КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Спеціальність 122 - Комп'ютерні науки,
програма "Інформаційна аналітика та впливи"

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему:

"ПРОГНОЗУВАННЯ ВИВЕРЖЕННЯ ВУЛКАНІВ НА ОСНОВІ ВИЯВЛЕННЯ ПРИХОВАНИХ ПРЕКУРСОРІВ У ГЕОФІЗИЧНИХ ДАНИХ"

Студентки 2-го курсу групи ІАВ-21

Толстокорової Анастасії Юріївни

(прізвище, ім'я, по батькові)

(підпис студента)

Науковий керівник:

доктор технічних наук, доцент

(науковий ступінь, вчене звання)

Єгорченков Олексій Володимирович

(прізвище, ім'я, по батькові)

(дата) *(підпис)*

Попередній захист:

(Висновок: "До захисту в Державній екзаменаційній комісії")

Завідувач кафедри технологій управління

(підпис) *(прізвище, ініціали)* *(дата)*

Київ – 2022

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій

Кафедра технологійуправління
Освітньо-кваліфікаційний рівень Магістр
Спеціальність 122-Комп'ютернінауки
Освітня програма Інформаційнааналітикатавпливи

ЗАТВЕРДЖУЮ
Завідувач кафедри
професор Морозов В.В.
«____» _____20__ року

ЗАВДАННЯ НА ВИКОНАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Студентка Толстокорова Анастасія Юріївна
Група ІАВ-21

1. Тема кваліфікаційної роботи «Прогнозування виверження вулканів на основі виявлення прихованих прекурсорів у геофізичних даних»

Затверджена наказом по від «__»_____2022р. №= = .
2. Строк подання студентом готової роботи – "__"= 2022р.

3. Цільова установка та вихідні дані до роботи прогнозування часу до виверження наступного вулкану за допомогою прогнозуючої моделі, побудованої на основі алгоритму машинного навчання CatBoost та навчена на даних сейсмічних сенсорів, зібраних італійськимнаціональннимінститутомгеофізикитавулканології.

4. Зміст роботи вступ; постановка задачі та аналіз її вирішення; математичні моделі та методи дослідження для вирішення задачі; реалізація пргнозуючої моделі; практичне застосуваннярозробленоїмоделі;висновки.

5. Перелік графічного матеріалу (слайдів) 5брисунків,9формул,3таблиці,6додатків.
6. Календарний план виконання роботи:

№ п/п	Назва частин роботи	%	Виконання роботи	
			За планом	Фактично
1.	Вибір теми дипломної роботи	3	01.10.21	01.10.21
2.	Протокол кафедри ТУ про затвердження тем дипломних робіт та призначення наукових керівників	2	17.11.21	17.11.21

3.	Формування переліку нормативних матеріалів, літератури з проблематики дипломної роботи	10	07.01.22	07.01.22
4.	Складання розгорнутого плану кваліфікаційної роботи	5	18.01.22	18.01.22
5.	Ознайомлення наукового керівника з розгорнутим планом кваліфікаційної роботи. Внесення змін.	5	19.01.22 - 20.01.22	20.01.22
6.	Підготовка розділу 1 «ПОСТАНОВКА ЗАДАЧІ ТА АНАЛІЗ ЇЇ ВИРІШЕННЯ»	10	14.02.22	15.02.22
7.	Підготовка розділу 2 «МАТЕМАТИЧНІ МОДЕЛІ ТА МЕТОДИ ДОСЛІДЖЕННЯ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ»	14	08.03.22	08.03.22
8.	Підготовка розділу 3 « РЕАЛІЗАЦІЯ ПРОГНОЗУЮЧОЇ МОДЕЛІ»	14	01.04.22	01.04.22
9.	Підготовка розділу 4 «ПРАКТИЧНЕ ЗАСТОСУВАННЯ РОЗРОБЛЕНОЇ МОДЕЛІ»	13	20.04.22	20.04.22
10	Оформлення кваліфікаційної роботи. Підготовка висновків і пропозицій	15	03.05.22	03.05.22
11	Передача кваліфікаційної роботи науковому керівникові	2	04.05.22	04.05.22
12	Передача кваліфікаційної роботи рецензенту для рецензування	2	11.05.22	11.05.22
13	Попередній захист кваліфікаційної роботи	5	17.05.22	17.05.22

Дата видачі завдання «___»= =_____ 2022р.

Керівник роботи доктор технічних наук, доцент Єгорченков Олексій Володимирович
(посада, прізвище, ім’я, по батькові)

_____
(підпис)

Завдання прийняла до виконання студентка групи ІАВ-21

Толсткорова Анастасія Юріївна
(прізвище, ім’я, по батькові)

ЗМІСТ

АНОТАЦІЯ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Спеціальність 122 - Комп’ютерні науки,

освітня програма “Інформаційна аналітика та впливи”

Дипломна робота магістра ТолстокоровоїАнастасіїЮріївни.

Тема роботи – «Прогнозування виверження вулканів на основі виявлення прихованих прекурсорів у геофізичних даних».

