

УДК 550.34.05.013.3

ХАРАКТЕРИЗАЦИЯ ГРУНТОВЫХ УСЛОВИЙ НА СЕЙСМОСТАНЦИЯХ СЕВЕРНОГО КАВКАЗА С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

© 2025 г. Т. С. Савадян¹, О. В. Павленко^{1, *}¹Институт физики Земли им. О.Ю. Шмидта РАН, г. Москва, Россия

*E-mail: olga@ifz.ru

Поступила в редакцию 28.11.2023 г.

После доработки 27.03.2024 г.

Принята к публикации 07.04.2024 г.

Для расширения возможностей использования записей местных землетрясений (для построения региональных уравнений прогноза движений грунта, оценки сейсмической опасности и др.) выполнена классификация сейсмостанций Северного Кавказа по грунтовым условиям. Разработана методика, позволяющая сделать оценку грунтовых условий посредством сравнения спектров слабых землетрясений, выбранных в узких диапазонах магнитуд и гипоцентральных расстояний, на разных станциях. Применение методов машинного обучения показало сложность задачи, но в то же время использование логических операций и методик позволило определить наиболее эффективные подходы для ее решения. В результате выполнена классификация 70-ти сейсмостанций Северного Кавказа по грунтовым условиям; грунтовые условия характеризуются одним безразмерным параметром, основанным на расчете спектральных характеристик. В будущем предполагается уточнить эти оценки.

Ключевые слова: характеристика грунтовых условий, Северный Кавказ, спектральные параметры, методы машинного обучения.

DOI: 10.31857/S0002333725010127, EDN: ABZUKI

ВВЕДЕНИЕ

Северный Кавказ является частью Альпийско-Гималайской складчатой системы и представляет собой сложную систему геологических структур и формаций, обусловленную длительной и многоступенчатой тектонической историей. Геологическое строение Северного Кавказа отражает процессы столкновения континентальных плит, которые привели к формированию горных хребтов, внутренних бассейнов и разломных систем. Северный Кавказ включает такие горные системы, как Главный Кавказский хребет, Северокавказское предгорье и Среднекавказская возвышенность [Рогожин, 2012]. Главный Кавказский хребет представляет собой протяженный горный хребет, состоящий из древних кристаллических пород и метаморфических слоев, сформировавшихся в палеозойскую эру. Северокавказское предгорье и Среднекавказская

возвышенность представляют собой складчатые структуры, образованные за счет столкновения континентальных плит [Пруцкий и др., 2005].

Сейсмическая активность в регионе Северного Кавказа является значительной; она обусловлена тектоническими процессами, связанными со столкновением Евразийской и Арабской плит. Землетрясения могут иметь высокую интенсивность и представлять большую опасность для населения и инфраструктуры [Габсатарова и др., 2018]. Территория подвержена высокой степени сейсмического риска вследствие высокой плотности населения, низкого качества массовой застройки, большого количества особо ответственных сооружений и высокого уровня исходной сейсмичности [Рогожин, 2012].

Основные зоны сейсмической активности на Северном Кавказе проходят вдоль границы между горными хребтами и предгорьями, а так-

же вдоль крупных разломных систем, таких как Копетдагский и Черноморский разломы (рис. 1).

Исследования сейсмической активности на Северном Кавказе направлены на изучение связи между тектоническими процессами и землетрясениями, а также на разработку методов прогнозирования и оценки сейсмической опасности и сейсмического риска для населения и инфраструктуры [Рогожин, 2012]. В рамках этих исследований проводятся наблюдения и мониторинг землетрясений с использованием сетей сейсмостанций, размещенных по всему региону. Данные, полученные от этих станций, анализируются и используются для создания карт сейсмической активности, которые служат основой для принятия решений о строительстве и проектировании инфраструктуры в зонах повышенного сейсмического риска. В рамках изучения геологии и сейсмологии Северного Кавказа активно разрабатываются и применяются новые методы и подходы, такие как микросейсмический мониторинг, дистанционное зондирование, геодинамические и гравиметрические измерения, геохимический анализ и методы машинного обучения, что позволяет лучше прогнозировать и оценивать сейсмические риски для населения и экономики [Акимов и др., 2019].

Записи местных землетрясений, полученные сетями сейсмостанций Северного Кавказа, являются основой для оценки сейсмической опасности в регионе. Однако эти записи имеют довольно ограниченное применение без знания грунтовых условий на сейсмостанциях, которые, к сожалению, остаются до сих пор практически неизученными.

Уточнение грунтовых условий под станциями необходимо, в первую очередь, для создания уравнений прогноза движений грунта (УПДГ), которые связывают параметры сильных движений земной поверхности при землетрясениях с магнитудой землетрясения и расстоянием до очага. Такие уравнения крайне необходимы в сейсмостойком строительстве, для проектирования и строительства сейсмостойких зданий.

Характеристики приповерхностных грунтов могут значительно влиять на усиление и спектральный состав сейсмических волн вблизи земной поверхности; знание грунтовых условий под станцией позволяет более точно определять положения очагов землетрясений и параметры колебаний поверхности при будущих сильных землетрясениях.

В настоящей работе была поставлена задача сделать оценки грунтовых условий на сейсмостанциях Северного Кавказа по записям местных землетрясений.

ДАННЫЕ И МЕТОД

На территории Северного Кавказа сейсмические наблюдения проводятся станциями четырех сейсмических сетей: OBGSR (телесеизмическая сеть Центрального отделения Федерального государственного бюджетного учреждения науки Федерального исследовательского центра “Единая геофизическая служба Российской академии наук” (ФИЦ ЕГС РАН), г. Обнинск), KMGSR (локальная сеть ФИЦ ЕГС РАН, г. Кисловодск), DAGSR (региональная сеть Дагестанского филиала ФИЦ ЕГС РАН, г. Махачкала) и NOGSR (региональная сеть Северо-Осетинского филиала ФИЦ ЕГС РАН, г. Владикавказ) (рис. 1). Все станции трехкомпонентные, оснащены цифровым оборудованием, подключены к интернету и передают информацию в центры обработки в режиме, близком к реальному времени [Габсатрова и др., 2018].

Станции расположены в стратегических точках региона, обеспечивая оптимальное покрытие сейсмически активных территорий и получение точных данных о землетрясениях. Они размещены вблизи основных и второстепенных разломов, вулканических структур и других геологических объектов, связанных с сейсмической активностью. Они могут быть расположены на горных склонах, в долинах или прибрежных районах, в зависимости от специфики местной сейсмической активности и доступности территории, или могут быть установлены в крупных населенных пунктах и вблизи таких инфраструктурных объектов, как дамбы, мосты и промышленные предприятия, чтобы обеспечить оперативное получение информации о землетрясениях и своевременное реагирование на них (рис. 1).

В настоящей работе исследовались записи (велосиграммы) местных землетрясений, происшедших в период с 2001 по 2020 г. и зарегистрированных 82-мя цифровыми сейсмостанциями (рис. 1). Каталог этих землетрясений включал в себя информацию о записях более чем 5000 сейсмических событий.

