмического набора данных

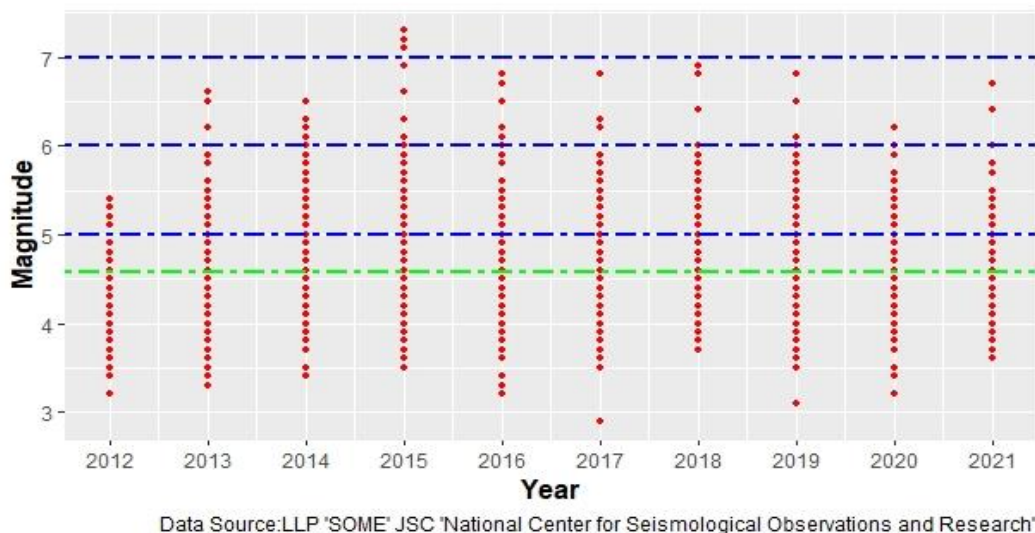
Дата	Время по Гринвичу	Широта, с.ш.	Долгота, в.д.	Глубина, км	Энерг. класс, Кр	Магнитуда MPVA
14.10.2012	16:30:00	40.88	74.17	15	10.8	4.7
12.10.2012	18:08:00	40.82	76.97	0	9.8	4.2
12.10.2012	7:18:00	43.05	78.22	20	9	4.4
11.10.2012	10:58:00	44.13	78.95	20	8.1	3.6

При статистическом анализе и исследовании набора данных о землетрясениях, были изучены количественные характеристики сейсмических событий по частоте их возникновения (по времени, месяцу, году), а также определены классы магнитуд. Сама величина магнитуды (M, от латинского *magnitudo* - величина), является безразмерной величиной, которая помогает классифицировать события по величине их очага. Иными словами, магнитуда землетрясения – это количественная характеристика величины очага землетрясения [18, с. 151]. «В каталоге ISC-GEM, в каталоге Центральной Азии, как и в каталоге Казахстана можно увидеть целый набор энергетических и магнитудных характеристик землетрясений, для определения которых используются различные шкалы» [18, с. 151; 148]. MPVA – региональная магнитуда по объемным волнам, определяемая по Р-волне, зарегистрированной короткопериодными приборами. Определяется по значениям амплитуды и периода смещений с помощью региональной калибровочной кривой для Северного Тянь-Шаня [149] для расстояний до 1000 км. Используется в практике работ СОМЭ МОН РК и Центра данных ИГИ [18, с. 151]. Статистический анализ набора сейсмических данных на основе инструментов технологии анализа данных по количественной характеристике величины очага землетрясения, выявил следующие классы магнитуд [150] и их численные показатели. Данные классы магнитуд с численными показателями отображены в таблице 14 [18, с. 152]. Таким образом, опираясь на результаты статистического анализа данных, выявлено, что самое наибольшее количество зарегистрированных сейсмических событий имеют классы магнитуд 4-5, а также 3-4 и 5-6.

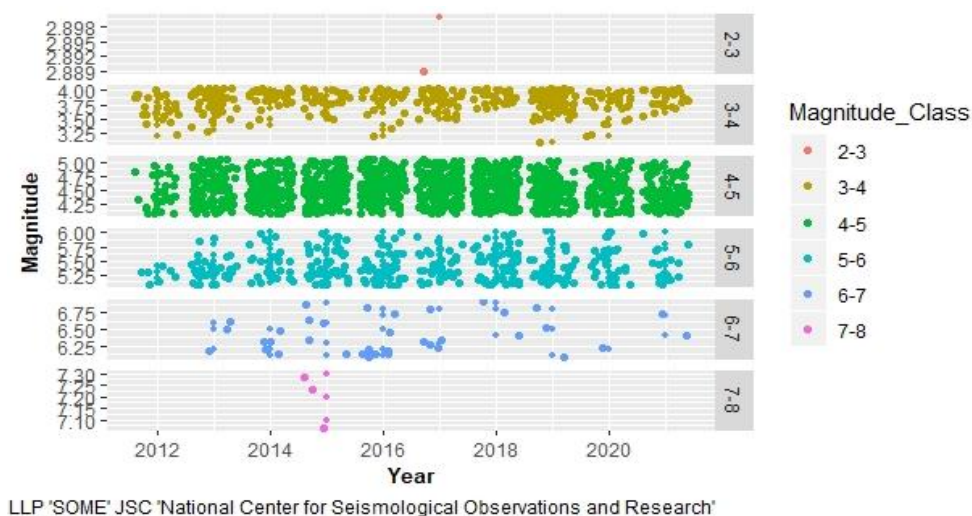
Таблица 14 – Классы магнитуд

Magnitude Class	n
2-3	1
3-4	343
4-5	1195
5-6	326
6-7	37
7-8	3

В продолжении исследования по статистическому анализу данных, зарегистрированных сейсмических событий, были изучены распределения данных по временному интервалу. Полученный график плотности в соответствии с рисунком 24а, 24б визуализирует распределение данных за непрерывный интервал с 2012 года по 2021 год.



а



б

а – распределение магнитуд сейсмических данных по годам; б – распределение классов магнитуд сейсмических данных по годам

Рисунок 24 – Графики визуализации плотности распределения данных сейсмических событий по годам

Более подробная интерпретация графиков плотности распределения землетрясений по годам представлена также в таблице 15 [18, с. 153]. Из графиков и табличной информации было определено, что >100 сейсмических событий с классом магнитуд 4-5, зафиксированы в 2013-2020 годах, и 3 землетрясения с классом магнитуд 7-8 в 2015 году. Сейсмические события с

классом магнитуд 6-7, имеют частоту возникновения от 1 до 9 раз, кроме 2012 года. Классы магнитуд 3-4, 5-6, определены во всех интервалах времени в наборе данных, тем самым показывая возникновение землетрясений по всем годам.

Таблица 15 – Распределение класса магнитуд по годам

Year	Magnitude Class					
	2-3	3-4	4-5	5-6	6-7	7-8
2012		25	27	7		
2013		44	121	32	3	
2014		36	121	27	5	
2015		29	134	48	6	3
2016		30	167	43	9	
2017	1	37	149	37	5	
2018		26	165	52	3	
2019		59	112	36	3	
2020		31	120	28	1	
2021		26	79	16	2	

Также, с помощью инструмента технологии анализа данных, были продолжены исследования зафиксированных сейсмических событий во временной шкале по месяцам, который представлен в соответствии с рисунком 25.

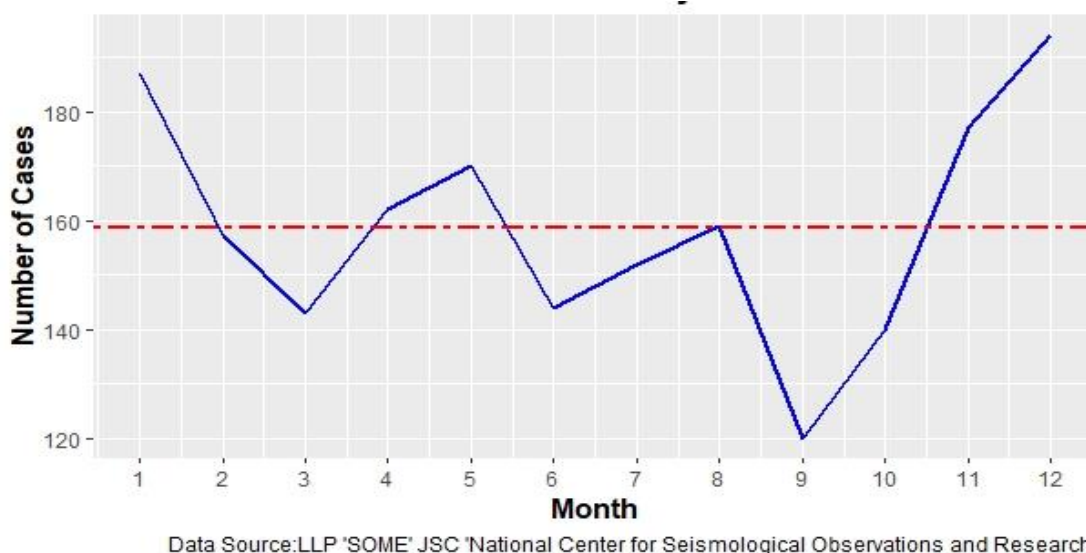
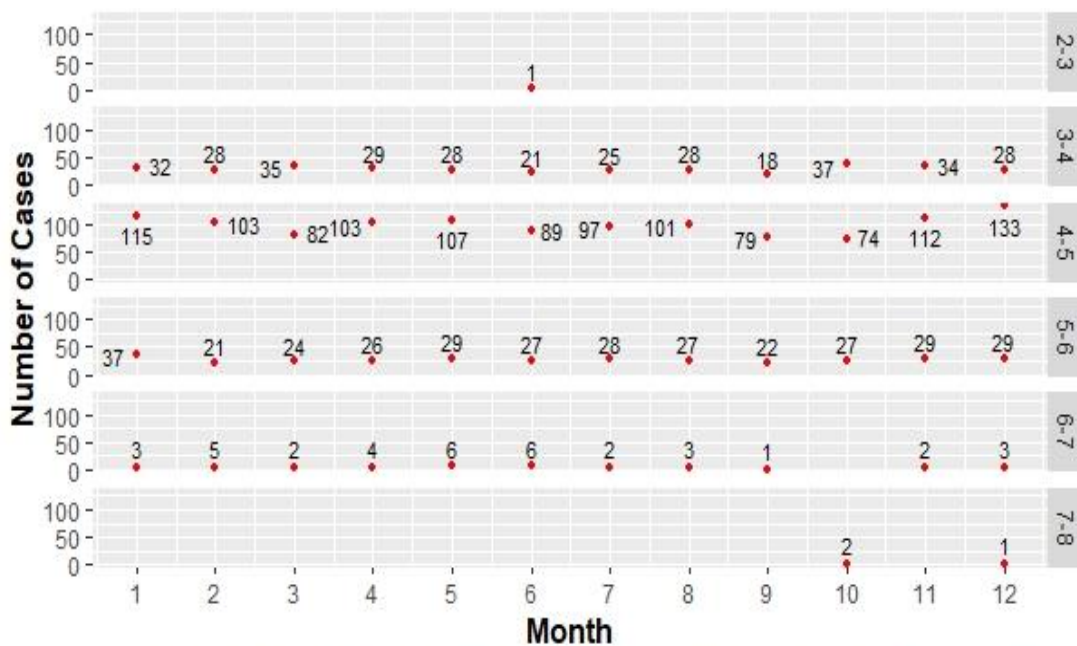


Рисунок 25 – График зафиксированных сейсмических событий во временной шкале по месяцам

Следовательно, статистический анализ сейсмического набора данных показывает, что возникновение сейсмических событий происходило в каждом месяце с определенным классом магнитуд, который также представлен в соответствии с рисунком 26.



Data Source: LLP 'SOME' JSC 'National Center for Seismological Observations and Research'

Рисунок 26 – График распределения зафиксированных сейсмических событий по месяцам и по классам магнитуд

Количество сейсмических событий по месяцам отображено в таблице 16, где условное разбиение ≥ 150 событий приходится на 1, 2, 4, 5, 7, 8, 11, 12 месяцы, среди которых наибольшее количество произошедших землетрясений приходится на месяцы зимнего периода, а также количество в интервале ≤ 150 сейсмических событий показывают 3, 6, 9 месяцы.

Таблица 16 – Количество сейсмических событий по месяцам

Month	1	2	3	4	5	6	7	8	9	10	11	12
n	187	157	143	162	170	144	152	159	120	140	177	194

Аналогичное исследование анализа данных проведено по времени возникновения зафиксированных сейсмических событий за сутки. Графическое представление и визуализация статистического исследования по времени возникновения землетрясений за сутки с классами магнитуд представлены в соответствии с рисунками 27, 28, а также в таблице 17 отражены их количественные характеристики.

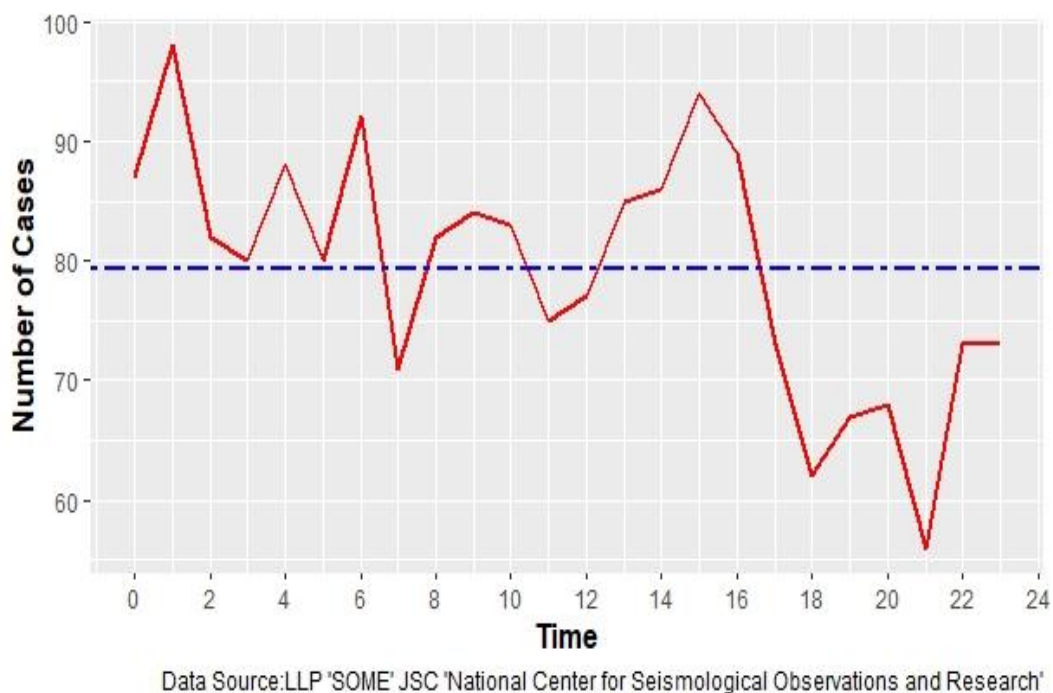


Рисунок 27 – Визуализация количества зафиксированных сейсмических событий по времени за сутки

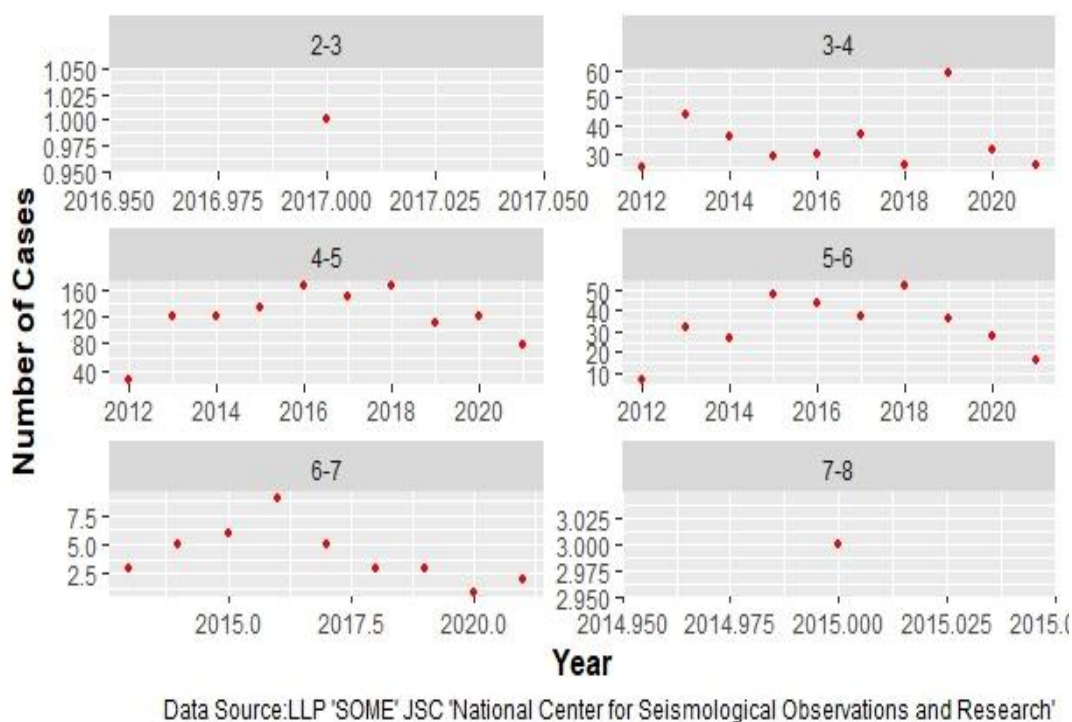


Рисунок 28 – График распределения зафиксированных сейсмических событий по времени и по классам магнитуд

Таблица 17 – Количественная характеристика распределения класса магнитуд во временной шкале за сутки

Hour	Magnitude Class					
	2-3	3-4	4-5	5-6	6-7	7-8
0		19	50	17	1	
1		21	63	13	1	
2		17	52	12	1	
3		15	56	7	2	
4		21	51	14	2	
5		9	55	15	1	
6		18	63	9	2	
7		10	44	15	2	
8		15	47	18	1	
9		17	44	18	4	1
10		11	51	18	3	
11		17	48	6	4	
12		13	47	16	1	
13		13	51	21		
14		21	47	14	4	
15		19	62	12	1	
16		16	60	11	2	
17	1	10	47	14	1	
18		6	49	7		
19		5	47	14		1
20		11	43	13		
21		7	30	19		
22		20	39	13	1	
23		12	49	10	2	

Как показывает первый подэтап статистического анализа набора сейсмических данных, частота возникновения землетрясений приходится на каждый год, на каждый месяц, на каждый час времени за сутки. В большей степени возникновение землетрясений происходит в зимний период времени. Важно отметить, что среди первых сведений о произошедшем сейсмическом событии, а именно о землетрясении, становится известной его магнитуда. Магнитуда землетрясения не является показателем интенсивности землетрясения, ощущаемой в конкретной точке на поверхности земли. Интенсивность проявления землетрясения в конкретном районе на поверхности зависит от магнитуды землетрясения, глубины гипоцентра, типа горных пород и расстояния до очага. В связи с этим, с помощью статистического анализа набора данных, были определены классы магнитуд. В дополнение к данному статистическому анализу набора данных, также были изучены распределения землетрясений по классам магнитуд на интерактивной карте, полученной с помощью инструмента технологии анализа данных. С помощью интерактивной карты проанализированы очаги возникновения землетрясений на основе их географических координат, т.е. проведен визуальный анализ и определены регионы. Визуальный анализ распространения землетрясений по классу

магнитуды в 7-8, не выявил их принадлежность к регионам РК. Визуальный анализ пространственных данных распространения землетрясений на интерактивной карте по классу магнитуды равной 6-7 показывает, что количество эпицентров землетрясений принадлежащих территории РК составляет 4 точки, 2 из которых лежат на границе с Республикой Кыргызстан. Географическая область принадлежности данных эпицентров – это Восточно-Казахстанская область, Алматинская область и г. Алматы. Расположение пространственных точек эпицентров землетрясений отображено в соответствии с рисунком 29, которые обозначены эллипсами.