Как известно, приповерхностные грунты существенно влияют как на уровень, так и на спектральный состав сейсмических колебаний на поверхности. Исходя из этого мы предпола-

гали, что грунтовые условия найдут отражение в спектрах записей местных землетрясений. При этом, чтобы спектры велосиграм на разных станциях можно было сравнивать, записи должны быть выбраны в узких диапазонах магнитуд и расстояний от очага. Также отклик грунта должен был оставаться линейным, т.е. сейсмические движения на поверхности должны быть достаточно слабыми, но все же заметно превышающими уровень шума [Павленко, 2009]. Поэтому для анализа выбирались записи на достаточно больших эпицентральных расстояниях.

При таких условиях уровень сигнала на скальных станциях должен быть наиболее низким, и спектры велосиграм должны иметь простую форму: спектральные амплитуды максимальны на некоторой низкой частоте (определяемой полосой частот сейсмоприемника) и плавно спадают с ростом частоты [Voore, 2003].

Поэтому для анализа из всей базы записей сначала были выбраны записи землетрясений с гипоцентрными расстояниями 100–200 км и магнитудами от 4.0 до 5.0. Эти записи были разбиты на группы с шагом по магнитуде 0.2 и с шагом по гипоцентрному расстоянию 25 км — всего 20 групп. Однако, поскольку выборки должны быть представительными, в итоге анализировались лишь выборки с наибольшей представительностью данных, т.е. содержащие достаточное количество записей: не меньше приблизительно 50–60 записей в каждой. Таких выборок набралось 8, с моментными магнитудами 4.2–4.4 и 4.4–4.6 и гипоцентрными расстояниями от 100 до 200 км, с шагом 25 км (магнитуды M_s , используемые в ФИЦ ЕГС РАН переведены в моментные магнитуды по зависимостям, описанным в работе [Гусев, Мельникова, 1990]). В процессе дальнейшей работы их пришлось объединить в 4 более крупные выборки для большей представительности, увеличив шаг по расстояниям до 50 км.

Рассчитывались спектры Фурье записей-велосиграм (после вычитания линейных трендов) по наиболее интенсивной части записей. Для анализа выбраны наиболее качественные записи с частотой оцифровки 40 Гц и 50 Гц (рис. 3), т.е. не анализировались многочисленные записи с частотой оцифровки 20 Гц, поскольку спектры таких сигналов существенно ограничены по частоте.

Спектры сглаживались, затем для каждого спектра были рассчитаны их огибающие с помощью преобразования Гильберта:

$$H(t) = \frac{1}{\pi} v.p. \int_{-\infty}^{\infty} \frac{H(\tau)}{t - \tau} d\tau, \quad (1)$$

где: $H(\tau)$ — исходная функция; $v.p.$ — главное значение несобственного интеграла по Коши. Расчет огибающей спектра позволяет получить обобщенное представление о форме и характере спектра [Oppenheim, 1999], что упрощает дальнейшее сравнение спектров записей разных станций (рис. 4). По огибающей в дальнейшем были рассчитаны спектральные атрибуты или параметры формы спектра. Далее, говоря о спектрах, мы будем иметь в виду сглаженные огибающие спектров.

В заключительной части работы для более глубокого анализа были применены методы машинного обучения.

ХАРАКТЕРИЗАЦИЯ ГРУНТОВЫХ УСЛОВИЙ ПО СПЕКТРАМ ЗАПИСЕЙ СЛАБЫХ ЗЕМЛЕТРЯСЕНИЙ

На первом этапе работы по уровню спектральных амплитуд и формам спектров выделялись станции, установленные на скальном грунте. Вероятно, на Северном Кавказе таких станций большинство, поскольку твердая скала — желательное место для установки станции (чтобы избежать искажений сигналов, связанных с влиянием грунта), и природные условия позволяют достаточно легко выбирать такие места.

Для записей, выбранных в узких диапазонах магнитуд и гипоцентрных расстояний, спектры и спектральные параметры на скальных станциях будут близки, поскольку грунтовые условия сходны.

Как сказано выше, для анализа были выбраны записи слабых сейсмических событий, т.е. записи землетрясений с небольшими магнитудами и большими эпицентрными расстояниями, чтобы избежать нелинейных искажений в записях на грунтовых станциях вследствие нелинейности отклика грунта.

Среди таких записей, ориентируясь на формы их спектров, были выбраны предположительно станции на скальных грунтах: они имели наиболее низкий уровень спектральных амплитуд (поскольку грунты усиливают сейсмические колебания при их линейном поведении) и простую форму спектров, показывающую снижение спектральных амплитуд с частотой. Пример такого спектра показан на рис. 5.

По уровню и форме спектра в пределах каждой из 4-х анализируемых выборок записей были найдены станции со скальным грунтом или близким к скальному. На рис. 6 показаны

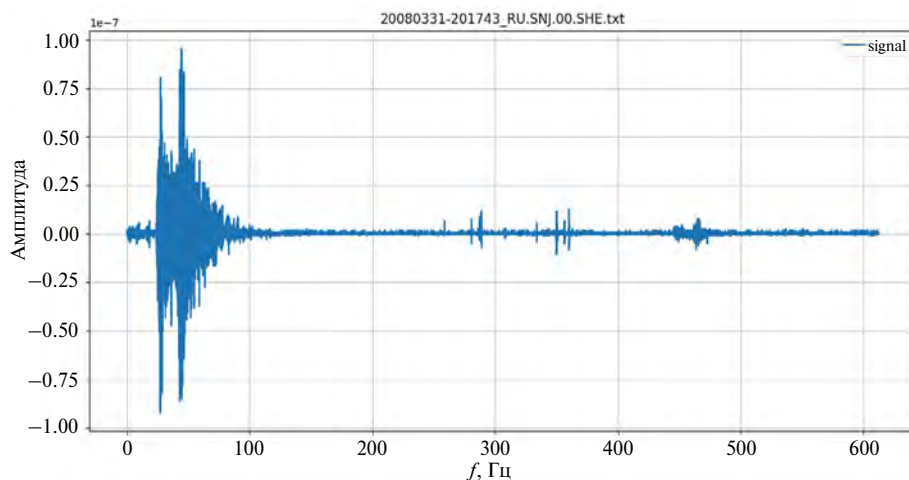


Рис. 3. Пример записи велосигаммы землетрясения на восточной компоненте (SHE) станции Сунжа (SNJ на рис. 1). По вертикальной оси — амплитуда в относительных единицах, по горизонтальной оси — время в секундах.

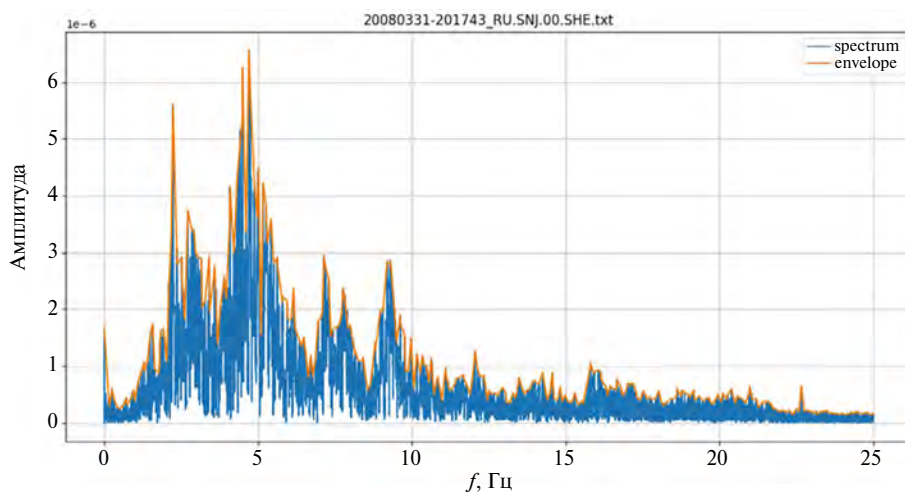


Рис. 4. Пример спектра велосигаммы и его огибающей. По вертикальной оси — амплитуда в относительных единицах, по горизонтальной оси — частота в герцах.