Рисунок 29 – Расположение пространственных точек эпицентров землетрясений на территории РК классом магнитуд 6-7

Визуальный анализ распространения землетрясений по классу магнитуды равной 5-6, также на интерактивной карте показывает эпицентры, принадлежащие территории РК. Такими регионами, являются Восточно-Казахстанская область, Алматинская область, г. Алматы, г. Талдыкорган, Жамбыльская область. Расположение пространственных точек эпицентров землетрясений отображено в соответствии с рисунком 30.

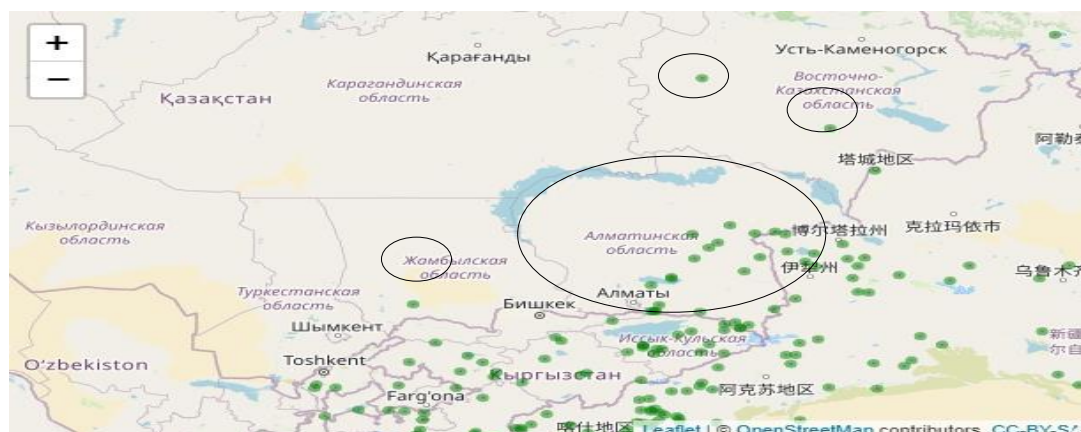


Рисунок 30 – Расположение пространственных точек эпицентров землетрясений на территории РК классом магнитуд 5-6

Полученные результаты статистического анализа данных сейсмических событий проводились на атрибутах землетрясений, дата и время, долгота, широта, магнитуда. Визуализация данных распространения эпицентров землетрясений на интерактивной карте, также помогло выявить географические принадлежности точек возникновения землетрясений. Следующий подэтап исследования каталога землетрясений проводился алгоритмом DBSCAN для получения кластеров землетрясений.

3.2 Оценка сейсмической опасности алгоритмом пространственной кластеризации на основе плотности с присутствием шума DBSCAN

Ранее было отмечено, что исследование набора данных о землетрясениях было проведено в два подэтапа. В первом подэтапе исследования каталога землетрясений проводился статистический анализ набора данных. Результаты статистического анализа данных показали количественные показатели по различным временным атрибутам, помогли определить классы магнитуд и их распределение на интерактивной карте. Однако, статистический анализ не в полной мере дает представление полного исследования сейсмического набора данных. В связи с этим, во втором подэтапе исследования набора данных землетрясений используется алгоритм DBSCAN для обнаружения участков с повышенной пространственной плотностью сейсмических событий.

Алгоритмы кластеризации на основе плотности определяют кластеры точек, расположенные в непосредственной близости, на основе заданного расстояния поиска. Особенностью алгоритмов плотностной кластеризации является то, что кластеры могут иметь произвольную форму. Наиболее популярным алгоритмом кластеризации на основе плотности является алгоритм DBSCAN [151]. Достоинства алгоритма заключаются в том, что он не требует спецификации числа кластеров в данных априори и устойчив к выбросам. Как известно, управление алгоритмом осуществляется на основе двух параметров: окрестность заданного радиуса (Eps) и минимальное количество объектов ($MinPts$). Алгоритм DBSCAN идентифицирует кластеры таким образом, что обнаруживает все основные точки и расширяет каждую до всех точек, достижимых по плотности. В случае, если в расширенном окружении больше не обнаружено узловых точек, то кластер завершен. Далее оставшиеся точки продолжают проверяться, чтобы увидеть, можно ли найти другую центральную точку для запуска нового кластера. После обработки всех точек, неосновные точки, которые не были отнесены к кластеру, считаются шумом. Касательно этого исследования точки данных, в первую очередь представляют точки зафиксированных сейсмических событий, а если быть точнее землетрясений. В связи с этим, некоторые обнаруженные алгоритмом изолированные точки, будут не достижимы по плотности из основной или узловой точек, тем самым расположены в регионах с низкой плотностью и не будут назначены ни одному кластеру. Следует обратить внимание, что эти точки также могут представлять по историческим данным некоторые участки сейсмических зон. Такие закономерности могут быть связаны с местными

геологическими рельефами, структурой пород земной коры и ее разломами.

Ранее уже было отмечено, что алгоритм DBSCAN может находить кластеры произвольной формы, тем самым обнаруживать взаимосвязи в различных сейсмических данных с точки зрения плотности распределения выборки [89, p.173; 152].

В алгоритме DBSCAN используется матрица схожести и как уже ранее отмечалось, реализация алгоритма осуществляется на основе окрестности заданного радиуса (Eps) и минимального количества объектов ($MinPts$, далее M). Также, алгоритм DBSCAN определяет число кластеров K в процессе работы. Более детально рассмотрим алгоритм: Eps – окрестность объекта X , формула (10).

$$D(X, Eps) = \{Y \in S: P(X, Y) \leq Eps\} \quad (10)$$

ядровой или базовой объект степени M при заданном значении Eps , будет таким объектом, где Eps которого содержит не менее M других объектов. В случае когда значение M заранее задано, то объект Y непосредственно плотно достижим из объекта X , если $Y \in S(X, Eps)$, а объект X является базовым. Также если, объект Y плотно достижим из объекта X , и если существуют такие объекты X_1, \dots, X_n , где $X_1=X$, $X_n=Y$, тогда объект X_{i+1} непосредственно плотно достижим из X_i при всех $i=1, \dots, n-1$.

Описание алгоритма DBSCAN выглядит следующим образом.

Шаг 1. Задаются значения параметров Eps и M (предположим, что $K:=0$).

Шаг 2. Если все объекты $X \in S$ уже были просмотрены, то остановка. Иначе выбирается любой из них и отмечается как просмотренный.

Шаг 3. Если X – это базовый объект, то создается новый кластер, полагая, что значение будет $K:=K+1$ и далее переходим к Шагу 4. В обратном случае точка X , помечается как «шум». Следует отметить, что в процессе выполнения данная точка вероятно может оказаться в Eps -окрестности какой-либо другой точки, тем самым может быть включена в другой кластер. Таким образом, далее переход осуществляется к Шагу 2.

Шаг 4. В полученный кластер будут включаться все объекты, которые являются плотно достижимыми из базового объекта X , далее происходит переход к Шагу 2.

Таким образом, алгоритм DBSCAN может объединить две отдаленные друг от друга точки, которые образуют кластер произвольной формы. Эти кластеры могут находиться недалеко от сейсмических разломов и относительно их окружающих участков. Тем самым, отображая высокоактивную сейсмическую зону, которая соединяет гипоцентры нескольких землетрясений.

3.3 Результаты экспериментальных исследований применения алгоритма DBSCAN для кластерного анализа набора сейсмических данных

В процессе кластерного анализа пространственных сейсмических данных алгоритмом DBSCAN, использованы географические координаты (долгота, широта). При предварительной обработке данных географические координаты были преобразованы в радианы для корректного вычисления расстояния между двумя точками на поверхности земли. В связи с этим, в качестве метрики расстояния между двумя точками использована формула ‘Haversine’. Формула ‘Haversine’ была впервые открыта James Andrew в 1805 году и впервые была использована Josef de Mendoza y Ríos в 1801 году [153]. Сам термин ‘Haversine’ был введен в 1835 году профессором James Inman [153, p. 3]. Josef de Mendoza y Ríos впервые использовал ‘Haversine’ в своем исследовании «Main Problems of Astronomical Nautical», Proc. Royal Soc, 22 декабря 1796 г. [153, p. 3]. ‘Haversine’ использовался в первую очередь для определения расстояний между звездами [154].

Формула ‘Haversine’ имеет свой собственный закон, который заключается в том, что все уравнения используются на основе формы сферической Земли за счет исключения фактора, что Земля имеет слегка эллиптическую форму (фактор эллипсоида) [153, p. 3]. Это частный случай общей формулы сферической тригонометрии, которая связана со сторонами и углами сферического треугольника [153, p. 3]. На расчет расстояния от одной точки до другой на земной поверхности влияет определенная степень кривизны [154, p. 272; 155-157].

В шаровой единице «треугольник» на поверхности сферы в соответствии с рисунком 31 [153, p. 3] определяется как большие круги, соединяющие три точки: u , v и w на шаре.

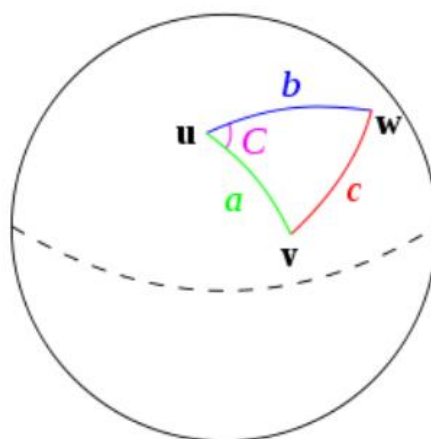


Рисунок 31 – Закон ‘Haversine’

Примечание – Рисунок адаптирован из работы [153, p. 3]

Если длина трех сторон равна a (от u до v), b (от u до w) и c (от v до w), а угол противоположного обзора c равен C , то закон ‘Haversine’ состоит из

следующих уравнений (11), как показано ниже:

$$\begin{aligned} \text{haversine}(c) &= \text{haversine}(a - b) \\ \sin(a)\sin(b)\text{haversine}(C) \end{aligned} \quad (11)$$

Эта формула (11) обычно записывается в терминах функции *haversine*, которая задается формулой (12):

$$\text{haversine}(\theta) = \sin^2(\theta/2) \quad (12)$$

где θ является центральным углом между любыми двумя точками на сфере и вычисляется по формуле (13):

$$\theta = \frac{d}{r} \quad (13)$$

где d – означает расстояние между двумя точками по большому кругу сферы;
 r – радиус Земли (средний радиус = 6 371 км).

Формула (12) *haversine* измеряет расстояние по дуге между двумя точками на сфере с учетом их долготы и широты, где расстояние по дуге большого круга является кратчайшим расстоянием между двумя точками на поверхности формула (14):

$$d = 2r \sin^{-1} \left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)} \right) \quad (14)$$

при этом φ_1 и φ_2 – это широта точки 1 и широта точки 2 (в радианах), а λ_1 и λ_2 – долгота точки 1 и долгота точки 2 (в радианах).

В большинстве исследований в качестве метрики расстояний между двумя точками, кроме Евклидова расстояния, часто используют формулу ‘*Haversine*’. Данная метрика применяется в основе алгоритмов машинного обучения при кластеризации таких наборов данных, как сейсмические данные [158], данные о вулканах, а также о данных цунами, вызванные землетрясениями [159].

Как уже упоминалось ранее (в пункте 3.2, Раздел 3), для DBSCAN необходимы параметры *Eps* и *MinPts*. Для определения оптимального значения параметра *Eps* используется метод *k-NN Distance* (*k-nearest neighbor distance*) [151, р. 11]. Суть метода заключается в вычислении расстояний в матрице точек до *k*-ближайших соседей. Вычисляется среднее расстояние от каждой точки до *k* ближайших соседей. Значение *k* устанавливается в соответствии со значениями *MinPts*. Данные *k*-расстояния отображаются в порядке возрастания. Выбирается значение, соответствующее порогу, на котором происходит резкое

изменение кривой расстояния k . Как видно в соответствии с рисунком 32, оптимальным значением для $Eps=0.005$.

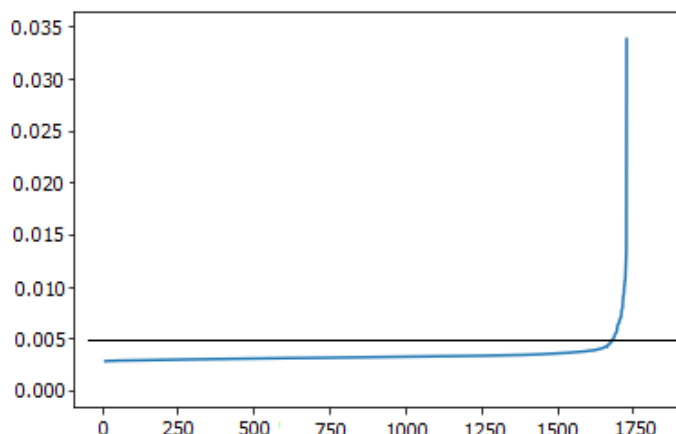


Рисунок 32 – Определение оптимального значения Eps

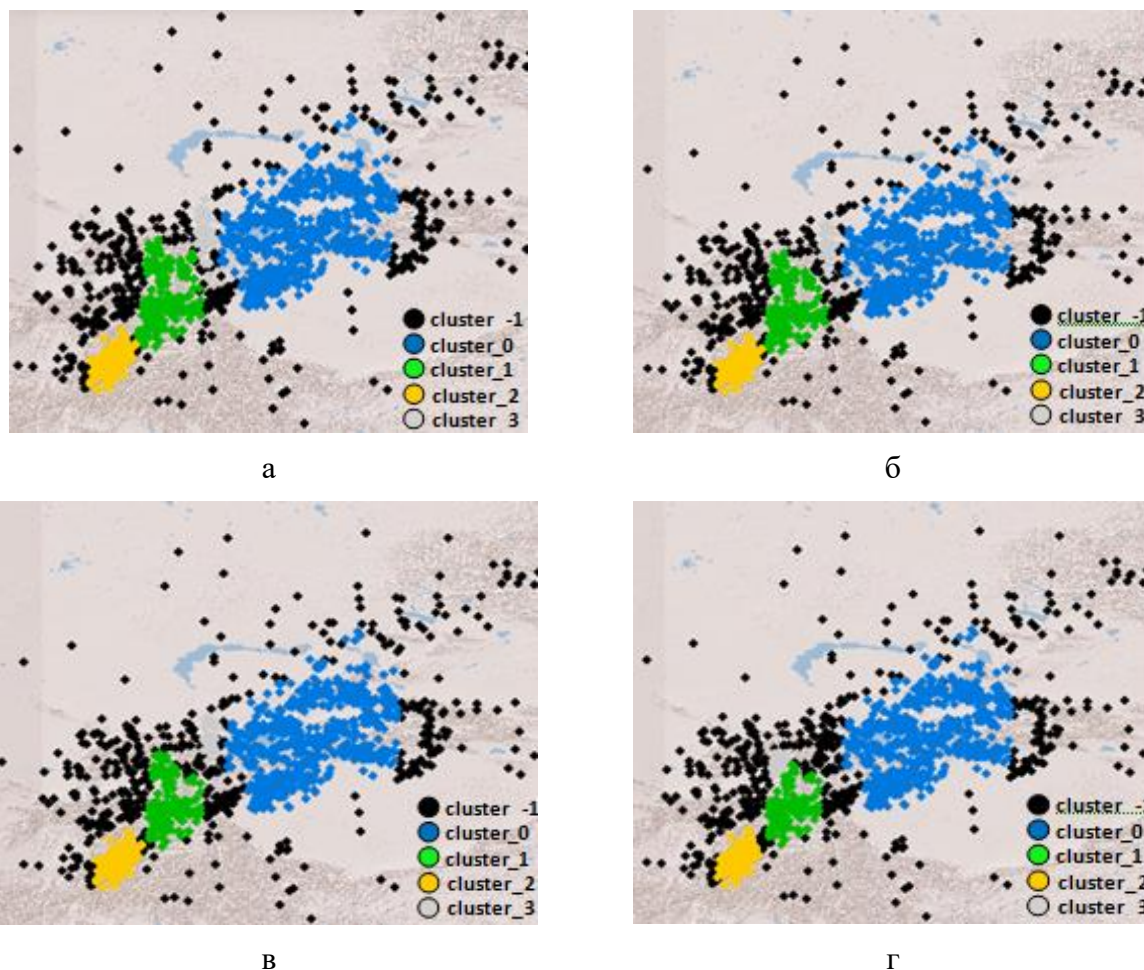
Физически сейсмические провинции не должны сильно меняться в разные периоды времени. Это означает, что количество кластеров должно быть практически одинаковым во все периоды времени [160]. На основе найденного параметра Eps и в результате проведения ряд экспериментов был осуществлен выбор оптимальных значений $MinPts$, результаты которых представлены в таблице 18.

Таблице 18 – Подбор параметров $MinPts$

Eps	$MinPts$	Number of Clusters	Noise Ratio, %
0.005	20	7	23,3
	21	8	26
	22	8	25,8
	23	8	27
	24	5	31.1
	25	5	32
	26	5	34,2
	27	5	36,5
	28	5	37,8
	29	5	40,3
	30	6	41,2
	31	6	43,1

По данным, представленным в таблице 17, можно увидеть, что при одном и том же значении Eps и с постепенным увеличением значений $MinPts$ число кластеров варьируется по-разному, что касается данных, идентифицированных, как шум $NoiseRatio$ постепенно увеличивается. Уменьшение объема зашумленных данных $NoiseRatio$ приводит к уменьшению количества кластеров и к снижению точности полученных результатов. В результате в качестве оптимальных значений выбраны именно те значения $MinPts$, при

которых число кластеров показывало постоянное значение. Сейсмические участки с повышенной пространственной плотностью с оптимальными значениями представлены в соответствии с рисунком 33.



а – Eps=0.005, MinPts=24, Number of clusters=5; б – Eps=0.005, MinPts=25, Number of clusters=5; в – Eps=0.005, MinPts=26, Number of clusters=5; г – Eps=0.005, MinPts=27, Number of clusters=5

Рисунок 33 – Результаты работы алгоритма DBSCAN при оптимальных параметрах Eps и MinPts

Результаты пространственной кластеризации на основе плотности точек, выявил 5 кластеров и показывает, что наибольшая часть плотных пространственных точек (cluster_0) приходится Джунгаро-Северо-Тянь-Шаньскому региону, т.е. принадлежит сейсмоактивному поясу Казахстана и к ней прилегающим сопредельным территориям других стран. Общеизвестно, что регион Тянь-Шаня и вся территория Центральной Азии остаются сейсмоопасными регионами из-за тысячи активных разломов, в том числе и из-за главного Джунгарского разлома. Данный факт, в первую очередь объясняет саму природу возникновения очагов землетрясений.

При сопоставлении выявленных кластеров с общей картой сейсмогенерирующих зон территории Казахстана [161], определены такие

зоны, как Южно-Прибалхашская, Дегересская, Алматинская, Заилийская, Кеминская, Северо-Кунгейская, Дубунская, Кетменская, Северо-Кетменская, Центрально-Кетменская, Басулытауская, Баянкольская. Также, cluster_0 включает в себя границы соседних государств. Cluster_1 и cluster_2 также относятся к регионам соседних стран. Cluster_3, также как Cluster_1 и Cluster_2 не принадлежит территории РК. Точки, не соотнесенные алгоритмом ни к одному кластеру, отмеченные как шумовыми точками (cluster_-1), а также отдельные отдаленные шумовые точки требуют дополнительных исследований.

Также важно отметить, что применение алгоритма пространственной кластеризации для исследования сейсмических наборов данных показал устойчивые значения результатов кластеризации (как и в случае использования набора данных в количестве 1905 записей зафиксированных сейсмических событий, так и в случае 1742 записей).

В настоящее время природную опасность в виде землетрясения все еще невозможно точно предсказать, однако анализ сейсмических данных на основе технологии анализа данных и методов машинного обучения дает возможность глубинного исследования этого явления. В этой части исследования предложен новый подход обнаружения и кластеризации сейсмических данных. Исходя из кластерного анализа, следует отметить, что большинство землетрясений не принадлежат территории РК. В будущем результаты кластерного анализа на основе плотности могут быть полезными для дальнейшего выделения сейсмогенерирующих зон и определения сейсмического потенциала земной коры Казахстана, а также способствовать улучшению количественной оценки пространственно-временного распределения землетрясений.

Выводы по третьему разделу:

1. В данном разделе представлены исследования сейсмического набора данных на основе технологии анализа данных и алгоритма машинного обучения. Исследование проводилось в два подэтапа. В первом подэтапе исследования сейсмического набора данных применен статистический анализ данных, который позволил выявить количественные показатели по временным интервалам, включая время возникновения, периодичность в месяцах, годах и сезонность. Кроме этого, определены классы магнитуд землетрясений, а также их распределения на интерактивной карте и выявлены принадлежности территории РК.

2. Во втором подэтапе исследования и анализа сейсмического набора данных предложен новый подход обнаружения и кластеризации сейсмических данных. Согласно результатам исследования и кластерного анализа, были обнаружены кластеры с повышенной пространственной плотностью по исследуемому набору сейсмических данных, где определены оптимальные значения входных параметров алгоритма $Eps=0.005$ и $MinPts=24$, $MinPts=25$, $MinPts=26$, $MinPts=24$. Применение данных значений в качестве входных параметров алгоритма DBSCAN для пространственной кластеризации сейсмических данных, были получены устойчивые показатели количества

кластеров Number of clusters=5. Изучение этих кластеров в соответствии с общей картой сейсмогенерирующих зон Казахстана, позволило определить надлежащие сейсмические регионы. Рассматриваемый подход кластерного анализа также применим и для других наборов данных в подобного рода исследованиях. Эффективность применения предложенного подхода доказана на данных РК. Таким образом, реализация пространственной кластеризации сейсмических данных является также важной частью в разработке способа оценки сейсмостойкости городских типовых объектов для проведения комплексной оценки сейсмической уязвимости зданий. Полученные результаты докладывались и опубликованы в работах [18, р. 149; 146, с. 36].

4 АРХИТЕКТУРА И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ИНТЕЛЛЕКТУАЛЬНОЙ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ (ИИАС) ОЦЕНКИ СЕЙСМОСТОЙКОСТИ ГОРОДСКИХ ОБЪЕКТОВ С УЧЕТОМ СЕЙСМИЧЕСКИХ ДАННЫХ

4.1 Общее описание и концептуальная основа ИИАС

Современные научные направления, связанные по исследованию вопросов обеспечения безопасности и оценки рисков, в настоящее время сложно представить без использования надёжных, точных и быстрых методов анализа данных и прогнозирования. Очевидно, что с ростом и накоплением большего объёма различных данных, в т.ч. и данных об объектах из городской инфраструктуры с параметрами их технического состояния, включающие экспертные оценки, а также данные, связанные с их местоположениями, характеризующие географические особенности той или иной местности, территории страны, требуют решения задач на основе обработки и экспресс анализа данных. Использование технологий интеллектуального анализа данных в совокупности с методами машинного обучения дают хорошую возможность для извлечения необходимых знаний из накапливаемых наборов данных. Такая интеграция современных подходов способствует быстрому принятию решений и оценки текущей ситуации на государственном уровне для стабилизации экономических последствий, а также в первую очередь обеспечению сохранности жизни людей.

В целом, аналитические системы на основе технологий анализа данных и алгоритмов машинного обучения успешно применяются в различных ГИС [162, 163], геопорталах, фреймворках данных и т.д. Выполнение аналитической работы без внедрения инструментов технологии анализа данных и алгоритмов машинного обучения, лишь только человеком не всегда рентабельно, так как анализ может оказаться трудоемким процессом и требовать привлечения специалистов высокой квалификации. К тому же, закономерности, которые скрыты в данных, не всегда могут быть обнаружены человеком.

В большинстве случаев любые наборы данных по истечении времени накапливают большой объем данных, которые в свою очередь требуют для обработки современные технологии анализа данных и новые IT-решения. Такой объем данных со временем не поддается обработке традиционными методами управления данными, таких как, системы управления реляционными базами данных (РСУБД). При работе с растущим набором данных РСУБД являются неэффективными и не пригодными для использования в системах хранения и анализа больших данных.

В связи с этим, в рамках диссертационной работы была разработана масштабируемая архитектура для анализа данных и оценки сейсмостойкости городских объектов, включая пространственные сейсмические данные о землетрясениях. По сравнению с уже существующими архитектурами платформ и информационно-аналитических систем, данная архитектура является экономичной, оперативной и открытой. Предлагаемая в контексте

диссертационного исследования информационная технология, была реализована в виде интеллектуальной информационно-аналитической системы (ИИАС) с веб-интерфейсом, которая анализирует данные, связанные с оценкой сейсмостойкости городских объектов с учетом сейсмических данных о землетрясениях. Реализация системы и автоматизация процесса, которую обеспечивает данная интеллектуальная информационно-аналитическая система (ИИАС), позволяет сократить время на выполнение анализа данных и сделать сам процесс дешевле, что подтверждает практическую значимость работы.

Концепция данной интеллектуальной информационно-аналитической системы (ИИАС) поддерживает следующие возможности:

- сбор и хранение необработанных данных в большом масштабе;
- хранение многих типов данных в одном хранилище данных;
- выполнение преобразований данных;
- определение структуры данных во время ее использования (схема при чтении);
- выполнение аналитических операций по двум предметам (оценка сейсмостойкости городских объектов и анализ сейсмических данных) на основе конкретных (описания алгоритмов были представлены в Разделах 2 и 3) алгоритмов машинного обучения.



Рисунок 34 – Концепция разработки ИИАС

Спроектированная архитектура предназначена для создания и реализации масштабируемых API-интерфейсов для анализа данных и оценки сейсмостойкости городских объектов с учетом пространственных сейсмических данных.

Также данная архитектура реализована при разработке аналитических сервисов в помощь исследователям и аналитикам из рассматриваемой предметной области, предоставляя им индивидуальные решения на основе интеллектуального анализа данных и применения методов и/или алгоритмов машинного обучения, результат которых отображается в виде отдельных моделей, а также интерактивной картой с отображением на ней кластеров. Архитектура ИИАС имеет отличительные особенности, которые отображены в соответствии с рисунком 34.

4.2 Инструментарий и технологии анализа данных в архитектуре ИИАС

В структуре системы ИИАС для работы с наборами данных или если рассматривать в общем понимании, то с работой базами данных, используется технология NoSQL. Данная технология по сравнению с традиционными системами, где в основном для обращения к базам данных используется язык запросов SQL, отличается тем, что NoSQL является более эффективной технологией для хранения и обработки большого количества данных, в т.ч. и данных любого масштаба и в режиме реального времени.

В проектировании внутренних компонентов архитектуры ИИАС для работы с базами данных используется технология Cassandra. Технология Cassandra – это база данных NoSQL с широкими столбцами [164]. Cassandra является линейно масштабируемой, долговременной и высокодоступной, согласованной и распределенной базой данных. Данная технология была разработана в 2008 году компанией Facebook. Технология имеет свой открытый исходный код. Технология является проектом Apache, а его корпоративная версия поддерживается Datastax. Применение данной технологии работы с базами данных будет отличным выбором для многих архитекторов и разработчиков, которые разрабатывают системы и приложения для работы с большими данными на основе структурированного набора данных.

Отличительной особенностью Cassandra, по сравнению с другими технологиями разработки приложений и работы с данными, особенно, в первую очередь работы с большими данными, является тот факт, что она превосходит распределенную файловую систему Hadoop (HDFS) [165]. Однако, для данной технологии необходимо интегрироваться с вычислительными платформами для хранения, объединения, агрегирования и группировки больших данных (Spark/Hadoop), чтобы можно было выполнять расширенные аналитические операции. Архитектурные особенности и достоинства Cassandra, представлены в соответствии с рисунком 35.



Рисунок 35 – Особенности в архитектуре Cassandra

В одноранговой архитектуре технологии Cassandra все узлы являются равными между собой, нет главного узла. Все узлы одновременно выполняют чтение и запись. Кроме этого, в архитектуре технологии Cassandra есть еще один важный аспект, такой как наличие двух ключевых компонентов, которые представлены в соответствии с рисунком 36.

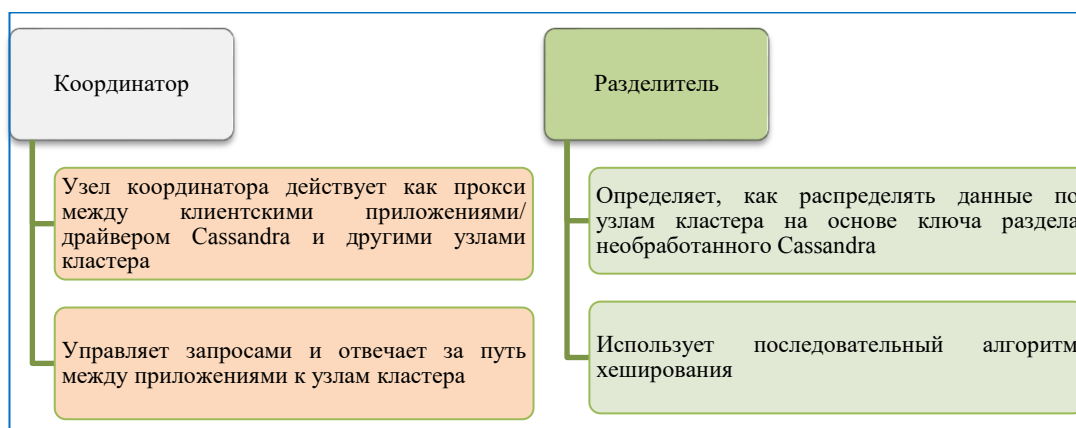


Рисунок 36 – Ключевые компоненты в архитектуре Cassandra

В проектировании архитектуры ИИАС, кроме системы хранилища данных и работы с набором данных, также важной составляющей является сама платформа для развертывания архитектуры системы. В качестве такой платформы использована платформа для обработки больших данных Spark.

Apache Spark является универсальным, унифицированным механизмом для обработки больших данных с открытым исходным кодом. Обоснованность выбора данной платформы, регламентируется тем, что данные, связанные с сейсмической обработкой растут по мере их возникновения и регистрации даже незначительных толчков в глубине земной коры в виде слабых землетрясений,

помимо и сильных. Кроме этого, имеются и исторические данные о землетрясениях с колоссальным временным промежутком в интервале многих лет существования человечества. С учетом таких данных и накопления новых данных, в будущем потребуются необходимость обработки больших данных и применения технологий для аналитики больших данных. В связи с этим, обосновывается такой выбор технологий для имеющихся данных в структуре ИИАС. Конечно же, по мере эксплуатации ИИАС при уже огромном количестве подобных данных, будут и модифицироваться алгоритмы и методы машинного обучения, а также технологии моделирования моделей и вычислений. Такая особенность в информационном и технологическом подходе, является отличительным признаком от всех других известных систем, решающие подобные задачи, что еще раз доказывает масштабируемость разработанной системы ИИАС.

Apache Spark прост в использовании, хорошо масштабируется и отказоустойчив. В Apache Spark ядро и библиотеки верхнего уровня являются основными компонентами. Также Apache Spark способен решать ряд проблем с большими данными в широком диапазоне рабочих нагрузок, такими как пакетная, итеративная и потоковая передача в одном кластере. Библиотеки верхнего уровня Apache Spark предназначены для потоковой обработки (Spark Streaming), обработки структурированных данных (SparkSQL), машинного обучения (MLlib) и обработки графов (GraphX).

Apache Spark дает возможность для разработки гибкой среды с интерактивной оболочкой и интегрированные API, которые могут быть разработаны с мощными языками программирования, как R, Python, Java. В случае с разработанной ИИАС и в контексте данной диссертационной работы, был использован язык программирования R, Rstudio и RShiny для интеграции API в общую архитектуру ИИАС.

Apache Spark не имеет собственной системы управления файлами, и поэтому его необходимо интегрировать с внешними источниками данных, такими как Cassandra. Существующие решения на основе Spark для пространственных данных могут выполнять пространственные операции только с наборами данных, которые доступны в текстовых форматах файлов и хранятся на локальном диске (или в HDFS).

В качестве инструментария для визуализации использован язык программирования с открытым исходным кодом R. По своей популярности данный язык находится пятым в рейтинге. В основном R используется для решения задач анализа данных, связанные с наукой о данных, также в статистическом моделировании данных и визуализации анализируемых наборов данных, которые помещаются в основную память. Так как масштабируемые решения для уже более больших наборов данных будут сложной задачей, в виду своей ограниченности памяти, то в разработке интерфейсной части ИИАС, использовался Sparklyr. Sparklyr [166] является R-интерфейсом для Apache Spark. Sparklyr позволяет преобразовывать данные на основе dplyr для объектов Spark, а также предлагает масштабируемые

библиотеки для сортировки, объединения, группировки и выборки данных. Кроме этого, Sparklyr применяется для доступа и анализа данных внутри Spark и отображает результаты только в R, а также Sparklyr используется для развертывания заданий Spark в кластере Spark удаленно.

При разработке архитектуры ИИАС для ее надежной работы и эффективной функциональности, были учтены и использованы интеграции двух структур Cassandra (хранилища данных NoSQL) и платформа Spark. В архитектуре ИИАС для обеспечения интеграции этих двух структур используется коннектор Spark-Cassandra в соответствии с рисунком 37.

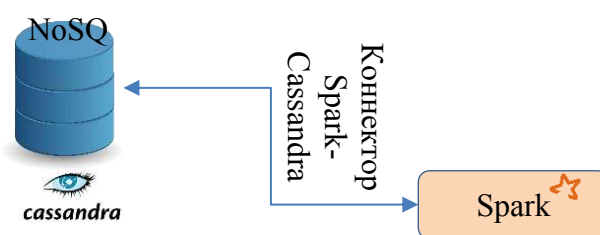


Рисунок 37 – Интеграция Cassandra и Spark

Данный коннектор в структуре архитектуры обеспечивает обработку данных для хранилища данных Cassandra с помощью Spark и поддерживает язык программирования R через Sparklyr, тем самым предоставляю доступ к ключевым данным, имеющиеся в самой базе данных Cassandra.