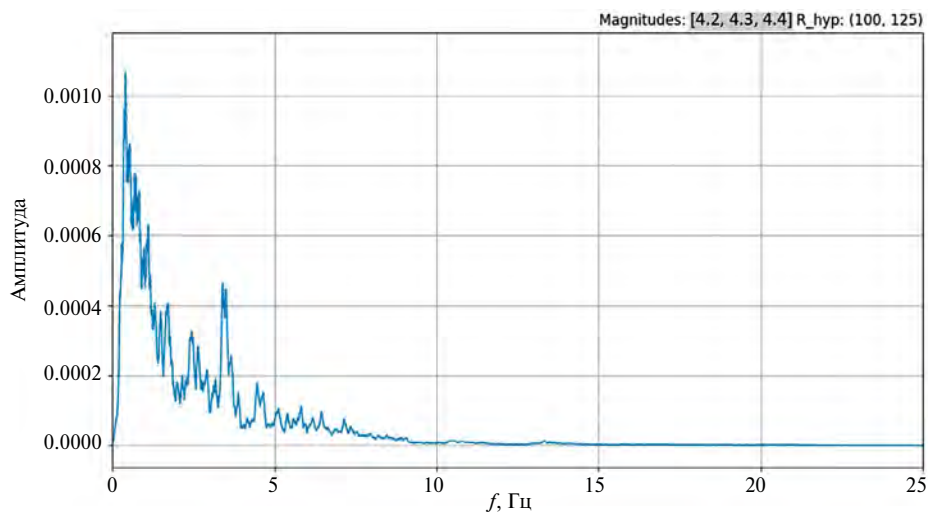


Рис. 5. Спектр велосигаммы местного землетрясения, записанного станцией Ставд-Дурт, установленной на скальном грунте. По вертикальной оси — амплитуда в относительных единицах, по горизонтальной оси — частота в герцах.

примеры спектров велосиграм группы землетрясений: спектр станции со скальным грунтом показан синим цветом.

Сейсмостанции Хунзах, Ботлих, Ставд-Дурт, Ведено, Грозный, Буйнакск, Караман, Сергокала, Унцукуль были таким образом идентифицированы как скальные. Другие станции, формы спектров которых отклоняются от форм спектров скальных станций, были идентифицированы как грунтовые.

Очевидно, что грунтовые станции можно классифицировать, т.е. разделить на классы, по степени отклонения их спектров от спектров

скальных станций, т.е. по степени влияния приповерхностных грунтов в месте установки станции на спектр. Для этого нужно выбрать некоторый набор параметров для описания формы и уровня спектра (спектральной огибающей). В качестве таких параметров были выбраны 7 параметров: амплитуда и частота максимума спектра ($A_{\max}$ и $f_{\max}$) и спектральные амплитуды на частотах 2 Гц, 5 Гц, 8 Гц, 11 Гц и 14 Гц, нормированные на $A_{\max}$. На наш взгляд, такой набор параметров достаточно полно описывает форму спектра и уровень спектральных амплитуд (рис. 7).

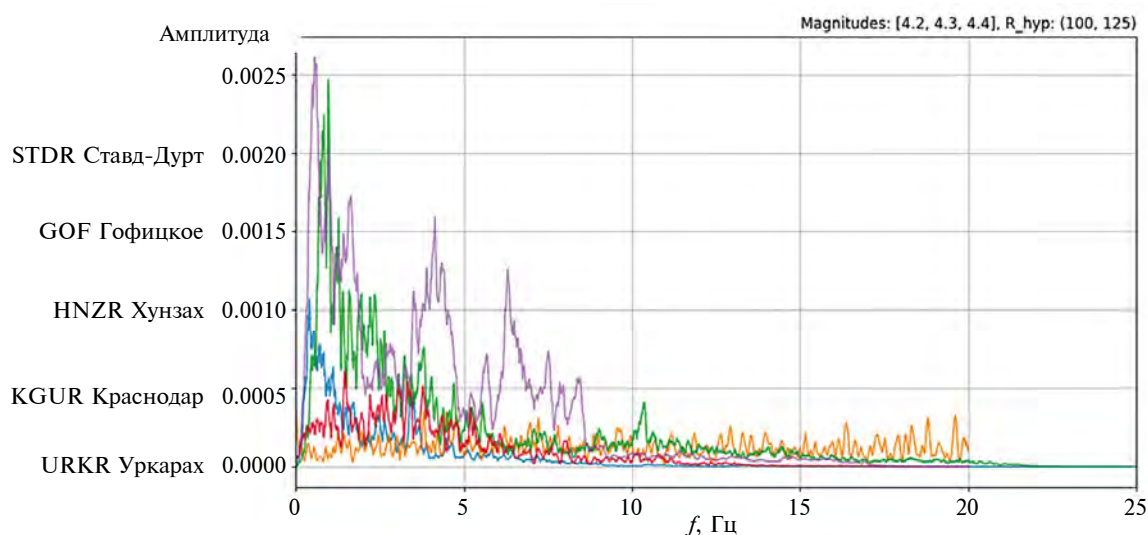


Рис. 6. Примеры спектров велосиграм группы землетрясений с магнитудами 4.2–4.4 и гипоцентрными расстояниями 100–125 км. По вертикальной оси — амплитуда в относительных единицах, по горизонтальной оси — частота в герцах. Спектры для станций Ставд-Дурт, Гофицкое, Хунзах, Краснодар и Уркарах показаны соответственно синим, зеленым, красным, желтым и сиреневым цветами.

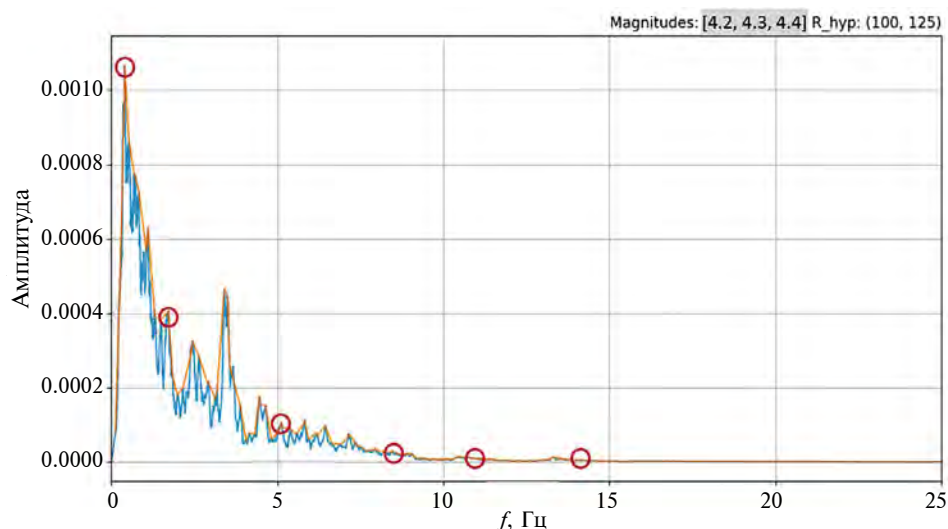


Рис. 7. Спектр станции Ставд-Дурт и его огибающая. Красными кружочками отмечены точки расчета спектральных параметров.