4.3 Структура и описание данных

Достаточное количество и объем сложных, разнородных данных в большинстве случаев, генерируются из различных источников и могут быть использованы для исследований заинтересованными лицами на практике. Множество поступающих данных, обычно распределены неоднородно и доступны из разрозненных источников и, как правило, обрабатываются разными методами и в разных форматах. Технология анализа данных используется для анализа данных и данных с большим объемом структурированных и неструктурированных данных для получения более быстрых и точных результатов, которые в дальнейшем помогут исследователям в более быстром и точном принятии решений.

Подходы и технологии, а также проблемы, связанные с разработкой систем и приложений для быстрой обработки данных, на сегодняшний день имеют свои принципиальные и технологические различия. С учетом растущей ежедневной информации и накапливающихся данных, каждый исследователь или разработчик, той или иной системы или веб-приложения, может использовать передовые технологии анализа данных, в т.ч. и применять технологии большой аналитики, с прицелом и учетом, на то, что в будущем возникнет необходимость и потребность в обработке объемов больших данных. Технология анализа данных обладает отличным потенциалом для обращения к

этой информации и получения исчерпывающих сведений с помощью алгоритмов машинного обучения, анализа пространственных данных, применения передовых, облачных ресурсов и систем хранения. За счет использования подобных технологий анализа данных, разработанная система легко сможет быть адаптирована и масштабирована до обработки огромных объемов информации. Так, к примеру, в проектировании и реализации рассматриваемой ИИАС, также был применен подход использования технологии Spark Apache и технологии хранения данных Cassandra.

В хранилище данных есть два вида разнородных структурированных данных, которые используются в двух модулях системы анализа данных и вычислительного процесса с помощью методов и алгоритмов машинного обучения для построения аналитических моделей с целью принятия решений специалистами и исследователями в этой сфере деятельности, науки.

Самая основная задача в реализации и разработке систем и приложений на основе технологии анализа данных является то, что необходимо привести эти данные в действие. Решение такой проблемы может быть осуществлена на широком спектре различных технологических платформ и механизмов, таких как управление реляционными базами данных (СУБД) и нереляционными данными (к примеру, база данных столбцов) или другими аналитическими СУБД, а также потоковой обработкой данных. Основная задача состоит в том, чтобы создать и внедрить надежную, эффективную систему для анализа и управления необходимым объемом данных и включая пространственные данные.

В технологии анализа данных одной из важных задач является выполнение таких аналитических операций, как:

- загрузка данных;
- предварительная обработка данных;
- интеграция и преобразование зашумленных данных.

Множество данных для анализа поступают из различных источников, таких как данные на основе Интернета вещей (IoT) или с сайтов с определенных служб деятельности, также данные, созданные машинами. Эти разнообразные данные необходимо моделировать и хранить так, чтобы их можно было эффективно обрабатывать и анализировать для получения ценной информации.

ИИАС реализована для сбора, обработки, анализа данных методами и алгоритмами машинного обучения и визуализации результатов анализа данных в масштабе с использованием открытого исходного кода.

Необходимо отметить, что в области сейсмостойкого строительства очень мало открытых наборов данных или цифровых данных, в частности в РК. Следовательно, в связи с этим, сбор данных также является одной из важнейших задач для разработки различных информационных систем и приложений данных в области сейсмостойкого строительства. Пространственные и структурированные (полуструктурированные) данные о землетрясениях могут собираться из различных источников, таких как

информационные сайты, веб-порталы и т.д., а также с веб-источников специализированных институтов, которые ранее были уже описаны (Раздел 1, в соответствии с рисунком 1). Сведения о сборе данных приведена в таблице 19.

Таблица 19 – Виды наборов данных ИИАС

Наборы данных	Источник данных	Описание	Формат данных
Набор данных 1 здания	Отчетная документация в открытом доступе	Данные собраны по районам г. Алматы	Таблица / .csv
Набор данных 2 землетрясения	http://some.kz/	Данные собраны с 2011 года по 2021 год включительно	Таблица / .csv

В проектировании хранилища данных ИИАС для получения согласованных наборов данных и прежде всего, чтобы начать процесс анализа данных на основе определенных методов (алгоритмов) машинного обучения, к наборам данных применяются методы предварительной обработки и интеграции данных, такие как: обработка отсутствующих значений, удаление дубликатов данных, проверка данных, поиск сходства и объединение. Разработанная для анализа данных ИИАС, включающая также пространственные данные реализована на основе платформы анализа данных.

Наборы данных 1 о городских объектах моделируются и сохраняются в базе данных Cassandra. Модели данных Cassandra для необработанных наборов данных и агрегированных наборов данных показаны в таблице 20 соответственно.

Таблица 20 – Модель данных Cassandra для необработанных наборов данных 1

Имя таблицы	Схема	Раздел ключ	Кластеризация ключ
raw_buildings	<object, location, pr_develop_year, year_constr, type_pr, seism_cat soils, seism_area, space-plan_sol, floors, antiseism_activ, gen_char, foundations, bearing_struct, floor_struct, wall_fence, partitions, height, total_area, construct_vol, conclusion, assessment, build_cost thousand.\$, build_cost thousand.KZT, people>	assessment	object type_pr foundations

Наборы данных 2 о землетрясениях также сохраняются в базе данных Cassandra. Модели данных Cassandra для необработанных наборов данных и агрегированных наборов данных показаны в таблице 21.

Таблица 21 – Модель данных Cassandra для необработанных наборов данных 2

Имя таблицы	Схема	Раздел ключ	Кластеризация ключ
raw_seismic_data	< data, greenwichmeantime, latitude, longitude, depth, energy_class, magnitude >	latitude longitude	depth magnitude

При проектировании архитектуры ИИАС процесс моделирования данных является важным аспектом для работы с базой данных NoSQL на основе технологии Cassandra. Cassandra реализует модель данных на основе запросов. Модель входного набора данных должна равномерно распределять данные между всеми узлами в кластере, а также минимизировать количество разделов во время чтения данных для дальнейшего их извлечения. Входные наборы данных представляются в формате .csv и подготавливаются для извлечения с помощью служб подготовки данных. После этого последовательно загружаются в базу данных Cassandra. Входные наборы данных с различными атрибутами данных хранятся в базе данных Cassandra на основе модели данных. В качестве примера в соответствии с рисунком 38 представлена модель хранения данных для атрибутов «долгота» и «широта» для зафиксированных сейсмических событий в наборе данных 2.

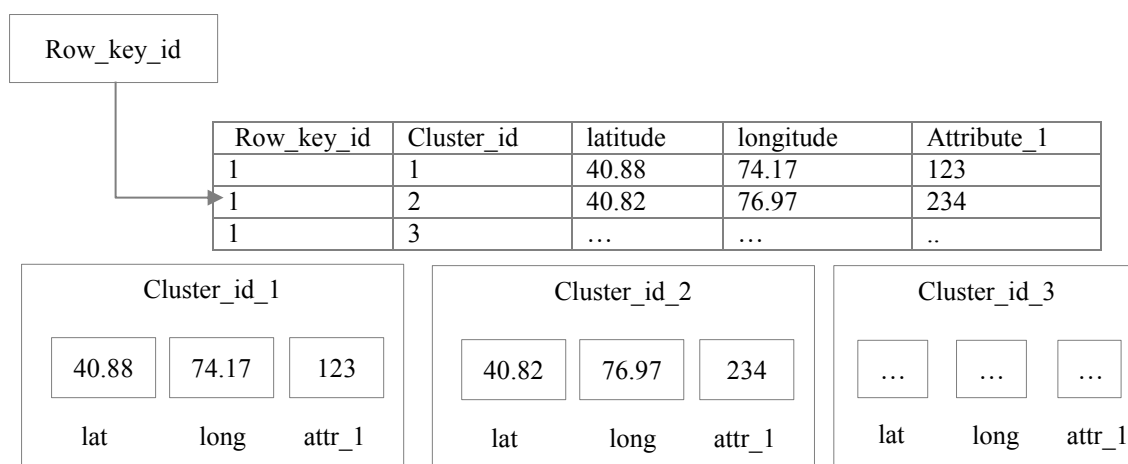


Рисунок 38 – Пример модели хранения данных Cassandra по атрибутам географических координат сейсмических событий из НД 2

Таким образом, разработанная методика оценки сейсмостойкости городских объектов и анализ пространственных сейсмических данных релизуются в структуре ИИАС на основе предложенной архитектуры (в соответствии с рисунком 40 в пункте 4.4 Раздела 4). Веб-службы аналитики и визуализации разработаны для получения модели оценки сейсмостойкости зданий и важных, информативных характеристик, влияющие на общую модель оценки по зданиям г. Алматы (РК). В целом Алматинская область относится к южной границе сейсмогенерирующих зон, к которым приурочены сильные группы землетрясений с магнитудой 5,6-7,5 баллов, включая историческое

сильное землетрясение Верненское [167] 1887 года. Исходя из таких важнейших факторов, изучение и интеллектуальный анализ данных по городским объектам является также одной из главных задач в исследовательской работе.

4.4 Проектирование архитектуры и программный интерфейс ИИАС

Предложенная в диссертационной работе ИИАС, в первую очередь реализует цель разработки методов оценки сейсмостойкости городских объектов и анализа пространственных сейсмических данных для сокращения технологического разрыва между специалистами и исследователями, а также аналитиками из этой предметной области. Прототипы информационных систем и приложений в области анализа данных, связанные с исследованиями по обеспечению безопасности и оценки рисков разрабатываются чаще всего на основе архитектуры аналитики больших данных или используют технологии, которые в будущем позволят легко интегрировать и масштабировать систему под обработку больших данных. В качестве алгоритмов оценки городского объекта на сейсмостойкость, используются методы и алгоритмы машинного обучения, такие как кластеризация (K-Means/НК-Means) и методы классификации (Decision Tree and Random Forest [168, 169]). Результаты исследования НД 1 по Модулю 1 и реализация оценки городских объектов на сейсмостойкость, также полностью изложены в Разделе 2, а также в других научных работах [16, р. 154; 122, р. 261; 124, с. 1029]. В архитектуре ИИАС, спроектированные модули обладают такими свойствами, как функциональная целостность и завершенность. Каждый модуль реализует одну функцию и самостоятельно выполняет полный набор операций для реализации своей функции. На входе программный модуль получает определенный набор исходных данных и выполняет содержательную обработку, тем самым возвращает один набор результатных данных. Также, в архитектуре ИИАС, есть логическая независимость, результат работы программного модуля зависит только от исходных данных, но не зависит от работы других модулей. В целом, реализация архитектуры ИИАС состоит из основных трех этапов, которые отражены в соответствии с рисунком 39. Общая схема и полная структура разработанной архитектуры ИИАС представлена в соответствии с рисунком 40.

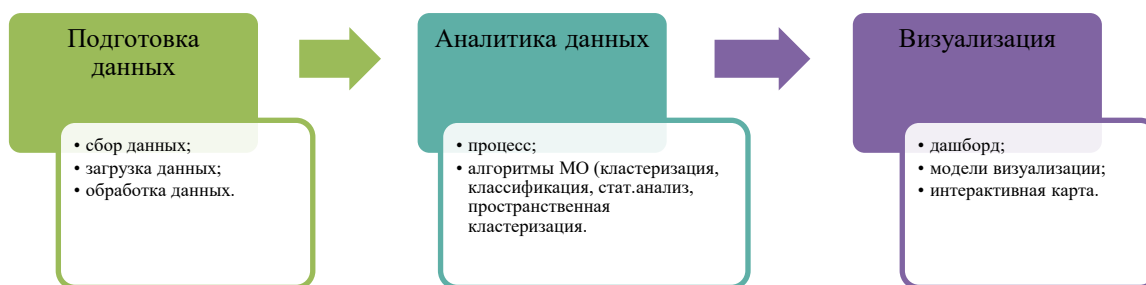


Рисунок 39 – Этапы реализации ИИАС

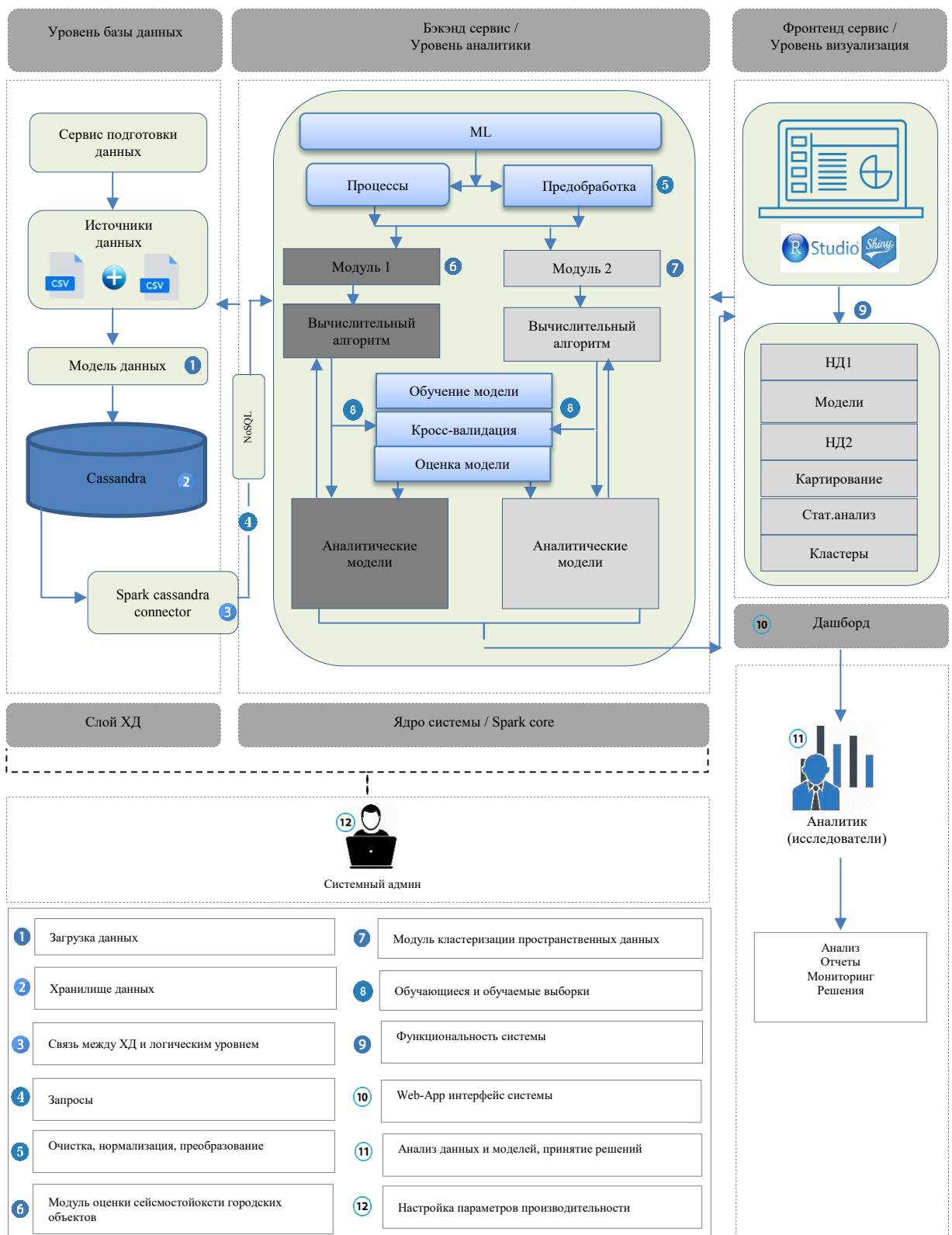


Рисунок 40 – Архитектура ИИАС

Программная архитектура программы или вычислительной системы - это структура или структуры системы, которые состоят из программных компонентов, видимых извне свойств этих компонентов и взаимосвязей между ними [170].

Уровень подготовки данных представляется источником данных, который служит для сбора и обработки данных. Так как данные, используемые в ИИАС, имеют два типа/два набора данных, то соответственно эти данные в качестве входных данных поступают из разнородных источников данных для выполнения соответствующей аналитики. В ИИАС набор данных 1 и набор данных 2 представляется в источнике данных в виде отчета. Службы подготовки данных получают чистые данные из разрозненных источников и сохраняют их в постоянной базе данных. Такими процессами представляются: извлечение и сбор данных; репозиторий данных; предварительная обработка и интеграция данных (в соответствии с рисунком 41).

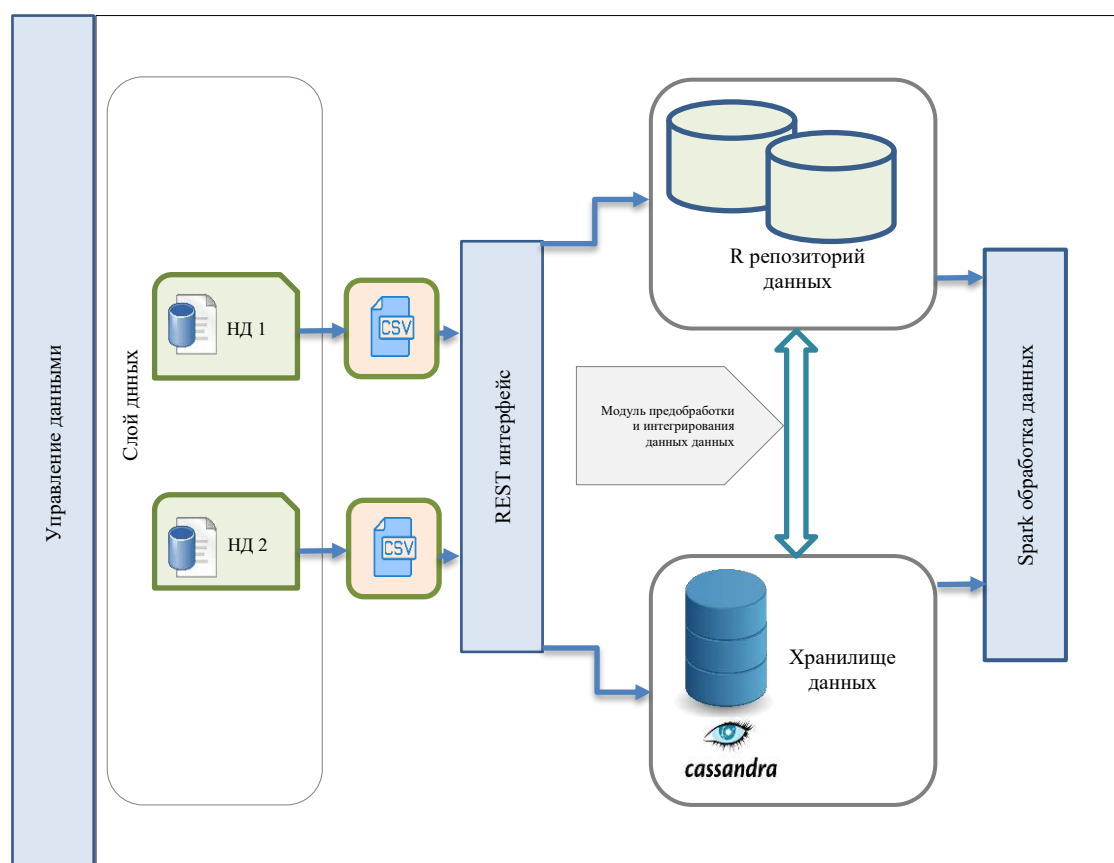


Рисунок 41 – Архитектура уровня работы с данными в ИИАС

После извлечения с источников данные, полученные интерфейсом REST, далее будут сохраняться в репозитории данных и в полуструктурированных форматах (JSON). В качестве базы данных выбрана MongoDB для репозитория данных и R, Nodejs используются для импорта данных и экспорта из MongoDB. По сути MongoDB является документо-ориентированной системой управления базами данных. Она не требует описания схемы таблиц и также является классическим примером NoSQL-систем, использующая похожие на JSON документы и схему базы данных.

Архитектура ИИАС разработана в виде платформы для аналитики данных. Основная задача ИИАС заключается в обработке, анализе данных и в построении моделей оценки сейсмостойкости городских объектов на примере

определенного набора данных 1, а также в обнаружении плотных участков кластеров землетрясений в пространстве (на карте) для анализа сейсмических данных и генерируя важную информацию из массива набора данных 2. В реализации таких сложных процессов, кроме информационного и технологического обеспечения ИИАС, используются инструменты и технологии анализа данных, как статистический анализ и машинное обучение. Более упрощенная, концептуальная модель ИИАС и ее реализация, представлена в соответствии с рисунком 42.

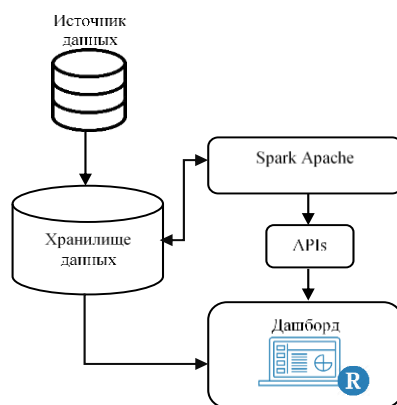


Рисунок 42 – Упрощенная концептуальная модель ИИАС

Спроектированная архитектура ИИАС предоставляет пользовательский веб-интерфейс, разработанный с помощью аналитических и визуальных сервисов посредством специальных API-интерфейсов Restful, в соответствии с рисунком 43 и интерактивной карты. Просматриваемая в структуре ИИАС интерактивная веб-карта реализована с использованием библиотек (leaflet, ggplot, Shiny) языка программирования R. Веб-службы вывода анализа и визуализации данных используют интерфейса SparkSQL/CQL поверх аналитической платформы.

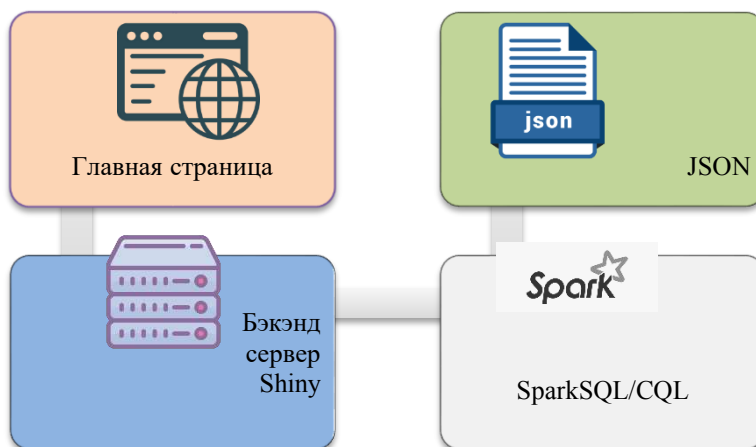


Рисунок 43 – Логическая модель реализации Restful API

В ИИАС визуализация исследованных наборов данных, полученных результатов анализа данных и моделей оценок, реализуется в виде приложения (дашборда). Панель управления приложения также разработана с помощью библиотек языка программирования R. Оконные интерфейсы и панели управления приложения представлены на следующих в соответствии с рисунками 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55.

Данная ИИАС не требует установки специального программного обеспечения на компьютер, так как имеет веб-ориентированную архитектуру, что позволяет беспрепятственно загружать ее с любой рабочей области, подключенной к сети Интернет.

Реализация сложных алгоритмов машинного обучения является частью более крупного процесса, в который входит задача оценки сейсмостойкости городских объектов с учетом сейсмических данных. Созданные модели с помощью технологии анализа данных и методами/алгоритмами машинного обучения развернуты на основе разработанной архитектуры ИИАС. Следует отметить, что еще одной небольшой особенностью в интерфейсной части ИИАС, является применение английского языка с целью расширения круга пользователей в лице аналитиков и специалистов из полиязычной среды.

Структура ИИАС анализирует и обрабатывает два совершенно разных наборов данных. В связи с этим, ниже следуют описания интерфейсных окон аналитики и моделей по модулю 1 оценки сейсмостойкости городских объектов. Окно модуля 1 ИИАС содержит фрейм меню (в соответствии с рисунком 44). Меню интеллектуальной ИИАС состоит из следующего:

- методы машинного обучения;
- набор данных по городским объектам.



Рисунок 44 – Аналитическое окно оценки сейсмостойкости городских объектов методами машинного обучения

При загрузке модуля 1 ИИАС сразу раскрывается окно, где отображается доступное меню с разделами, а также отображаются кластерная и

классификационные модели оценки сейсмостойкости городских объектов в соответствии с рисунком 45.



Рисунок 45 – Кластерная оценка сейсмостойкости городских объектов ИИАС

На основной странице модуля 1 в ИИАС реализованы алгоритмы машинного обучения: кластеризация K-Means, классификация методами Decision Tree и Random Forest. В соответствии с рисунком 45 представлена визуализация кластерной модели городских объектов. Количество полученных кластеров равно 4. В процессе работы ИИАС в модуле 1 при щелчке на вкладку «Models of Decision Trees» отображается классификационная модель, построенная методом дерева решений, и которая позволяет осуществлять оценку сейсмостойкости городских объектов с загруженного набора данных. Первая классификационная модель на основе метода DT представлена в соответствии с рисунком 46.

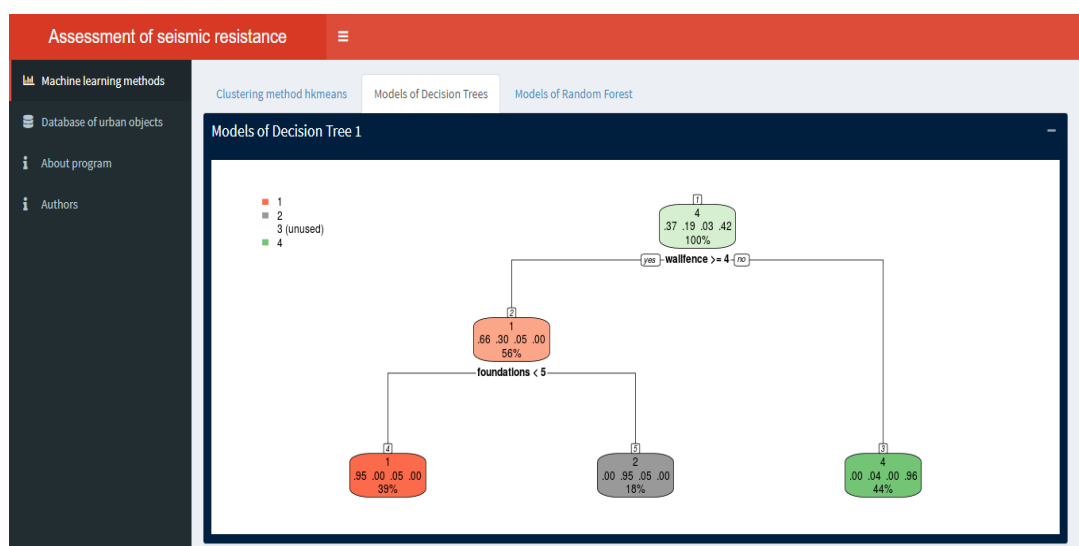


Рисунок 46 – Первая классификационная оценка сейсмостойкости городских объектов ИИАС

С помощью инструмента прокрутки вниз также можно увидеть общую полную классификационную модель на основе метода DT. Визуализация общей классификационной модели оценки сейсмостойкости городского объекта представлена в соответствии с рисунком 47.

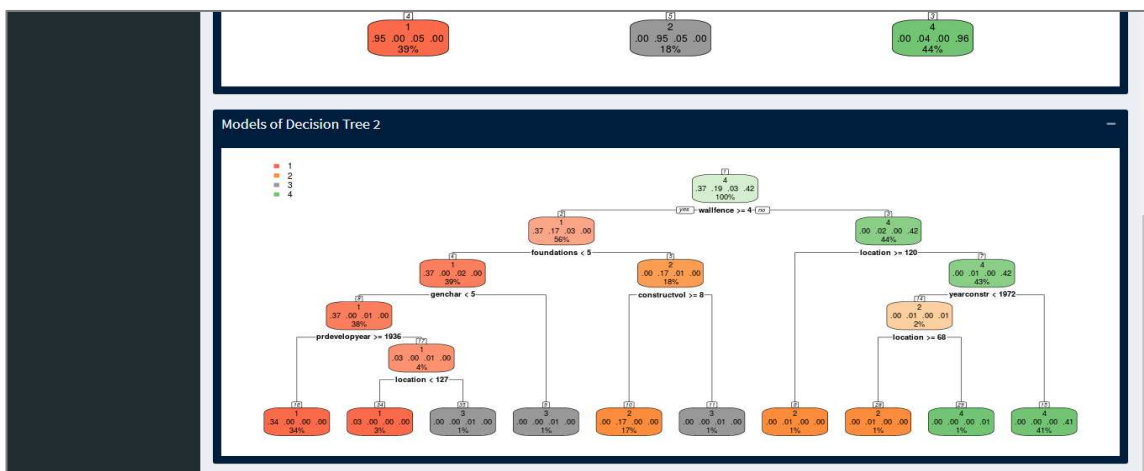


Рисунок 47 – Общая классификационная оценка сейсмостойкости городских объектов ИИАС

В следующей реализации оценки сейсмостойкости городских объектов, после того как были построены классификационные модели на основе метода RF выявляются важные информативные характеристики, участвующие в решении оценки сейсмостойкости городских объектов. Так как, RF является более сильным классификатором по сравнению с DT, то соответственно и алгоритм является более сложным и требует больше времени для обработки данных и загрузки модели. При щелчке на вкладку «Models of Random Forest» в данном окне загружается модель, которая определяет важные переменные необходимые для оценки сейсмостойкости городских объектов. Визуализация построения модели представлена в соответствии с рисунком 48.

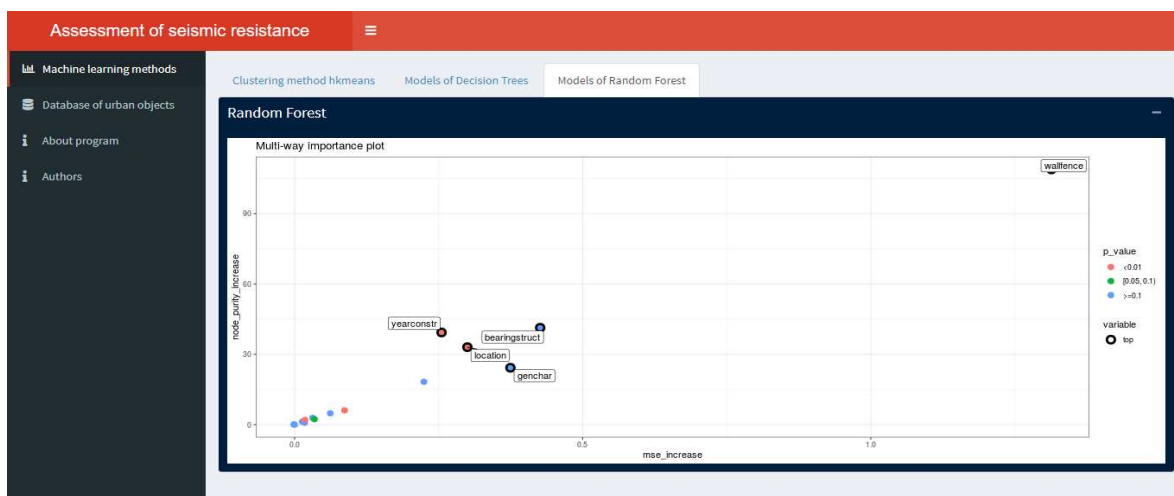


Рисунок 48 – Определение важных информативных характеристик для оценки сейсмостойкости городских объектов

Модуль 1 ИИАС в панели меню включает также раздел «Database of urban objects», который загружает и отображает набор данных городских объектов, представленный в соответствии с рисунком 49. Каждая запись набора данных представляет сведения об определенном городском объекте. В текущем окне предусмотрены функции поиска необходимой записи, а также отображение необходимого количества записей.

object	location	prdevelopyear	yearconstr	typepr	seismcatsoils	seismarea	spaceplansol	floors	antiseismactiv
1 жилое здание	мкр. Жетысу-2, д.37	1979	1992	типовой - серия 158	1	9	прямоугольной формы в плане, с перепадами по высоте, подвал	9	есть
2 жилое здание	мкр. Жетысу-2, д.47	1979	1993	типовой - серия 158	1	9	прямоугольной формы в плане, без перепадов по высоте, подвал	9	есть
3 жилое здание	мкр. Жетысу-2, д.46	1979	1993	типовой - серия 158	1	9	прямоугольное в плане, без перепадов по высоте, подвал	9	есть
4 жилое здание	мкр. Жетысу-2, д.52	1979	1993	типовой - серия 158	1	9	прямоугольное в плане, без перепадов по высоте, подвал	9	есть
5 жилое здание	мкр. Жетысу-3, д.17	2003	2005	индивидуальный	1	9	прямоугольное в плане, без перепадов по высоте, подвал	9	есть

Рисунок 49 – Окно загрузки «Database of urban objects» набора данных городских объектов

Следующий модуль 2 ИИАС представляет статистический анализ сейсмических данных о зарегистрированных землетрясениях и результат применения алгоритма пространственной кластеризации DBSCAN. Окно модуля 2 ИИАС также содержит фрейм меню (в соответствии с рисунком 50).

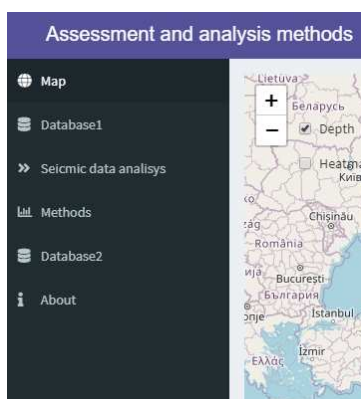


Рисунок 50 – Панель меню модуля 2 ИИАС

Меню модуля 2 состоит из следующего:

- интерактивная карта визуализации сейсмических данных, в соответствии с рисунком 51;
- набор данных 2 по землетрясениям;
- статистический анализ сейсмических данных;

– пространственная кластеризация алгоритмом DBSCAN.

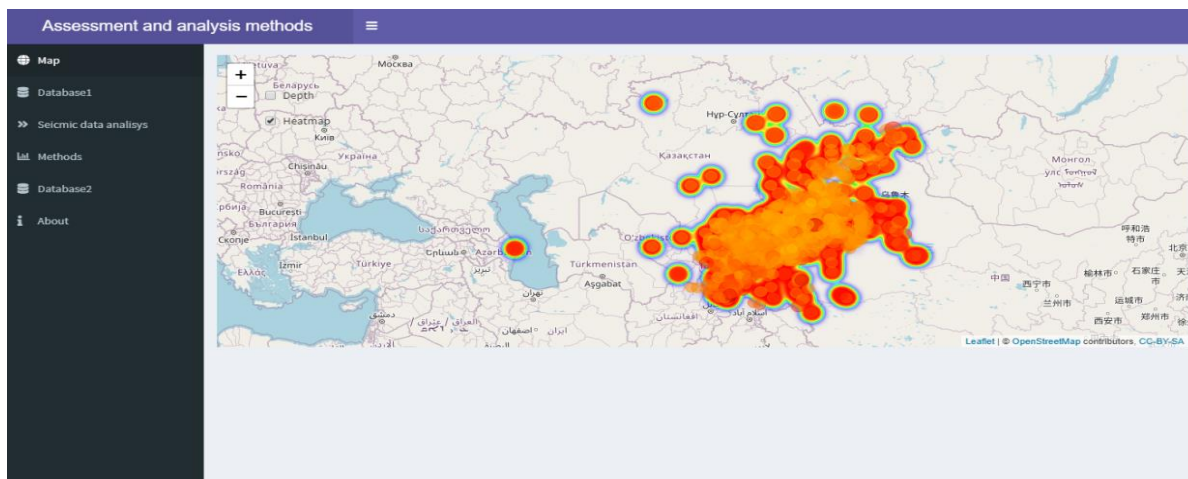


Рисунок 51 – Карта визуализации очагов землетрясений

При загрузке интерактивной карты сразу раскрывается окно приложения, где отображается доступное меню с разделами. С помощью знаков **+** и **-** можно увеличивать или уменьшать масштаб карты. Также на карте расположены 2 опции: Depth и Heatmap. При выборе опции Depth на карте в соответствии с рисунком 52, отображается значение глубины землетрясения. Показатели глубины до 70 (shallow) отображаются красным цветом, от 71 до 300 (intermediate) отображаются синим цветом.