В таблице 1 представлены примеры рассчитанных параметров для некоторых скальных станций: Хунзах, Ботлих и Ставд-Дурт, Буйнакск, Ведено, Грозный, Караман в одной из выборок записей землетрясений. Спектральные амплитуды даны в м/с, частоты — в герцах. Можно сказать, что параметры, оцененные на трех скальных станциях, оказались достаточно близкими между собой, если учесть их отличия от параметров на других, нескальных станциях.

В пределах каждой из 4-х анализируемых групп записей спектральные параметры, полученные для скальных станций, были усреднены и, таким образом, получены средние значения 7-ми параметров, описывающих формы спектров на скальных станциях.

Затем для всех остальных станций (предположительно грунтовых) были рассчитаны отклонения их спектральных параметров от усредненных спектральных параметров скальных станций в пределах каждой из 4-х анализируемых выборок записей землетрясений. Из каждого параметра для грунтовой станции вычитался соответствующий усредненный параметр скальной станции, при этом значения разницы между нормированными амплитудами на частотах 2, 5, 8, 11 и 14 Гц были взяты по модулю.

Для получения общей оценки отклонения для каждой грунтовой станции полученные значения отклонений по всем семи параметрам суммировались.

По величине отклонений все станции были разбиты на категории; выделены следующие категории станций: при отклонениях меньше 1 — скальный грунт (категория 1); при отклонениях 1–3 — грунт категории 2; при отклонениях 3–5 — грунт категории 3; при отклонениях больше 5 — грунт категории 4.

Примеры полученных оценок отклонений спектров грунтовых станций от скальных станций показаны в табл. 2.

Таким образом было выполнено разделение сейсмостанций по категориям грунта на основе отклонений спектров грунтовых станций от спектров скальных станций.

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ХАРАКТЕРИЗАЦИИ ГРУНТОВЫХ УСЛОВИЙ

В заключительной части работы для классификации станций по категориям грунта были применены методы машинного обучения.

Таблица 1. Результат расчета спектральных параметров на скальных станциях

Станция	$A_{\max}$	$f_{\max}$	$A_{2\text{Гц}}$	$A_{5\text{Гц}}$	$A_{8\text{Гц}}$	$A_{11\text{Гц}}$	$A_{14\text{Гц}}$
HNZR	0.002827	1.357	0.32416	0.137558	0.037735	0.050336	0.059335
BTLR	0.001171	1.139	0.44529	0.105442	0.04723	0.011695	0.008547
STDR	0.001799	0.568	0.244855	0.050698	0.020983	0.01234	0.008545
BUJR	0.003110	1.356	0.496030	0.090696	0.008216	0.002922	0.004440
DVE	0.001205	1.142	0.130415	0.054676	0.015438	0.013088	0.011661
GRO	0.002268	0.714	0.408918	0.037700	0.014530	0.001365	0.002016
KANR	0.002612	0.535	0.231593	0.036666	0.010993	0.006025	0.003951

Таблица 2. Примеры отклонений спектральных параметров грунтовых и скальных станций

Станция	Отклонение	Категория грунта
STDR (Ставд-Дурт)	–0.1	0
NCK (Нальчик)	0.885	1
MAK (Махачкала)	2.318	2
DRN (Дербент)	3.769	3
AGYR (Агой)	8.507	4

Как известно, машинное обучение — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение за счет применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, математического анализа, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме. Многие методы тесно связаны с извлечением информации и интеллектуальным анализом данных, так что применение методов машинного обучения в решении такой непростой и ответственной задачи, как оценивание грунтовых условий на сейсмостанциях по записям землетрясений, может дать определенные преимущества.

Основная идея машинного обучения заключается в том, чтобы система могла самостоятельно “обучаться” на основе предоставленных данных и делать предсказания или принимать решения. Существует множество алгоритмов, с помощью которых можно обучить систему [Brink et al., 2016].

Для обучения модели необходимы данные, представленные в виде набора объектов, каждый из которых описывается набором признаков (*features*) и целевой переменной (*target*). Объекты в данных называются образцами (*samples*). Признаки (*features*) — переменные или характеристики, используемые для описания объектов или событий. Целевая переменная (*target*) — значение, которое модель должна предсказывать или классифицировать на основе признаков.

Необходимо оценивать качество модели; для этого используются определенные метрики — числовые показатели для измерения и оценки производительности моделей машинного обучения. В машинном обучении нужно сравнить ответ модели и реальный *target*, и метрики позволяют оценить, насколько модель приблизилась к правильному ответу. Метрики также позволяют сравнивать разные модели, определять, насколько точно и адекватно они выполняют свою задачу, а также контролировать процесс обучения и внедрения моделей в реальных приложениях.

В зависимости от типа задачи машинного обучения (регрессия, классификация, кластеризация и т. д.) используются различные метрики. Для задачи классификации, решаемой в данной работе, используются следующие метрики:

1. *Accuracy* (доля правильных ответов) — доля правильно классифицированных объектов относительно всех объектов:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (2)$$

где: TP (*true positive*) — истинно положительный результат; TN (*true negative*) — истинно отрицательный результат; FP (*false positive*) — ложно-положительный результат; FN (*false negative*) — ложноотрицательный результат.

2. *Precision* (точность) — доля истинно положительных результатов среди всех положительных результатов, предсказанных моделью:

$$precision = \frac{TP}{TP + FP}. \quad (3)$$

3. *Recall* (полнота) — доля истинно положительных результатов среди всех реальных положительных результатов. *Recall* показывает, как хорошо модель идентифицирует положительные случаи:

$$recall = \frac{TP}{TP + FN}. \quad (4)$$

4. F_1 -мера — среднее гармоническое между *precision* и *recall*. F_1 -score является более сбалансированной метрикой, учитывающей обе характеристики модели и обеспечивающей компромисс между ними:

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}. \quad (5)$$

Выбор подходящей метрики зависит от специфики задачи и требований к модели. Важно учитывать, что одна метрика может недостаточно полно описывать производительность модели, поэтому часто используются комбинации метрик для оценки моделей машинного обучения [Виноградова, Головин, 2017].

Подготовка данных для обучения

Набор данных, необходимых для обучения модели, включал в себя информацию о станции и рассчитанные для нее спектральные параметры (пример данных приведен в табл. 1).

Оценка исходных данных и их взаимоотношений — важный этап в процессе обучения модели классификации. Она позволяет улучшить производительность и точность модели, сократить время на обучение и подбор оптимальных параметров, а также снизить вероятность переобучения и ошибок.

Анализировать исходные данные и их взаимоотношения можно с помощью визуализации распределения классов в осях двух признаков, корреляционного анализа, методов понижения размерности или других статистических методов.

Это помогает определить наиболее релевантные и информативные признаки, которые могут ускорить процесс обучения и улучшить точность предсказаний. Кроме того, анализ данных также может привести к созданию новых признаков или комбинаций признаков, которые могут быть более эффективными для задачи классификации [Наумов и др., 2020].

Все объекты были построены в осях признак № 1, признак № 2 для каждой пары признаков.