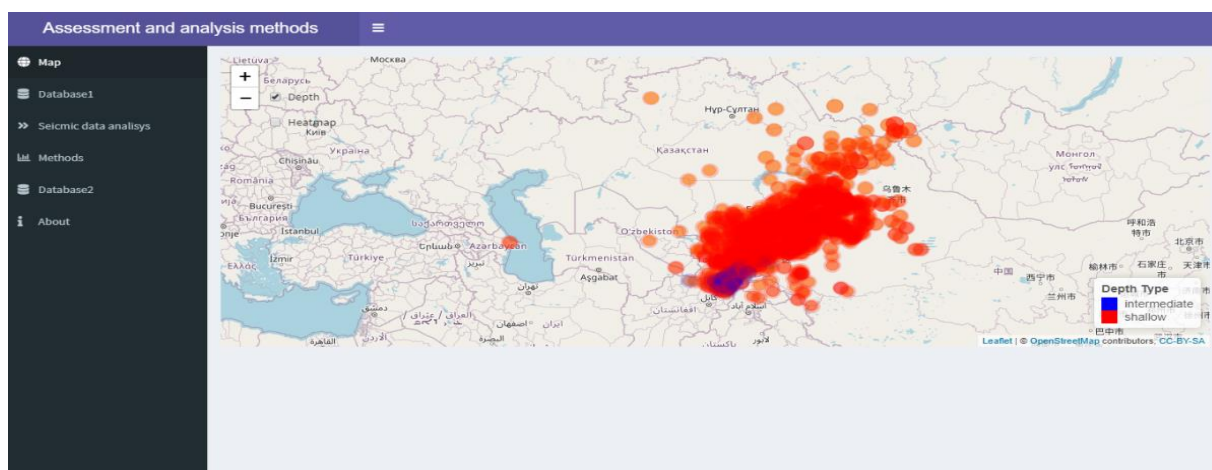


Рисунок 52 – Отображение землетрясений на карте по значениям глубины

Heatmap, т.е. тепловая карта показывает плотность значений магнитуды по землетрясениям. Значения представлены с помощью интуитивно понятных цветов; красный цвет представляет наиболее высокие значения, оранжевый цвет наиболее низкие значения. Все, что находится между этими границами, может быть закодировано по цвету с различными вариациями красного и оранжевого.

Содержание меню модуля 2, также может загружать и отображать набор

сейсмических данных, который представлен в соответствии с рисунком 53. Каждая запись набора данных представляет собой зарегистрированное событие землетрясения.

	date	time	lat	long	depth	energy_cl	mag
1	09.09.2020	16:16:54	37.9	72.13	190	10.5	4.7
2	01.09.2020	2:57:00	37.17	71.02	120	10.2	4.7
3	20.08.2020	15:21:03	42.83	79.03	15	9.7	4.8
4	20.08.2020	0:32:22	47	84.11	10	9.6	3.9
5	16.08.2020	13:06:33	42.51	79.66	25	8.7	4.2
6	16.08.2020	2:31:51	40.01	75.14	15	9.5	4.1
7	15.08.2020	21:58:09	39.76	74.01	5	10.3	4.6
8	12.08.2020	10:03:17	41.87	75.41	10	10.7	4.9
9	08.08.2020	0:35:11	42.44	87.11	10	12.2	5.2
10	06.08.2020	6:41:29	42.66	73.7	5	9.2	4.1

Рисунок 53 – Окно загрузки набора сейсмических данных

Визуализации результатов статистического анализа набора данных сейсмических событий представлены в соответствии с рисунком 54.

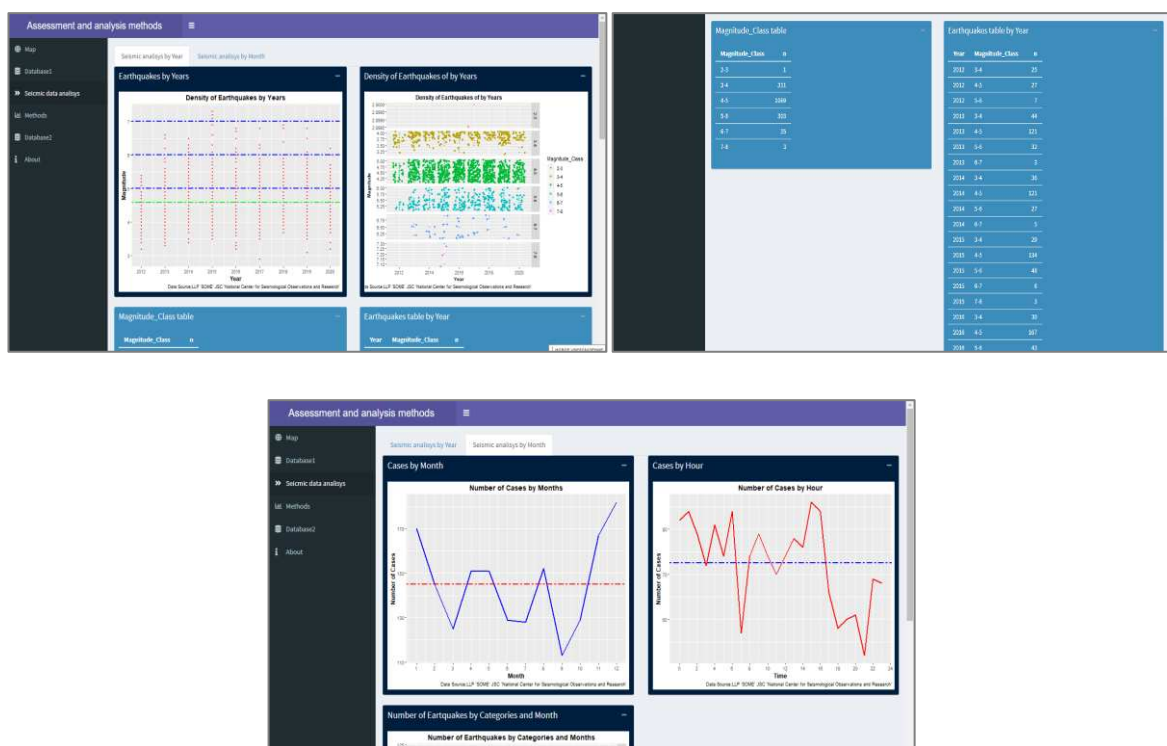


Рисунок 54 – Визуализации результатов статистического анализа данных

Краткие сведения о возможностях ИИАС об авторах и разработчике представлены в соответствии с рисунком 55.



Рисунок 55 – Окно с краткой аннотацией об ИИАС, авторах и разработчике

Таким образом, спроектированная архитектура ИИАС реализует разработанный способ оценки сейсмостойкости городских объектов с учетом сейсмических данных на основе методов интеллектуального анализа данных. На основе спроектированной архитектуры ИИАС реализована программная реализация системы, которая позволяет осуществить «быструю» оценку данных, повышая эффективность процессов обработки, анализа и оценки данных. Работоспособность ИИАС протестирована на исследованных наборах данных в соответствии с разработанным способом, что показывает точные результаты анализа наборов данных, согласующиеся с экспертными оценками.

Выводы по четвертому разделу

1. Представлено полное описание концептуальной основы и результатов экспериментальных исследований по разработанной интеллектуальной информационно-аналитической системе, осуществляющая оценку сейсмостойкости городских объектов с учетом сейсмических данных на основе технологии анализа данных и методов машинного обучения.

2. Обоснован выбор программного инструментария и информационной технологии для проектирования и реализации интеллектуальной информационно-аналитической системы оценки сейсмостойкости городских объектов с учетом сейсмических данных.

3. Разработана и спроектирована архитектура интеллектуальной информационно-аналитической системы оценки сейсмостойкости городских объектов с учетом сейсмических данных, а также подробно описаны концептуальные, логические модели ввода-вывода данных интеллектуальной информационно-аналитической системы и приведен алгоритм работы системы.

4. Работоспособность разработанной интеллектуальной информационно-аналитической системы была проверена на исследованных в диссертационной работе наборах данных по городским объектам и на примере данных сейсмических событий (землетрясений), зафиксированные станциями на территории РК. Результаты оценок сейсмостойкости городских объектов с учетом сейсмических событий в процессе опытной эксплуатации, согласуются с результатами экспертных оценок.

ЗАКЛЮЧЕНИЕ

В ходе выполнения диссертационной работы и проведенного исследования, выполнен ряд задач, позволяющих сделать следующие выводы. Исследованы и проанализированы текущее состояние и предпосылки в области современного сейсмостойкого строительства и сейсмической обстановки в целом, и в стране. Выполнен расширенный аналитический обзор и анализ существующих подходов и методик оценки сейсмостойкости городских объектов и анализа сейсмических данных, основывающихся на технологиях интеллектуального анализа данных и методах машинного обучения. Получены выводы о применимости и эффективности использования методов машинного обучения в решении задач по оценке сейсмической уязвимости и рисков. Кроме этого, также изучены существующие информационные технологии по оценке сейсмической уязвимости объектов и прогнозированию сейсмических опасностей.

В рамках диссертационной работы реализованы основные задачи по разработке способа оценки сейсмостойкости городских объектов с учетом сейсмических данных на основе технологии анализа данных и методов машинного обучения, а также разработана функциональная модель данного способа. Разработанный способ оценки сейсмостойкости городских объектов включает в себя технологию анализа данных и комплексное применение методов кластеризации и классификации. С помощью кластерного анализа данных для городских объектов получены 4 кластера. На основе кластерных решений методом Decision Tree построены деревья решений для оценки сейсмостойкости городского объекта. В структуре дерева решений методом Random Forest определены наиболее информативные признаки, влияющие на итоговое решение сейсмостойкости объекта. Оценка качества полученных решений показала высокий результат точности. В обработке и анализе сейсмических данных использован другой алгоритм машинного обучения, пространственная кластеризация на основе плотности с присутствием шума DBSCAN. Данный алгоритм позволил обнаружить участки с повышенной пространственной плотностью сейсмических событий. Важная особенность применения данного алгоритма, заключалась в подборе входных параметров алгоритма DBSCAN. Также для сейсмического набора данных применен статистический анализ данных с целью выявления и анализа количественных характеристик сейсмических событий. На основе разработанного способа оценки сейсмостойкости городских объектов с учетом сейсмических данных, спроектирована архитектура ИИАС, в структуре которой реализованы модули анализа наборов данных и алгоритмов машинного обучения. Разработанная архитектура позволяет повысить эффективность процессов масштабирования, обеспечения безопасности, гибкости системы, а также является удобным для сопровождения. Работоспособность разработанной ИИАС проверена на исследованных наборах данных и описаны результаты программной реализации алгоритмов машинного обучения кластеризации и классификации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 Президент Республики Казахстан – Лидер Нации Н.А. Назарбаев. Стратегия «Казахстан-2050»: новый политический курс состоявшегося государства: послание народу Казахстана // <https://www.primeminister.kz/ru/documents/gosprograms/stratplan-2025>. 14.12.2012.
- 2 Президент Республики Казахстан – Лидер Нации Н.А. Назарбаев. Новые возможности развития в условиях четвертой промышленной революции: послание народу Казахстана // <https://www.akorda.kz/ru/addresses>. 10.01.2018.
- 3 Постановление Правительства Республики Казахстан. Об утверждении Государственной программы жилищного строительства «Нұрлы жер»: утв. 22 июня 2018 года, №372 // <https://online.zakon.kz/>. 31.12.2019.
- 4 Жунусов Т.Ж. Землетрясение и сейсмостойкое строительство: монография. – Алматы: ЛЕМ, 2008. – 74 с.
- 5 Жунусов Т.Ж., Шахнович Ю.Г., Горовиц И.Г. и др. Исследования каркасного здания на сейсмоизолирующих опорах // Бетон и железобетон. – 1985. – №7. – С. 42-43.
- 6 Абаканов М.С. К вопросу применения свайных фундаментов с высоким ростверком в сейсмостойком строительстве // Матер. междунар. науч.-практ. конф. «Новые строительные тренды в XXI веке». – Алматы, 2017. – С. 27-35.
- 7 Абаканов М.С., Тулеев А.Т., Тулеев Т.Д. Прочность железобетонных изгибаемых элементов по наклонному сечению без поперечного армирования по расчетным методикам СП РК EN 1992-1-1:2004/2011 и СНиП 2.03.01-84* // Вестник АО «КазНИИСА». – 2019. – №3(891). – С. 16-24.
- 8 Лапин В.А., Ержанов С.Е. Проблемы определения сейсмического риска для населенных пунктов Республики Казахстан // Вестник АО «КазНИИСА». – 2016. – №7. – С. 20-24.
- 9 Ержанов С.Е., Лапин В.А. Системы сейсмоизоляции зданий и сооружений в Республике Казахстан и развитых стран мира // Исследования сейсмостойкости сооружений и конструкций. – 2015. – №23(33). – С. 193-208.
- 10 Курскеев А.К., Тимуш А.В., Сыдыков А.С. и др. Общее сейсмическое районирование территории Казахстана // Карта ОСР территории Казахстана: объяснительная записка. – Алма-Ата, 2003. – С. 84.
- 11 Абаканов Т.Д., Бегалинов А.Б., Абаканов А.Т. Оценка сейсмостойкости подземных сооружений водосбросов Капчагайской ГЭС // Гидротехническое строительство. – 2016. – №2. – С. 54-60.
- 12 Nuguzhinov Zh., Vatin N., Bakirov Zh. et al. Stress-strain state of bending reinforced beams with cracks // Magazine of Civil Engineering. – 2020. – Vol. 97(5). – P. 1-15.
- 13 Постановление Правительства Республики Казахстан. Об утверждении Государственной программы «Цифровой Казахстан»: утв. 12 декабря 2017 года, №827 (с изм. и доп. по состоянию на 29.07.2019 г.) // https://online.zakon.kz/document/?doc_id=3716805. 24.08.2019.

14 Карменова М.А., Попова Г.В. Анализ современных программных комплексов по расчету зданий и сооружений на сейсмические воздействия. World Science: Problems and Innovations: сб. ст. – Пенза: МЦНС «Наука и Просвещение». – 2018. – С. 111-115.

15 Карменова М.А., Попова Г.В. Методики получения 3D-моделей городских объектов для информационного моделирования // Творчество молодых – инновационному развитию Казахстана: матер. 5-й междунар. научн.-техн. конф. студен., магистр. и молод. уч. – Усть-Каменогорск, 2019. – С. 100-103.

16 Karmenova M., Nugumanova A., Tlebaldinova A. et al. Seismic Assessment of Urban Buildings Using Data Mining Methods // полное название конференции (ICSTA). – Antalya, 2020. – P. 154-159.

17 Карменова М.А., Тлебалдинова А.С. Обзор подходов оценки сейсмоустойчивости объектов городского типа // Творчество молодых – инновационному развитию Казахстана: матер. 6-й междунар. научн.-техн. конф. студен., магистр. и молод. уч. – Усть-Каменогорск, 2020. – С. 108-112.

18 Karmenova M., Tlebaldinova A., Madiyarov M. et al. Spatial Analysis of the Earthquakes Distribution in Kazakhstan for Risk Mapping // Вычислительные технологии. – 2020. – №3, ч. 1. – С. 149-157.

19 Laasri H.A., Akhouayri E.-S., Agliz D. et al. A fuzzy expert system for automatic seismic signal classification // Expert Syst. Appl. – 2015. – Vol. 42. – P. 1013-1027.

20 Andrić J.M., Lu D.-G. Fuzzy probabilistic seismic hazard analysis with applications to Kunming city, China // Nat. Hazards. – 2017. – Vol. 89. – P. 1031-1057.

21 Asim K.M., Awais M., Martínez-Álvarez F. et al. Seismic activity prediction using computational intelligence techniques in northern Pakistan // Acta Geophys. – 2017. – Vol. 65. – P. 919-930.

22 Asim K.M., Martínez-Álvarez F., Basit A. et al. Earthquake magnitude prediction in Hindukush region using machine learning techniques // Nat. Hazards. – 2016. – Vol. 85. – P. 471-486.

23 Абакаров А.Д., Курбанов И.Б. К оценке сейсмического риска территорий // Вестник Дагестанского государственного технического университета. – 2014. – №1(32). – С. 68-77.

24 Мажиев Х.Н. Материалы и конструкции для повышения сейсмостойкости зданий и сооружений: дис. ... док. техн. наук: 05.23.05, 05.23.01. – Махачкала, 2011. – 555 с.

25 Omori F. Seismic Experiments on the Fracturing and Overturning of Columns // In book: Publ. Earthquake Invest. Comm. in Foreign Languages. – Tokyo, 1900. – 37 p.

26 Сеитов Б.М., Ордобаев Б.С. Сейсмостойкость зданий и сооружений: практикум. – Бишкек: Изд-во КРСУ, 2016. – 96 с.

27 Акатьев В.А., Метелкин Е.В., Нигметов Г.М. Методы и средства оценки и снижения сейсмического риска: обзор // Безопасность в техносфере. –

2016. – Т. 5, №5. – С. 75-86.

28 Ковальчук С.В. Интеллектуальные технологии создания, исследования и применения композитных моделей сложных систем: дис. ... док. техн. наук: 05.13.18. – СПб., 2018. – 270 с.

29 Tlebaldinova A., Nugumanova A., Karmenova M. et al. Temporal Network Approach to Explore Bike Sharing Usage Patterns // Proceed. of the 6th internat. conf. on Vehicle Technology and Intelligent Transport Systems (VEHITS, 2020). – Prague, 2020. – P. 129-136.

30 He M., Li Q., Li X. Injection-Induced Seismic Risk Management Using Machine Learning Methodology – A Perspective Study // Front. Earth Sci. – 2020. – Vol. 8. – P. 1-18.

31 Menga T., Jinga X., Yanab Zh. et al. A survey on machine learning for data fusion // Information Fusion. – 2020. – Vol. 57. – P. 115-129.