Примеры таких диаграмм показаны на рис. 8 — рис. 11.

В настоящей работе решалась задача классификации станций по 5-ти различным категориям грунта, но, как видно из рис. 8 — рис. 11, отличить одну категорию от другой оказалось достаточно сложно. Это связано с тем, что часто встречаются пересечения между категориями для каждой пары признаков: например, 2-я имеет пересечения с 1-й, и с 3-й категориями. Категории 0 и 1 пересекаются практически полностью, что не позволяет однозначно визуальным образом их разделять (рис. 8 — рис. 11).

Поскольку визуально разделить классы (категории) однозначно не получается, было принято решение использовать модели, которые прогнозируют принадлежность к классу исходя из логических

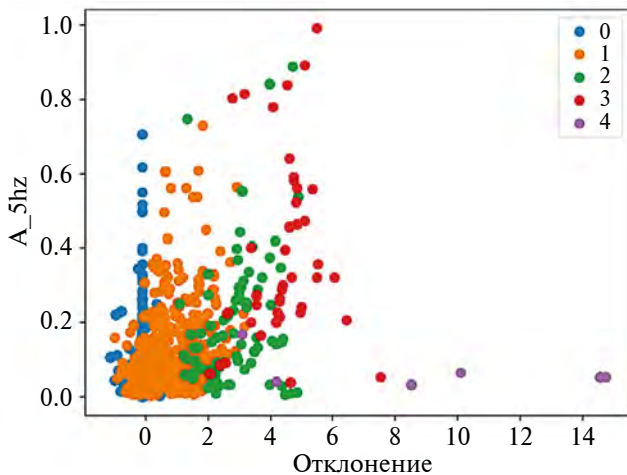


Рис. 8. Диаграмма рассеяния объектов с признаками “Отклонение” и “Амплитуда на 5 Гц” для 5-ти классов.

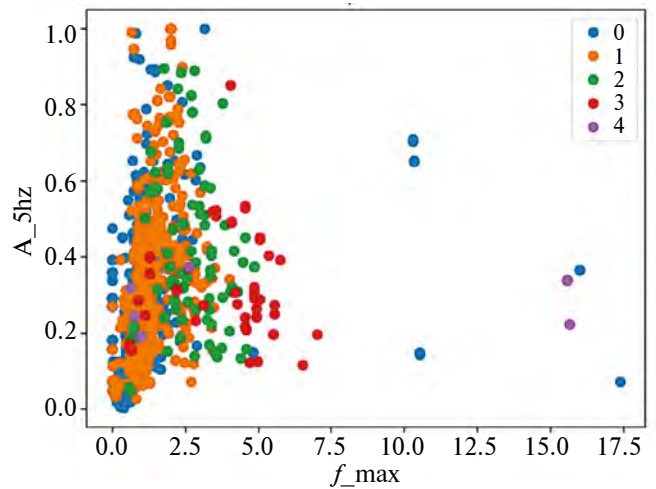


Рис. 9. Диаграмма рассеяния объектов с признаками “Частота максимума” и “Амплитуда на 2 Гц” для 5-ти классов.

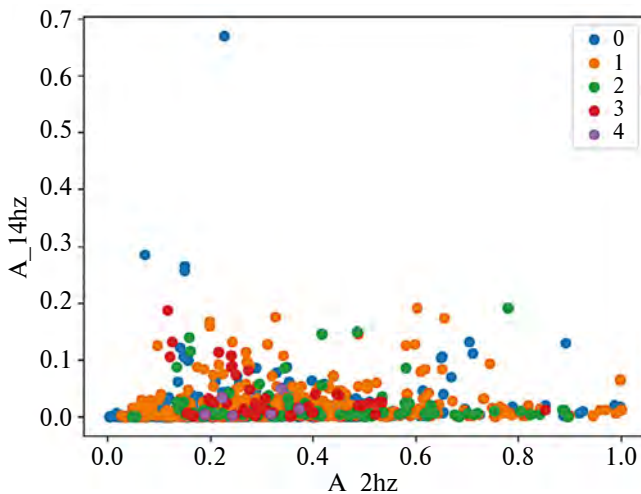


Рис. 10. Диаграмма рассеяния объектов с признаками “Амплитуда на 2 Гц” и “Амплитуда на 14 Гц” для 5-ти классов.

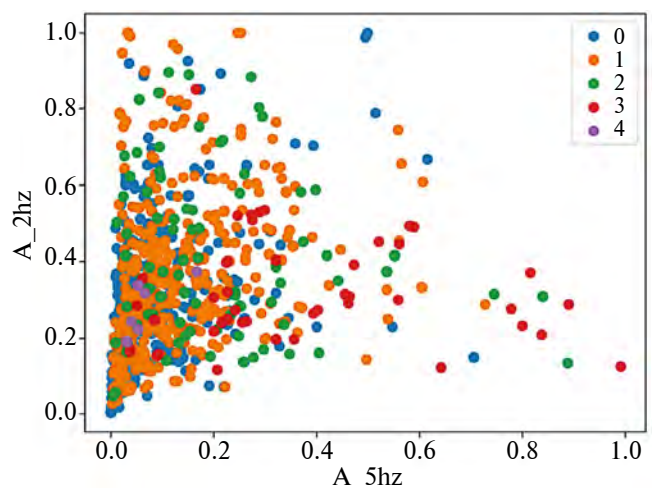


Рис. 11. Диаграмма рассеяния объектов с признаками “Амплитуда на 5 Гц” и “Амплитуда на 2 Гц” для 5-ти классов.

операций. Это такие модели, как “решающие деревья”, и их ансамбль — градиентный бустинг. Также мы попробовали проверить качество выполнения прогнозирования с помощью самой простой среди моделей машинного обучения — “ k -ближайших соседей” [Кузьмина, Ефимов, 2018].

Прежде всего, набор данных был собран и очищен от аномальных значений, также были обработаны пропущенные значения. Важным шагом являлось восстановление баланса классов — корректировка соотношения классов в обучающем наборе данных для устранения дисбаланса между ними. Дисбаланс классов возникает, когда один или несколько классов представлены значительно большим или меньшим количеством образцов по сравнению с остальными классами. Это может привести к плохому качеству обучения модели, поскольку модель может пренебрегать меньшими классами и сосредотачиваться на больших, что ухудшает ее обобщающую способность. В данной работе для восстановления баланса классов были применены веса классов. Этот метод заключается в присвоении различных весов классам, чтобы компенсировать неравномерное представление классов в данных. Веса позволяют модели уделить больше внимания менее представленным классам, без явного добавления дополнительных образцов.

После обработки набор данных был разделен на обучающую выборку и тестовую выборку. Это позволяет обучать модель на одной части данных и оценивать ее производительность на другой, независимой части [Наумов и др., 2020].

Кроме того, был предварительно отделен набор данных для последующей валидации обученной модели. В этом наборе было 400 вручную размеченных объектов с целью последующего сравнения их с предсказаниями модели. Этот резервный набор данных позволяет провести дополнительную оценку эффективности модели, что поможет лучше понять ее способность прогноза на новых данных.

Использованные модели машинного обучения

Первой была использована модель “ k -ближайших соседей” (kNN) — алгоритм, основанный на идее, что объекты с похожими признаками имеют похожие значения целевой переменной. Алгоритм определяет k -объектов, ближайших к новому объекту, и предсказывает значение целевой переменной на основе значений ближайших соседей.