32 Riedel I., Guéguen Ph., Mura M.D. et al. Seismic vulnerability assessment of urban environments in moderate-to-low seismic hazard regions using association rule learning and support vector machine methods // Natural Hazards. – 2015. – Vol. 76(2). – P. 1111-1141.

33 Карменова М.А., Тлебалдинова А.С., Жантасова Ж.З. Қалалық объектілердің 3d үлгілерін құруға арналған үш өлшемді фотограмметрия. // Вестник КазНТУ. – 2020. – №1(137). – С. 322-328.

34 Zhang Z., Hsu T.-Y., Wei H.-H. et al. Development of a Data-Mining Technique for Regional-Scale Evaluation of Building Seismic Vulnerability // Applied Sciences. – 2019. – Vol. 9(7). – P. 1-17.

35 Chen C.S., Cheng M.Y., Wu Y.W. Seismic assessment of school buildings in Taiwan using the evolutionary support vector machine inference system // Expert Systems with Applications. – 2012. – Vol. 39(4). – P. 4102-4110.

36 Chen H.M., Kao W.K., Tsai H.C. Genetic programming for predicting aseismic abilities of school buildings // Engineering Applications of Artificial Intelligence. – 2012. – Vol. 25(6). – P. 1103-1113.

37 Kao W.K., Chen H.M., Chou J. Aseismic ability estimation of school building using predictive data mining models // Expert Systems with Applications. – 2011. – Vol. 38. – P. 10252-10263.

38 Liu Y., Li Z., Wei B. et al. Seismic vulnerability assessment at urban scale using data mining and GIScience technology: application to Urumqi (China) // Geomatics, Natural Hazards and Risk. – 2019. – Vol. 10(1). – P. 958-985.

39 Ortega J., Vasconcelos G., Rodrigues H. et al. Development of a Numerical Tool for the Seismic Vulnerability Assessment of Vernacular Architecture // Journal of Earthquake Engineering. – 2019. – Vol. 25(1). – P. 2926-2954.

40 Brando G., De Matteis G., Spacone E. Predictive model for the seismic vulnerability assessment of small historic centres: application to the inner Abruzzi Region in Italy // Engineering Structures. – 2017. – Vol. 153. – P. 81-96.

41 Drago C., Ferlito R., Zucconic M. Clustering of damage variables for masonry buildings measured after L'Aquila earthquake // SIS, Statistics and Demography: the Legacy of Corrado Gini (Submitted). – 2015. – Vol. 3. – P. 1-6.

42 Irwansyah E., Winarko E. Spatial Data Clustering and Zonation of Earthquake Building Damage Hazard Area // The European Physical Journal Conferences. – 2014. – Vol. 68. – P. 1-6.

43 Guettiche A., Gueguen Ph. Seismic vulnerability assessment using association rule learning: application to the city of Constantine, Algeria // Natural Hazards. – 2017. – Vol. 86(3). – P. 1223-1245.

44 Riedel I., Gueguen P., Dunand F. et al. Macroscale vulnerability assessment of cities using association rule learning // Seismol Res Lett. – 2014. – Vol. 85(2). – P. 295-305.

45 Sari D.P., Rosadi D., Effendie A.R. et al. Application of Bayesian network model in determining the risk of building damage caused by earthquakes // Procced. Internat. con. on Information and Communications Technology. – Yogyakarta, 2018. – P. 131-135.

46 Sari D.P., Rosadi D., Effendie A.R. et al. K-means and bayesian networks to determine building damage levels // Computer Science. – 2019. – Vol. 17(2). – P. 719-727.

47 Zhang R., Chen Z., Chen S. et al. Deep long short-term memory networks for nonlinear structural seismicresponse prediction // Computers & Structures. – 2019. – Vol. 220(3). – P. 55-68.

48 Gökkaya K. Geographic analysis of earthquake damage in Turkey between 1900 and 2012 // Geomatics, Natural Hazards and Risk. – 2016. – Vol. 7(6). – P. 1948-1961.

49 Leon F., Atanasiu G.M., Gâlea D. Using Data Mining Techniques for the Management of Seismic Vulnerability // Key Engineering Materials. – 2006. – Vol. 326-328. – P. 501-504.

50 Istyaq S., Riaz M.M. A Clustering Method For Seismic Zone Identification And Spatial Data Mining // International Journal of Advanced Research in Computer Science Engineering and Information Technology. – 2013. – Vol. 1. – P. 66-88.

51 Karri N.A., Ansari M.Y., Pathak A. Identification of Seismic Zones of India using DBSCAN // Procced. internat. conf. Computing, Power and Communication Technologies (IEEE. 2018). – Greater Noida, 2018. – P. 65-69.

52 Edelani R., Barakbah A.R., Harsono T. et al. A. Association analysis of earthquake distribution in Indonesia for spatial risk mapping // Procced. internat. electron. sympos. on Knowledge Creation and Intelligent Computing (IEEE 2017). – Surabaya, 2017. – P. 231-238.

53 Yang P., Yin X., Zhang G. Seismic Data Analysis Based on Fuzzy Clustering // Procced. 8th internat. conf. on Signal Processing (IEEE, 2006). – Beijing, 2006. – P. 1-4.

54 Shodiq M.N., Kusuma D.H., Rifqi M.G. et al. Spatial analisys of magnitude distribution for earthquake prediction using neural network based on automatic clustering in Indonesia // Procced. internat. electron. sympos. on Knowledge Creation and Intelligent Computing (IEEE 2017). – Surabaya, 2017. – P. 246-251.

55 Ahumada A., Altunkaynak A., Ayoub A. Fuzzy logic-based attenuation

relationships of strong motion earthquake records // *Expert Syst. Appl.* – 2015. – Vol. 42. – P. 1287-1297.

56 Aboonasr S.F.G., Zamani A., Razavipour F. et al. Earthquake hazard assessment in the Zagros Orogenic Belt of Iran using a fuzzy rule-based model // *Acta Geophys.* – 2017. – Vol. 65. – P. 589-605.

57 Laasri H.A., Akhouayri E.-S., Agliz D. et al. A fuzzy expert system for automatic seismic signal classification // *Expert Syst. Appl.* – 2015. – Vol. 42. – P. 1013-1027.

58 Hossain M.S., Al Hasan A., Guha S. et al. A Belief Rule Based Expert System to Predict Earthquake under Uncertainty // *J. Wirel. Mob. Netw. Ubiquitous Comput. Dependable Appl.* – 2018. – Vol. 9. – P. 26-41.

59 Ikram A., Qamar U. A rule-based expert system for earthquake prediction // *J. Intell. Inf. Syst.* – 2014. – Vol. 43. – P. 205-230.

60 Asim K.M., Moustafa S.S., Niaz I.A. et al. Seismicity analysis and machine learning models for short-term low magnitude seismic activity predictions in Cyprus // *Soil Dyn. Earthq. Eng.* – 2020. – Vol. 130. – P. 1-11.

61 Vasti M., Dev A. Classification and Analysis of Real-World Earthquake Data Using Various Machine Learning Algorithms // In book: *Lecture Notes in Electrical Engineering.* – Singapore: Springer, 2019. – P. 1-14.

62 Mukhopadhyay U.K., Sharma R.N.K., Anwar S., Dutta A.D. Correlating Thermal Anomaly with Earthquake Occurrences Using Remote Sensing // Mukhopadhyay U.K., Sharma R.N.K., Anwar S., Dutta A.D. *Correlating Thermal Anomaly with Earthquake Occurrences Using Remote Sensing* // ??????. – Berlin; Heidelberg: Springer, 2019. – P. 863-875.

63 Karimzadeh S., Matsuoka M., Kuang J. et al. Spatial Prediction of Aftershocks Triggered by a Major Earthquake: A Binary Machine Learning Perspective // *ISPRS Int. J. Geo-Inf.* – 2019. – Vol. 8. – P. 1-13.

64 Zhou Z., Lin Y., Zhang Z. et al. Earthquake Detection in 1D Time-Series Data with Feature Selection and Dictionary Learning // *Seism. Res. Lett.* – 2019. – Vol. 90. – P. 563-572.

65 Corbi F., Sandri L., Bedford J. et al. Machine Learning Can Predict the Timing and Size of Analog Earthquakes // *Geophys. Res. Lett.* – 2019. – Vol. 46. – P. 1303-1311.

66 Kong Q., Trugman D.T., Ross Z.E. et al. Machine Learning in Seismology: Turning Data into Insights // *Seism. Res. Lett.* – 2018. – Vol. 90. – P. 3-14.

67 Galkina A., Gafeeva N. Machine Learning Methods for Earthquake Prediction: A Survey // *Proced. of the 4th conf. on Software Engineering and Information Management (SEIM-2019).* – SPb., 2019. – P. 25-32.

68 Gitis V.G., Derendyaev A. Machine Learning Methods for Seismic Hazards Forecast // *Geosciences.* – 2019. – Vol. 9. – P. 308.

69 Al-Najjar H.A.H., Kalantar B., Pradhan B. et al. Conditioning factor determination for mapping and prediction of landslide susceptibility using machine learning algorithms // In book: *Earth Resources and Environmental Remote Sensing/GIS Applications X.* – Strasbourg, 2019. – Vol. 11156. – P. 1-13.

70 Ganter T., Sundermier A., Ballard S. Alternate Null Hypothesis Correlation: A New Approach to Automatic Seismic Event Detection // *Bull. Seism. Soc. Am.* – 2018. – Vol. 108. – P. 3528-3547.

71 Mosavi A., Salimi M., Ardabili S.F. et al. State of the Art of Machine Learning Models in Energy Systems, a Systematic Review // *Energies.* – 2019. – Vol. 12. – P. 1-42.

72 Mosavi A., Ozturk P., Chau K.-W. Flood Prediction Using Machine Learning Models: Literature Review // *Water.* – 2018. – Vol. 10. – P. 1-44.

73 Dineva A., Mosavi A., Ardabili S.F. et al. Review of Soft Computing Models in Design and Control of Rotating Electrical Machines // *Energies.* – 2019. – Vol. 12. – P. 1-28.

74 Tehseen R., Farooq Sh. et al. Earthquake Prediction Using Expert Systems: A Systematic Mapping Study // *Sustainability.* – 2020. – Vol. 12(6). – P. 1-32.

75 Mirrashid M. Earthquake magnitude prediction by adaptive neuro-fuzzy inference system (ANFIS) based on fuzzy C-means algorithm // *Nat. Hazards.* – 2014. – Vol. 74. – P. 1577-1593.

76 Abayon R.C., Apilado J.R., Pacis D.B. et al. A Weather Prediction and Earthquake Monitoring System // *Proced. of the conf. on Systems, Process and Control (ICSPC).* – Malacca, 2018. – P. 203-208.

77 Tosunoglu N.G., Apaydin A. A New Spatial Algorithm Based on Adaptive Fuzzy Neural Network for Prediction of Crustal Motion Velocities in Earthquake Research // *Int. J. Fuzzy Syst.* – 2018. – Vol. 20. – P. 1656-1670.

78 Shibli M. A novel approach to predict earthquakes using adaptive neural fuzzy inference system and conservation of energy-angular momentum // *Int. J. Comp. Inf. Syst. Ind. Manag. Appl.* – 2011. – Vol. 2150. – P. 371-390.

79 Torres V.M., Castillo O. A Type-2 Fuzzy Neural Network Ensemble to Predict Chaotic Time Series // *In book: Studies in Computational Intelligence.* – Berlin; Heidelberg: Springer, 2015. – Vol. 601. – P. 185-195.

80 Shi P., Seydoux L., Poli P. Unsupervised Learning of Seismic Wavefield Features: Clustering Continuous Array Seismic Data During the 2009 L'Aquila Earthquake // *Journal of Geophysical Research: Solid Earth.* – 2021. – Vol. 126 – P. 1-15.

81 Seydoux L., Balestriero R., Poli P. et al. Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning // *Nature Communications.* – 2020. – Vol. 11(1) – P. 1-12.

82 Jozinović D., Lomax A. et al. Rapid prediction of earthquake ground shaking intensity using raw waveform data and a convolutional neural network // *Geophysical Journal International.* – 2020. – Vol. 222, Issue 2. – P. 1379-1389.

83 Asim K.M, Idris A., Iqbal T. et al. Earthquake prediction model using support vector regressor and hybrid neural networks // *PLoS One.* – 2018. – Vol. 13(7). – P. 1-22.

84 Uchide T. Focal mechanisms of small earthquakes beneath the Japanese islands based on first-motion polarities picked using deep learning // *Geophysical Journal International.* – 2020. – Vol. 223, Issue 3. – P. 1658-1671.

- 85 Chen Y. Automatic microseismic event picking via unsupervised machine learning // *Geophysical Journal International*. – 2018. – Vol. 212, Issue 1. – P. 88-102.
- 86 Asim K.M., Idris A. et al. Seismic indicators based earthquake predictor system using Genetic Programming and AdaBoost classification // *Soil Dynamics and Earthquake Engineering*. – 2018. – Vol. 111. – P. 1-7.
- 87 Cesca S. Seiscloud, a tool for density-based seismicity clustering and visualization // *J Seismol.* – 2020. – Vol. 24. – P. 443-457.
- 88 Shang X., Li X. et al. Means Cluster for Seismicity Partitioning and Geological Structure Interpretation, with Application to the Yongshaba Mine (China) // *Shock and Vibration* – 2017. – Vol. – P. 1-12.
- 89 Fan Z., Xu X. Application and visualization of typical clustering algorithms in seismic data analysis // *Procedia Computer Science*. – 2019. – Vol. 151. – P. 171-178.
- 90 Liu G., Li R. et al. Scaling relation of earthquake seismic data // *Physica A: Statistical Mechanics and its Applications*. – 2018. – Vol. 492. – P. 2092-2102/
- 91 Lindsey N.J., Martin E.R., Dreger D.S. et al. Fiber-Optic Network Observations of Earthquake Wavefields // *Geophysical Research Letters*. – 2017. – Vol. 44, Issue 23. – P. 11792-11799.
- 92 Langenbruch C., Weingarten M., Zoback M.D. Physics-based forecasting of man-made earthquake hazards in Oklahoma and Kansas // *Nat Commun*. – 2018. – Vol. 9. – P. 1-10.
- 93 Гордеев Е.И., Чебров В.Н., Дрознин Д.В. и др. Сбор, обработка и хранение сейсмологической информации. // *Комплексные сейсмологические и геофизические исследования Камчатки к 25-летию Камчатского филиала ГС РАН – Петропавловск-Камчатский: Геофиз. служба РАН, 2004. – С. 43-61.*
- 94 Berrais A. Using Expert Systems Technology to Teach Earthquake Resistant Design of Buildings / 2000 Annual Conference. – St. Louis, Missouri, 2000. – P. 5692-1-5692-9.
- 95 Website of Opensees // <https://opensees.berkeley.edu/>. May 24, 2021.
- 96 McKenna F. OpenSees: A Framework for Earthquake Engineering Simulation // *Computing in Science and Engineering*. – 2011. – Vol. 13(4). – P. 58-66.
- 97 Psyrrasa N.K., Sextos A.G. Build-X: Expert system for seismic analysis and assessment of 3D buildings using OpenSees // *Advances in Engineering Software*. – 2018. – Vol. 116. – P. 23-35.
- 98 Berrais A., Watson A.S. Expert systems for seismic engineering: the state-of-the-art // *Engineering Structures*. – 2003. – Vol. 15, Issue 3. – P. 146-154.
- 99 Assatourians K., Atkinson G.M. EqHaz: An Open-Source Probabilistic Seismic-Hazard Code Based on the Monte Carlo Simulation Approach // *Seismological Research Letters*. – 2013. – Vol. 84(3). – P. 516-524.
- 100 Borzi B., Onida M., Faravelli M. et al. IRMA platform for the calculation of damages and risks of Italian residential buildings // *Bulletin of Earthquake Engineering*. – 2021. – Vol. 19. – P. 3033-3055.

- 101 Cimellaro G.P., Marasco S. A computer-based environment for processing and selection of seismic ground motion records: OPENSIGNAL // *Front. Built Environ.* – 2015. – Vol. 1. – P. 1-13.
- 102 Website of RadExPro is a seismic processing software // <https://radexpro.com/ru/>. 14 April, 2021.
- 103 Papazafeiropoulos G., Plevris V. OpenSeismoMatlab: A new open-source software for strong ground motion data processing // *Heliyon.* – 2018. – Vol. 4. – P. 1-40.
- 104 Jambalos Ar-Ar S., Kyle R. et al. Seismic Vulnerability Assessment of San Agustin Church in Intramuros – Manila Using Fragility Curves // *Second Annual Intramuros Young Scholars Conference.* – 2021. – Vol. 156. – P. 55-76.
- 105 Jones J., Kalkan E., Stephens C. et al. PRISM software: processing and Review Interface for strong-motion data // *Seismological Research Letters.* – 2017. – Vol. 88(3). – P. 851-866.
- 106 Barclay W. Ogden. PRISM: Software for Risk Assessment and Decision-Making in Libraries // *Collections: A Journal for Museum and Archives Professionals.* – 2012. – Vol. 8(4). – P. 323-329.
- 107 Kwiatkowska M., Norman G., Parker D. PRISM 4.0: Verification of Probabilistic Real-Time Systems // *Lecture Notes in Computer Science.* – 2011. – Vol. 6806. – P. 585-591.
- 108 Kashima T. Viewwave 2.2.0 Manual. – Tsukuba. – January 2016. – P. 1-65.
- 109 Серебряков М.А. Исследование архитектур построения и алгоритмов работы систем хранения и анализа больших данных: магистр. дис. – Тольятти, 2019. – 97 с.
- 110 Аналитическая компания IDC // <http://idc.com>ю 15.03.2019.
- 111 Greenfield D. How Technology and People Are Steering Manufacturing's Future David Greenfield // <https://www.automationworld.com/how-technology-and-people-aresteering-manufacturings-future>. 27.05.2019.
- 112 Пальмов С.В. Исследование возможностей аналитической системы на основе метода машинного обучения // *Радиопромышленность.* – 2020. – Т. 30, №3. – С. 112-126.
- 113 Замятин И.В. Использование технологий машинного обучения в операционной деятельности производственной компании: проблемы и возможности // *Материалы: матер. междунар. науч.-практ. конф.* – М.: Перо, 2020. – С. 350-357.
- 114 Деловой портал // [https://www.tadviser.ru/index.php/Статья:Большие_данные_\(Big_Dat\)](https://www.tadviser.ru/index.php/Статья:Большие_данные_(Big_Dat)). 2017.10.24.
- 115 Мамедова Э.М. Исследование Data Mining - технологий, основанных на кластерном анализе: магист. ... дис. – Баку, 2018. – 81 с.
- 116 Alizadeh M., Ngah I., Hashim M. et al. Hybrid analytic network process and Artificial neural network (anp-ann) model // *Remote Sens.* – 2018. – Vol. 10. – P. 1-34.
- 117 Al-Heety E.A.M. Spatial Analysis of earthquakes in Iraq using statistical

and Data Mining techniques // Iraqi Geological Journal. – 2016. – Vol. 39-49, №2. – P. 1-15.