Алгоритм k -ближайших соседей (kNN) является одним из наиболее простых из популярных ме-

тодов классификации в машинном обучении. Его преимущество заключается в том, что он не требует предварительного обучения на большом объеме данных, а решение о классификации нового объекта принимается на основе сравнения его характеристик с характеристиками ближайших объектов из обучающей выборки. Таким образом, если нужно классифицировать новый объект на основе схожих характеристик с объектами из обучающей выборки, то kNN может быть хорошим выбором. Однако kNN может быть неэффективным при работе с большими объемами данных, так как требует хранения и вычисления расстояний до всех объектов в обучающей выборке. Кроме того, выбор оптимального значения k может быть нетривиальным в зависимости от характера данных [Векслер, 2019].

Модели kNN были переданы значения признаков и классов в виде таблицы для обучения. Эти данные были случайным образом оверсемплированы, поскольку начальные данные были неравномерно распределены между классами. Гиперпараметры модели:

- количество ближайших соседей $k = 3$, потому что облака точек разных классов пересекались;
- расстояние между точками измерялось по уравнению Евклида.

После нескольких итераций обучения модели был получен результат в описанных выше метриках (табл. 3).

Поскольку количество объектов неравномерно распределено между классами, значения точности и полноты для нулевого и первого класса выше, чем для остальных. Точность для 2-го класса ниже, чем для 3-го и 4-го, несмотря на то, что у этих классов меньше выборка. Это связано с тем, что признаки объектов 2-го класса пересекаются с признаками как 1-го, так и 3-го классов. Поэтому алгоритм kNN может присваивать объект к неправильному классу. Также общая доля правильных ответов (*accuracy*) данной модели составила 72%, что считается неудовлетворительным для дальнейшего использования.

Вторая использованная в этой работе модель машинного обучения — “решающие деревья”. Решающие деревья (*Decision Trees*) — алгоритм, который строит иерархическую структуру в виде дерева, где каждый узел представляет собой условие на одном из признаков, а листья содержат предсказания для целевой переменной. Решающие деревья делают предсказания, последовательно проходя по узлам дерева от корня до листа. Модель основана на применении иерархического

Таблица 3. Сводная таблица результатов обучения со сравнением качества моделей

<i>Precision</i>			
Класс	<i>kNN</i>	Решающие деревья	Градиентный бустинг
0	0.78	0.8	0.89
1	0.81	0.91	0.83
2	0.4	0.74	0.74
3	0.53	0.63	0.67
4	0.67	0.68	0.63
<i>Recall</i>			
Класс	<i>kNN</i>	Решающие деревья	Градиентный бустинг
0	0.88	0.98	0.87
1	0.68	0.78	0.85
2	0.5	0.81	0.69
3	0.57	0.67	0.73
4	0.5	0.5	0.57
<i>F₁-мера</i>			
Класс	<i>kNN</i>	Решающие деревья	Градиентный бустинг
0	0.83	0.86	0.88
1	0.74	0.83	0.84
2	0.44	0.77	0.72
3	0.55	0.65	0.7
4	0.58	0.58	0.6
<i>Accuracy</i>			
	<i>kNN</i>	Решающие деревья	Градиентный бустинг
	0.72	0.81	0.89

Примечания: *kNN* — *k*-ближайших соседей, *accuracy* — доля правильных ответов, *precision* — точность, *recall* — полнота. Цветом выделены лучшие результаты.

разбиения пространства признаков на области, соответствующие различным классам или значениям целевой переменной. Проще говоря, решающие деревья представляют собой последовательность условий или правил, позволяющих классифицировать объекты или предсказывать их свойства. Принцип работы решающего дерева:

1. Изначально все данные находятся в корне дерева.
2. Выбирается признак и пороговое значение для этого признака, которые наилучшим образом разбивают данные на подмножества. Критерий разбиения зависит от типа задачи (классификации или регрессии) и может быть основан, например, на мере неопределенности (энтропии) или среднеквадратичной ошибке.
3. Для каждого полученного подмножества данных повторяется процесс разбиения, пока

не будет достигнут некоторый критерий остановки, такой как максимальная глубина дерева, минимальное количество объектов в листе или незначительное улучшение критерия разбиения.

В результате формируется дерево, состоящее из внутренних узлов с условиями разбиения и листьев, которые содержат прогнозы для объектов, попавших в соответствующую область пространства признаков (рис. 12).

Логика работы решающего дерева заключается в том, что алгоритм последовательно проверяет условия в узлах дерева, начиная от корня, и направляет объекты в соответствующие ветви. При достижении листа дерева алгоритм выдает прогноз на основе главного класса объектов, попавших в данный лист. Таким образом, решающее дерево обеспечивает интуитивно понятное и интерпретируемое представление принятых

решений и может быть легко визуализировано и объяснено [Генрихов, Дюкова, 2012].

Модель решающего дерева обучалась на тех же данных, что и kNN , и были получены аналогичные оценки. Гиперпараметры модели (были заданы веса для каждого класса, поскольку данные распределены неравномерно между ними):

- Класс 0 — 0.1
- Класс 1 — 0.1
- Класс 2 — 0.2
- Класс 3 — 0.3
- Класс 4 — 0.3
- Максимальная глубина дерева = 3

После обучения модели решающих деревьев были получены результаты, представленные в табл. 3.

Общая доля правильных ответов (*accuracy*) составила 81%, и оценки точности и полноты лучше, чем у модели kNN . Поскольку результат обучения алгоритма решающие деревья лучше, было принято решение использовать еще одну модель машинного обучения, которая является ансамблем нескольких деревьев.

Третья использованная модель — градиентный бустинг. Это ансамблевый метод машинного обучения, который строит комбинацию из слабых моделей (решающих деревьев) с целью улучшить точность и устойчивость прогнозов. Используется итеративный процесс, в ходе которого каждая последующая модель стремится исправить ошибки, допущенные предыдущими моделями. Гра-

диентный бустинг использует градиентный спуск для оптимизации функции потерь, что позволяет найти веса каждой из базовых моделей.

Принцип работы градиентного бустинга заключается в следующем:

1) инициализация ансамбля с начальной моделью;

2) на каждой итерации обучается новая базовая модель (решающее дерево), которая стремится предсказать ошибки, допущенные ансамблем на предыдущих итерациях;

3) во время обучения базовой модели используется градиентный спуск для оптимизации функции потерь, что позволяет находить направление, в котором ошибка уменьшается наиболее быстро;

4) найденная модель добавляется в ансамбль с определенным коэффициентом, который также оптимизируется в процессе обучения, что позволяет контролировать вклад каждой базовой модели в ансамбль и предотвращать переобучение;

5) итерации продолжаются до тех пор, пока не будет достигнут предопределенный критерий остановки, такой как максимальное количество итераций или порог улучшения функции потерь.

Логика работы градиентного бустинга заключается в комбинации прогнозов слабых моделей, обученных на разных этапах итеративного процесса, чтобы получить более точный и устойчивый прогноз. В результате градиентный бустинг может обеспечить достаточно высокую точность предсказаний и хорошую обобщающую способ-

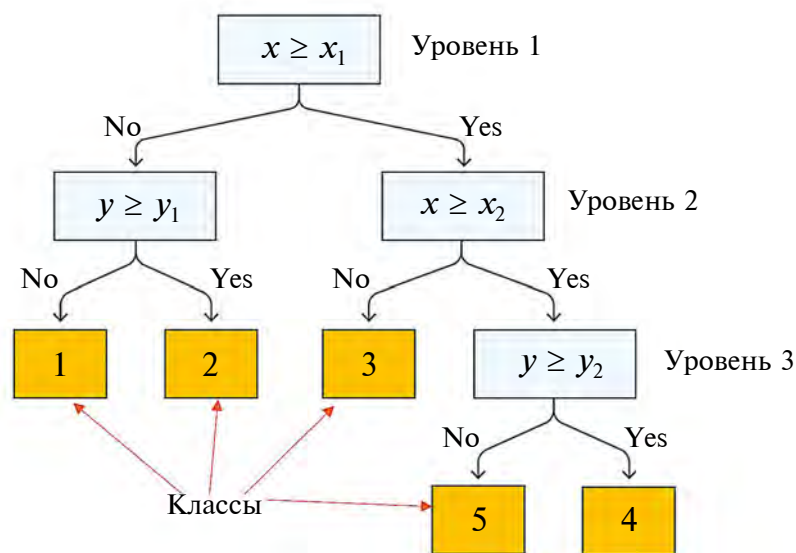


Рис. 12. Логика работы алгоритма решающих деревьев. Здесь x , y — признаки объектов.

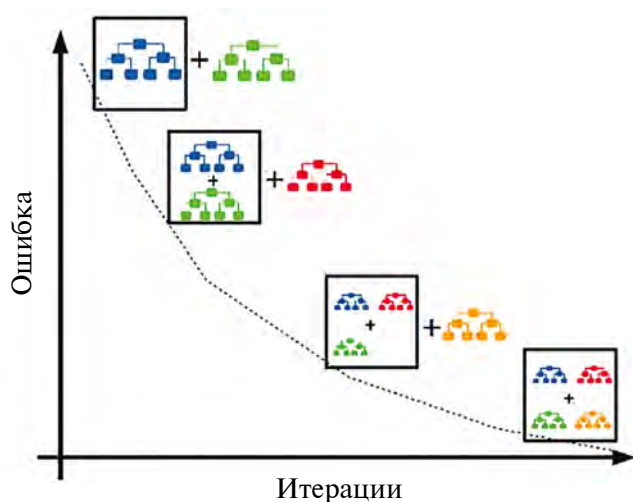


Рис. 13. Логика работы алгоритма градиентный бустинг. С ростом числа итераций возрастает количество решающих деревьев.

ность, что делает его одним из самых популярных и эффективных методов машинного обучения в решении различных задач (рис. 13).

Градиентный бустинг обладает рядом преимуществ по сравнению с другими методами машинного обучения. Во-первых, он способен обрабатывать разные типы данных, включая категориальные и числовые признаки. Во-вторых, он устойчив к выбросам и может работать с небольшим количеством данных. В-третьих, он позволяет контролировать сложность модели, предотвращая переобучение с помощью регуляризации и настройки гиперпараметров, таких как глубина деревьев и скорость обучения.

Тем не менее у градиентного бустинга есть и некоторые недостатки. Одним из них является вычислительная сложность, так как он требует построения большого количества базовых моделей в итеративном процессе. Это может замедлить обучение и предсказания, особенно на больших наборах данных. В целом градиентный бустинг является мощным и гибким методом машинного обучения, который позволяет строить ансамблевые модели с высокой точностью и обобщающей способностью [Дьяконов, Новикова, 2018].

Модель градиентного бустинга обучалась на тех же данных, что и два предыдущих алгоритма. Гиперпараметры модели:

- количество деревьев решений ($n_{estimators}$) = 10 000;
- скорость обучения (*learning rate*) = 0.005 — этот гиперпараметр определяет, как быстро модель будет обучаться на тренировочных данных; он контролирует размер шага, с которым гради-

ентный спуск изменяет веса модели на каждой итерации обучения;

- максимальная глубина дерева = 1.

Как видно из табл. 3, результаты обучения данной модели качественнее двух предыдущих. Все метрики выше и при этом реалистичные, что свидетельствует об отсутствии переобучения алгоритма. Доля правильных ответов для всех классов составила 89%, что является достаточно хорошим результатом для дальнейшего использования модели для прогноза.

Согласно анализу данных, представленных в табл. 3, модели решающих деревьев и градиентного бустинга показывают лучшие результаты с точки зрения метрик. Несмотря на то что градиентный бустинг обладает самой высокой точностью, модели, основанные на решающих деревьях, иногда превосходят его по другим параметрам. Это может быть связано с особенностями работы ансамбля моделей, когда некоторые деревья могут давать лучший прогноз, в то время как другие — хуже, что приводит к усредненному результату.

В заключение была проведена дополнительная работа с 400 неразмеченными объектами из первоначального набора данных. Этот набор был обработан по тому же алгоритму, что и обучающая выборка. Затем было проведено ручное разбиение объектов на классы, основанные на отклонениях грунтовых условий от стандартных. Эта работа заняла два рабочих дня.

Для проверки качества и эффективности работы обученной модели градиентного бустинга, она была применена к тем же 400 неразмеченным объектам. Модель демонстрировала впечатляющие результаты, правильно классифицируя 355 из 400 объектов, что эквивалентно точности предсказаний в 88,7%. Помимо этого, стоит отметить, что модель справилась с этой задачей всего за 5 минут, что подтверждает ее высокую эффективность и потенциал для применения на практике.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ И ВЫВОДЫ

Таким образом, анализ и обработка записей местных землетрясений, полученных сейсмостанциями Северного Кавказа, позволили получить актуальную информацию о грунтовых условиях на станциях.

Выполнена классификация 70-ти (для остальных 12-ти из 82-х станций для классификации не хватило записей землетрясений) сейсмостанций Северного Кавказа по грунтовым условиям (табл. 4); грунтовые условия характеризуются

Таблица 4. Результат классификации станций по типам грунтов

Станция	Расположение	Тип грунта	Грунт (по данным ФИЦ ЕГС РАН)
AD2	Ардон	0–1	лессовидный суглинок, глина
AGYR	Агой	3	
АНМР	Ахметовская	2	
АКТ	Ахты	2	аргиллит
ALER	Александровское	3	
ANN	Анапа	0	суглинок, аллювий, глины сланцеватые, песчаники
ARKR	Аракани	2	известняк
ARX	Архыз	2	
BEY	Белый Уголь	2	мергелистые известняки
BTK	Батакоюрт	0–1	суглинки и супеси, ниже по разрезу глины
BTLR	Ботлих	0	
BUJR	Буйнакск	0	
DBK	Дубки	1–2	известняк
DIG	Дигорское ущелье	1	алевролиты, аргиллиты с редкими прослоями песчаников и глинистых сланцев, ниже по разрезу гранитоиды
DLMR	Дылым	2	делювий
DOM, DOMR	Домбай	3	
DRN	Дербент	1–2	известняк
DVE	Ведено	0	
FSTR	Фишт	1	
GNB	Гуниб	1–2	известняк
GOF, GOFR	Гофицкое	2	песчано-глинистые осадки
GOYR	Гойтх	2	
GRO	Грозный	0	
GUZR	Гузерибль	2	
HNZR	Хунзах	0	скала, известняк
KANR	Караман	0	
KBT, KBTC	Куба-Таба	2–3	глина
KGU, KGUR	Краснодар	0–2	
KIVB	Кисловодск	1	известняк
KMG, KMGR	Комгарон	1	супесь, суглинок
KMKR	Кумух	3–4	аргиллит