118 Nivedhitha U., Krishna A.S. Development of a Predictive System for Anticipating Earthquakes using Data Mining Techniques for Urban Earthquake Vulnerability Assessment // Indian Journal of Science and Technology. – 2016. – Vol. 9(48). – P.1-7.

119 Shang X., Li X., Morales-Esteban A. et al. Data Field-Based K-Means Clustering for Spatio-Temporal Seismicity Analysis and Hazard Assessment // Remote Sens. – 2018. – Vol. 10. – P. 1-22.

120 Sohrab M.Md. et al. Improvement of K-means Clustering algorithm with better initial centroids based on weighted average // Proceed. 7th internat. conf. on Electrical and Computer Engineering. – Dhaka, 2012. – P. 647-650.

121 Deelers S. et al. Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance // International Journal of Computer Science. – 2007. – Vol. 2, №4. – P. 247-252.

122 Wójcik W., Karmenova M., Smailova S. et al. Development of data-mining technique for seismic vulnerability assessment // Journal of Electronics and Telecommunications. – 2021. – Vol. 67, Issue 2. – P. 261-266.

123 Margaret H Dunham. Data Mining-Introductory and Advanced Concepts, Pearson Education, 2006. – 328 p.

124 Карменова М.А., Нугуманова А.Б., Тлебалдинова А.С. Кластерный анализ данных в решении задач по оценке сейсмической уязвимости объектов городской среды // Вестник АУЭС. – 2020. – №1(48). – С. 102-109.

125 Lance G.N., Willams W.T. A general theory of classification sorting strategies. 1. Hierarchical systems // Computer J. – 1967. – №9. – P. 373-380.

126 Kasyanov V.N., Evstigneev V.A. Graphs in programming: processing, visualization and application. – SPb.: BHV-Petersburg, 2003. – 1104 p.

127 Tan P.J., Dowe D.L. MML inference of decision graph with multi-way and dynamic attributes // Lecture Notes in Computer Science. – Perth. – 2003. – Vol. 2903. – P. 269-281.

128 Quinlan J.R. Induction of decision trees // Mach Learn. – 1986. – Vol. 1(1). – P. 81-106.

129 Khoshnevis N., Taborda R. Prioritizing ground-motion validation metrics using semisupervised and supervised learning // Bull Seismol Soc Am. – 2018. – Vol. 108(4). – P. 2248-2264.

130 Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. – 2006. – Vol. 27, Issue 8. – P. 861-874.

131 Кажемский М.А., Шелухин О.И. Многоклассовая классификация сетевых атак на информационные ресурсы методами машинного обучения // Труды учебных заведений связи. – 2019. – Т. 5, №1. – С. 107-115.

132 Thomas C.W. Landgrebe R., Duin P.W. Approximating the multiclass ROC by pairwise analysis // Pattern Recognition Letters. – 2007. – Vol. 28. – P. 1747-1758.

133 Хрящев В.В., Приоров А.Л., Стефаниди А.Ф. и др. Разработка и исследование алгоритмов обработки и распознавания речевых сигналов и изображений для систем мультимодальной биометрии // Цифровая Обработка Сигналов. – 2017. – №3. – С. 45-49.

134 Prajwala T.R. A Comparative Study on Decision Tree and Random Forest Using R Tool // International Journal of Advanced Research in Computer and Communication Engineering. – 2015. – Vol. 4, Issue 1. – P. 196-199.

135 Kabir E., Guikema S., Kane B. Statistical Modeling of Tree Failures during Storms // Reliability Engineering & System Safety. – 2018. – Vol. 177. – P. 68-79.

136 Gehl P., Cavalieri F., Franchin P. Approximate Bayesian Network Formulation for the Rapid Loss Assessment of Real-World Infrastructure Systems // Reliability Engineering & System Safety. – 2018. – Vol. 177. – P. 80-93.

137 Breiman L. Random forests // Machine Learning. – 2001. – Vol. 45(1). – P. 5-32.

138 Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – California: Springer-Verlag, 2009. – 746 p.

139 Pal M. Random forest classifier for remote sensing classification // International Journal of Remote Sensing. – 2015. – Vol. 26, Issue 1. – P. 217-222.

140 Breiman L., Friedman R., Olshen R. et al. Classification and Regression Trees. – Belmont, California: Wadsworth International, 1984. – 342 p.

141 Quinlan J.R. Simplifying decision trees // International Journal of ManMachine Studies. – 1987. – Vol. 27. – P. 221-234.

142 Туманов А. В. Анализ сейсмологических данных для строительства / А. В. Туманов // Модели, системы, сети в экономике, технике, природе и обществе. – 2015. – № 3 (15). С. 151–158.

143 Информационный сайт // <https://works.doklad.ru/view/8rbCaaqNd5o.html>. 05.03.2020.

144 Буртиев Р.З. Методология оценки сейсмической опасности на основе вероятностных моделей сейсмичности: геофизика: 04.00.22 – Киев, 2017. – 322 с.

145 Сыдыков А. Сейсмический режим и прогнозирование сейсмической опасности в Казахстане: дис. Геофизика, геофизические методы поисков полезных ископаемых: 25.00.10. – Алматы, 2002. – 291 с.

146 Карменова М.А., Глебалдинова А.С., Смаилова С.С., Попова Г.В. Применение метода пространственной кластеризации для сейсмических данных // Вестник ВКТУ им. Д. Серикбаева. – 2020. – №4(90). – С. 36-40.

147 Сайт ТОО «СОМЭ» АО «Национальный центр сейсмологических наблюдений и исследований» // <http://some.kz/index.php?p=card&outlang=1&oid=9>. 2011.02.15.

148 Mukambayev A.S., Mikhaylova N.N. Resheniye problemy neodnorodnosti magnitud v rabotakh po seysmicheskomu zonirovaniyu territorii Respubliki Kazakhstan // Vestnik NYATS RK. – 2014. – Вып. 4. – P. 86-92.

149 Mikhailova N.N., Neverova N.P. Calibration function $s(d)$ for

determining the MPVA of the Northern Tien Shan earthquakes // Comprehensive research at the Alma-Ata prognostic site: col. of articles. – Alma-Ata: Nauka, 1986. – P. 41-47.

150 Bulut T. Analysis of Turkey Earthquake Data // https://rstudio-pubs-static.s3.amazonaws.com/572775_e243d7dd04c641399c8f30affdfa132c.html. 2020-02-08.

151 Schubert E., Sander J., Ester M. et al. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN // ACM Transactions on Database Systems. – 2017. – Vol. 42(3). – P. 1-21.

152 Yu Y., Jia Z., Cao L. et al. Fast density-based clustering algorithm for location big data // Journal of Software. – 2018. – Vol. 29(8). – P. 2470-2484.

153 Dauni P., Firdaus M.D., Asfariani R. et al. Implementation of Haversine formula for school location tracking // Journal of Physics: Conference Series. – 2019. – Vol. 1402, Issue 7. – P. 1-6.

154 Basyir M., Nasir M.S., Mellyssa W. 2018 Determination of Nearest Emergency Service Office using Haversine Formula Based on Android Platform // Emit Int J Eng Technol. – 2018. – Vol. 5, Issue 2. – P. 270-278.

155 Khairina D.M., Ramadhinata F.W., Hatta H.R. 2017 Pencarian Lokasi Jalur Nugraha Ekakurir (JNE) Terdekat Menggunakan Haversine Formula (Studi Kasus Kota Samarinda) // Pros SENIATI. – 2017. – Vol. 3. – P. 1-5.

156 Ganesh L., Kumar D.V. Indoor Wireless Localization using Haversine Formula // Int Adv Res J Sci Eng Technol. – 2015. – Vol. 2, Issue 7. – P. 59-63.

157 Hartanto S., Furqan M., Putera A. et al. 2017 Haversine Method in Looking for the Nearest Masjid // Int J Recent Trends Eng Res. – 2017. – Vol. 3, Issue 8. – P. 187-195.

158 Amalia A., Harmoko U., Yuliyanto G. Clustering of seismicity in the Indonesian Region for the 2018-2020 Period using the DBSCAN Algorithm // Journal of Physics and Its Applications. – 2021. – Vol. 4(1). P. 1-6.

159 Yoon C.E., Bergen K.J., Rong K. et al. Unsupervised Large-Scale Search for Similar Earthquake Signals // Bulletin of the Seismological Society of America. – 2019. – Vol. 109(4). – P. 1451-1468.

160 Kazemi-Beydokhti M., Abbaspour R.A. & Mojarab M. Spatio-Temporal Modeling of Seismic Provinces of Iran Using DBSCAN Algorithm. Pure Appl. Geophys., 174, 1937–1952 (2017). <https://doi.org/10.1007/s00024-017-1507-0>.

161 Лаборатория неотектоники и современной геодинамики // http://neotec.ginras.ru/neomaps/M050_Kazakhstan_2003_Seism-region_Karta-obshchego-seysmicheskogo-rayonirovaniya-respubliki-kazahstan.html. 2015-2021.

162 Yariyan P., Zabihi H., Wolf I.D. et al. Earthquake risk assessment using an integrated Fuzzy Analytic Hierarchy Process with Artificial Neural Networks based on GIS: A case study of Sanandaj in Iran // International Journal of Disaster Risk Reduction. – 2020. – Vol. 50. – P. 1-62.

163 Somodevilla M.J., Priego A.B., Castillo E. et al. Decision support system for seismic risks // Journal of Computer Science and Technology. – 2012. – Vol. 12, №2. – P. 71-77.

- 164 Website of MongoDB // <http://www.mongodb.org>. June 15, 2015.
- 165 Borthakur Dh. HDFS architecture guide // Hadoop Apache Project. – 2008. – Vol. 53. – P. 1-13.
- 166 Website of Sparklyr // <http://spark.rstudio.com/>. June 15th, 2021.
- 167 Lyakhov.KZ – Большая энциклопедия Казнета // https://vernoye-almaty.kz/nature/1887_verny.shtml. 25.09.2003.
- 168 Mangalathu S., Sun H., Nweke C.C. et al. Classifying earthquakedamage to buildings usingmachine learning // Earthquake Spectra. – 2019. – Vol. 36(1). – P. 1-39.
- 169 Hwang S.-H., Mangalathu S., Shin J. et al. Machine learning-based approaches for seismic demand and collapse of ductile reinforced concrete building frames // Journal of Building Engineering. – 2021. – Vol. 34. – P. 1-38.
- 170 Bass L., Clements P., Kazman R. Software Architecture in Practice. – Pittsburgh: Addison-Wesley, 25 сент. 2012 г. - Всего страниц: 624.

ПРИЛОЖЕНИЕ А

Свидетельство о государственной регистрации

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН



СВИДЕТЕЛЬСТВО
О ВНЕСЕНИИ СВЕДЕНИЙ В ГОСУДАРСТВЕННЫЙ РЕЕСТР
ПРАВ НА ОБЪЕКТЫ, ОХРАНЯЕМЫЕ АВТОРСКИМ ПРАВОМ
№ 14910 от «5» февраля 2021 года

Фамилия, имя, отчество, (если оно указано в документе, удостоверяющем личность) автора (ов):
**Т.ЛЕБА.ЛИНОВА АЙЖАН СОЛТАНГАЛIEВНА, КАРМЕНОВА МАРХАБА АХМЕТО.ЛИНОВНА,
УВАЛИЕВА ИНДИРА МАХМУТОВНА**

Вид объекта авторского права: **программа для ЭВМ**

Название объекта: **Web-приложение «Визуализация сейсмических данных»**

Дата создания объекта: **14.09.2020**





Құжат түпнұсқасының <http://www.kazpatent.kz/ru> сайтының
"Авторлық құқық" бөлімінде тексеруге болады <https://copyright.kazpatent.kz>
Подлинность документа возможно проверить на сайте [kazpatent.kz](http://www.kazpatent.kz)
в разделе «Авторское право» <https://copyright.kazpatent.kz>

Подписано ЭЦП

Оспанов Е.К.

Оценка сейсмостойкости городских объектов на основе технологии анализа данных (программа для ЭВМ)

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

СВИДЕТЕЛЬСТВО
О ВНЕСЕНИИ СВЕДЕНИЙ В ГОСУДАРСТВЕННЫЙ РЕЕСТР
ПРАВ НА ОБЪЕКТЫ, ОХРАНЯЕМЫЕ АВТОРСКИМ ПРАВОМ
№ 15298 от «18» февраля 2021 года

Фамилия, имя, отчество, (если оно указано в документе, удостоверяющем личность) автора (ов):
КАРМЕНОВА МАРХАБА АХМЕТОЛ-ИШНОВНА, ТЛЕБА ДИШОВА АЙЖАН СОЛТАНГА-ШЕВНА

Вид объекта авторского права: **программа для ЭВМ**

Название объекта: **Интеллектуальная информационно-аналитическая система «Оценка сейсмостойкости городских объектов на основе технологии анализа данных»**

Дата создания объекта: **05.10.2020**





Құжат түпнұсқасының: <http://www.kazpatent.kz/ru> сайтының
"Авторлық құқық" бөлімінде тексеруге болады. <https://copyright.kazpatent.kz>
Подлинность документа возможно проверить на сайте kazpatent.kz
в разделе «Авторское право» <https://copyright.kazpatent.kz>

Подписано ЭЦП

Оспанов Е. К.

ПРИЛОЖЕНИЕ Б

Акт внедрения результатов


МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное
бюджетное образовательное учреждение
высшего образования
«Алтайский государственный университет»
пр-т Ленина, 61, г. Барнаул, 656049
Тел. (385-2) 291-291, Факс (385-2) 66-76-26
E-mail: rector@asuu.ru



10.09.2021 № 10-2-21/03/4940
на № _____ от _____

УТВЕРЖДАЮ:

Первый проректор
по учебной работе АлтГУ,
канд. техн. наук, доцент
Жданова Е.А.

2021 г.



АКТ

о внедрении на базе ФГБОУ ВО «Алтайский государственный университет»
(г. Барнаул, Российская Федерация) результатов диссертационной работы
Карменовой М.А. «Разработка методов оценки сейсмостойкости городских
объектов на основе применения технологии анализа данных»

Мы, нижеподписавшиеся, директор Института математики и информационных технологий Журавлев Евгений Владимирович; научный руководитель научной-исследовательской работы и заведующий кафедрой теоретической кибернетики и прикладной математики Понькина Елена Владимировна, составили настоящий акт о том, что результаты диссертационного исследования Карменовой М.А. внедрены в учебный процесс в 2021 году.

Результаты исследований по теме диссертационной работы были положены в основу учебных курсов с элементами проектных научных исследований по дисциплинам «Модели и методы интеллектуального анализа», «Цифровые технологии анализа пространственных разновременных данных», «Цифровые технологии пространственного анализа и моделирования динамики природных процессов», реализуемых в рамках образовательной программы магистратуры по направлению 09.04.03 Прикладная информатика, профиль «Цифровые технологии анализа данных для устойчивого развития регионов Северной и Центральной Азии». Разработанное Карменовой М.А. web-приложение «Визуализация сейсмических данных (программа для ЭВМ)» и интеллектуальная информационно-аналитическая система «Оценка сейсмостойкости городских объектов на основе технологии анализа данных (программа для ЭВМ)» внедрены в образовательный процесс.

Научные результаты исследований опубликованы в рецензируемых изданиях, получили апробацию на международных конференциях:

а) М. Карменова, А. Нугуманова, А. Глебалинова. Кластерный анализ данных в решении задач по оценке сейсмической уязвимости объектов городской среды. Научно-технический журнал «Вестник АУЭС». 1(48), 2020.

б) M. Karmenova, A. Nugumanova, A. Tlebaldinova, A. Beldeubaev, G. Popova, A. Sedchenko. Seismic assessment of urban buildings using data mining methods. ICCTA, April 14-16, 2020.

в) Karmenova M., Tlebaldinova A., Madiyarov M., Zhantassova Zh. Spatial Analysis of the Earthquakes Distribution in Kazakhstan for Risk Mapping. ВЕСТНИК Национальной инженерной академии Республики Казахстан (НИАРК). ВЫЧИСЛИТЕЛЬНЫЕ ТЕХНОЛОГИИ. Федеральный исследовательский центр информационных и вычислительных технологий, 9-10 октября 2020 г., Часть 1, №3, С. 149-157.

г) W.Wójcik, M.Karmenova, S.Smailova, A.Tlebaldinova, A.Belbeubaev. Development of data-mining technique for seismic vulnerability assessment. Intl Journal of Electronics and Telecommunications, 2021, Vol.67, No.2, pp.261-266. DOI: 10.24425/ijet.2021.135974.

Эффективность внедрения: использование научных результатов исследования по диссертационной теме позволяют повысить качество учебно-методического обеспечения и способствовать дальнейшему совершенствованию научно-исследовательской работы обучающихся, позволяя им приобрести профессиональные компетенции в области применения современных подходов и технологий анализа данных, методов машинного обучения для решения проблем в исследуемой области, в частности оценки сейсмостойкости объектов городской среды.

Заведующий кафедрой
теоретической кибернетики
и прикладной математики,
к.т.н., доцент


Понькина Е.В.

Директор
института математики
и информационных технологий,
к.ф.-м.н., доцент


Журавлев Е.В.