Окончание таблицы

Станция	Расположение	Тип грунта	Грунт (по данным ФИЦ ЕГС РАН)
KOR, KORR	Кора	0–1	суглинки, глины с прослоями песков
KRNR	Каранай	1–2	известняк
KSMR	Касумкент	2	аллювий
LABN	Лабинск	0	
LAC, LACR	Лац	2	глина, суглинок
LSN	Лескен	1–2	глины, суглинки, супесь.
LZRR	Лазаревское	1	
MAK	Махачкала	2–3	аллювиальные отложения
NCK	Нальчик	1	
NEY, NEUR	Нейтрино	1–2	
NVN	Невинномысск	1	
PRT, PRTR	Притеречная	0–1	глина, суглинок, песчано-валунно-галечные отложения
PYA1	Пятигорск	4	
RPOR	Красная Поляна	2	
SGKR	Сергокала	0	известняк
SHA	Шиджатмаз	1–2	известняк
SNJ	Сунжа	1	песчано-валунно-галечные отложения
SOC	Сочи	1	глинистые сланцы
STDR	Ставд-Дурт	0	песчано-валунно-галечные отложения
TLTR	Тлярата	2	
TRK, TRKR	Терская	0	глина, суглинок, песчано-валунно-галечные отложения
UNCR	Унцукуль	0	песчаник
URKR	Уркарах	0–1	скальные породы
VLK, VLKG, VLKR	Владикавказ	1	песчано-валунно-галечные отложения
VSLR	Весёлое	1	
XNZR	Хунзах	0	скала, известняк
ZEI	Цей	1	ледниковые отложения, ниже по разрезу гранитоиды и сланцы

одним безразмерным параметром, основан-
ным на расчете спектральных характеристик.
Хотя такая информация о грунтовых условиях
на станциях и довольно ограничена, ее уже
можно использовать при разработке уравнений
прогноза движений грунта для Кавказского

региона. В будущем предполагается уточнять
данные о грунтовых условиях на станциях.

Нужно отметить, что описания грунтовых
условий на сейсмостанциях, предоставляемые
ФИЦ ЕГС РАН, практически не дают никакой
информации о грунтовых условиях, поскольку

содержат описания лишь поверхностной геологии. В то время как на уровень и спектральный состав сильных движений на поверхности при землетрясениях оказывает влияние грунтовая толща мощностью в несколько десятков метров приповерхностных грунтов. Поэтому никакие сравнения полученных в настоящей работе численных оценок с описаниями ФИЦ ЕГС РАН невозможны.

Применение методов машинного обучения показало сложность задачи характеристики грунтовых условий на сейсмостанциях по записям местных землетрясений, но в то же время использование логических операций и методик, таких как k -ближайших соседей, решающие деревья и градиентный бустинг, позволило определить наиболее эффективные подходы для решения задачи характеристики грунтовых условий.

Отметим, что ранее методы машинного обучения не применялись для характеристики грунтовых условий на сейсмостанциях.

Обучение моделей машинного обучения на основе выбранных методов показало их применимость, эффективность и достаточно высокую точность в решении поставленной задачи. Показаны перспективы и потенциал использования методов машинного обучения в сейсмологии, что может стать основой для дальнейшего развития этого направления исследований и применения новых технологий в конечном счете для повышения безопасности населения и инфраструктуры в сейсмоактивных регионах.

СПИСОК ЛИТЕРАТУРЫ

Акимов В. А. и др. Карты сейсмической опасности Северо-Западного и Центрального Кавказа в детальном масштабе // Вопросы инженерной сейсмологии. 2019. Т. 46. №. 4. С. 57–74.

Векслер В. А. Машинное обучение на основе алгоритма “ k -ближайших соседей”. Вызовы цифровой экономики: итоги и новые тренды. 2019. С. 110–115.

Виноградова Е. П., Головин Е. Н. Метрики качества алгоритмов машинного обучения в задачах классификации. Научная сессия ГУАП. 2017. С. 202–206.

Габсатарова И. П. и др. Северный Кавказ // Землетрясения Северной Евразии. 2018. №. 21 (2012). С. 79–94.

Габсатарова И. П. и др. Северный Кавказ. Землетрясения России в 2020 году. Обнинск: ФИЦ ЕГС РАН. 2022. 204 с.

Генрихов И. Е., Дюкова Е. В. Классификация на основе полных решающих деревьев // Журнал вычислительной математики и математической физики. 2012. Т. 52. №. 4. С. 750–761.

Гусев А. А., Мельникова В. Н. Связи между магнитудами — среднемировые и для Камчатки // Вулканология и сейсмология. 1990. № 6. С. 55–63.

Дьяконов И. Д., Новикова С. В. Решение задачи прогнозирования при помощи градиентного бустинга над решающими деревьями. Научный форум: технические и физико-математические науки. 2018. С. 9–12.

Кузьмина С. В., Ефимов А. И. Актуальные методы машинного обучения в области классификации. Актуальные проблемы современной науки и производства. 2018. С. 34–38.

Наумов В. Н., Жиряева Е. В., Падерно П. И. Анализ данных и машинное обучение. Методы и инструментальные средства. 2020.

Павленко О. В. Сейсмические волны в грунтовых слоях: нелинейное поведение грунта при сильных землетрясениях последних лет. Науч. мир. 2009.

Пруцкий Н. И. и др. Геология и минералогия Северного Кавказа-современное состояние (Геологический атлас Северного Кавказа м-ба 1: 1 000 000) // Региональная геология и металлогения. 2005. №. 25. С. 27–38.

Рогожин Е. А. Сейсмическая опасность на Северном Кавказе // Экологический Вестник научных центров ЧЭС. 2012. № 1. С. 124–128.

Boore D. M. Simulation of Ground Motion Using the Stochastic Method // Pure Appl. Geophys. 2003. V. 160. P. 635–676.

Brink H., Richards J., Fetherolf M. Real-world machine learning. Simon and Schuster. 2016.

Oppenheim A. V. Discrete-time signal processing. Pearson Education India. 1999.

Characterization of Ground Conditions at Seismic Stations in the North Caucasus Using Machine Learning Methods

T. S. Savadyan^a and O. V. Pavlenko^{a,*}

^a*Schmidt Institute of Physics of the Earth, Russian Academy of Sciences, Moscow, Russia*

^{*}*e-mail: olga@ifz.ru*

Received November 28, 2023; revised March 27, 2024; accepted April 7, 2024

Abstract — To extend the capabilities of using records of local earthquakes (for constructing regional ground motion prediction equations, assessing seismic hazard, etc.), the classification of seismic stations in the North Caucasus by the ground conditions was performed. A technique has been developed that allows assessment of ground conditions by comparing spectra of weak earthquakes selected in narrow ranges of magnitudes and hypocentral distances, at different stations. The use of machine-learning methods showed the complexity of the problem, but at the same time, the application of logical operations and techniques allowed us to find the most effective approaches to solve it. As a result, 70 seismic stations of the North Caucasus were classified according to the ground conditions; the conditions were characterized by one dimensionless parameter based on the calculation of spectral characteristics. We are planning to refine the estimates in the future.

Keywords: characterization of ground conditions, North Caucasus, spectral parameters, machine-learning methods