Мета дипломної роботи магістра – дослідити поведінку вулканів, розробити математичний апарат для тлумачення предикторів незвичайних подій в контексті вулканів для систематизації залежності від них часу до виверження.

Об’єкт дослідження – патерни зміни поведінки сейсмічних сигналів напередодні виверження вулканів.

Предмет дослідження – інформаційні засоби, методи, моделі управління даними при аналізі і прогнозуванні зміни патернів поведінки сейсмічних сигналів.

Наукова новизна роботи – розроблено математичний апарат для прогнозування вулканів, який відрізняється від існуючих тим, що потребує мінімальну кількість витрачених ресурсів при високій оцінці ефективності прогнозування. Використаний нетривіальний підхід та досліджена ефективність роботи CatBoost при вирішенні даного типу задач. Може масштабуватися на більш об’ємні дослідження та використовуватися за прямим та непрямим призначенням.

У роботі досліджуються існуючі підходи до вирішення задач регресії часових рядів. Розробляється нова методика вирішення на прикладі даних сейсмічних сигналів, а також проводиться аналіз її ефективності. Наводяться рекомендації щодо практичної імплементації методики.

Дипломна робота складається зі вступу, основної частини, яка включає чотири розділи, висновків, використаних джерел та додатків. Всього налічує 145 сторінок та перелік посилань з 47 джерел на 7 сторінках.

Ключові слова: сейсмічні сигнали, регресор, часові ряди, інтелектуальний аналіз даних, прогнозуюча математична модель, методика.

ВСТУП

Вулкани викидають гарячі небезпечні гази, попіл, лаву та гірські породи, які є потужними та руйнівними. Люди масово гинуть від вибухів вулканів.

Виверження вулканів можуть призвести до життєво небезпечних загроз для здоров'я, таких як повені, зсуви, відключення електроенергії, забруднення питної води та лісові пожежі. Проблеми зі здоров'ям після виверження вулкана включають інфекційні захворювання, респіраторні захворювання, опіки, травми від падінь та автомобільні аварії, пов'язані з слизькими та туманними умовами, викликаними попелом. При завчасному попередженні ймовірність негативних наслідків для здоров'я від виверження вулкана дуже низька.

Вплив вулканічного попелу вкрай шкідливий. Немовлята, літні люди та люди з респіраторними захворюваннями, такими як астма, емфізема та інші хронічні захворювання легенів, не уникнуть проблем, якщо вони вдихають вулканічний попіл. Попіл зернистий, абразивний, іноді їдкий і завжди неприємний. Дрібні частинки золи можуть потерти (подряпати) передню частину ока. Частинки золи можуть містити кристалічний кремнезем, матеріал, який викликає респіраторне захворювання, яке називається силікозом.

Більшість газів з вулкана швидко здувається. Однак важкі гази, такі як вуглекислий газ і сірководень, можуть накопичуватися в низинних районах. Найпоширенішим вулканічним газом є водяна пара, за ним слідують вуглекислий газ і діоксид сірки. Діоксид сірки може викликати проблеми з диханням як у здорових людей, так і у людей з астмою та іншими респіраторними проблемами. Інші вулканічні гази включають хлористий водень, монооксид вуглецю та фторид водню. Кількість цих газів сильно варіюється від одного виверження вулкана до іншого.

Хоча гази зазвичай швидко викидаються, можливо, що люди, які знаходяться поблизу вулкана або які знаходяться в низинних районах за вітром, можуть піддаватися впливу рівнів, які можуть погіршити здоров'я. При низьких рівнях гази можуть дратувати очі, ніс і горло. При більш високих рівнях гази можуть викликати прискорене дихання, головний біль, запаморочення, набряк і спазм горла, а також задуху.

Актуальність. Прогнозування вивержень вулканів заздалегідь до їх фактичного виверження — дуже важлива та актуальна проблема на сьогоднішній день, яка історично здавалася дуже важкою та майже нерозв'язною. Лише одне непередбачене виверження може призвести до десятків тисяч загиблих. Якби вчені могли достовірно передбачити, коли наступне виверження вулкана, евакуація могла б бути більш вчасною, а збитки від катастрофи значно б зменшилися.

Дослідження поведінки вулканів — дуже важлива тема багатьох наукових робіт, найкращі світові університети витрачають десятиліття, щоб принести хоча б мінімальний вклад в дослідженні цієї теми. Дана робота стане безцінним доповненням серед академічних робіт за даною тематикою для вирішення найглобальніших проблем людства.

Мета: дослідити поведінку вулканів, розробити математичний апарат для тлумачення предикторів незвичайних подій в контексті вулканів для систематизації залежності від них часу до виверження.

Методи дослідження: використання теоретичної гіпотези можливості передбачення виверження вулканів за сейсмічними сигналами, теоретичне моделювання можливостей машинного навчання в підході до подібних проблем, вивчення наукових матеріалів кращих світових університетів за даним питанням, емпіричні експерименти для перевірки допущених гіпотез, оптимізація побудованої моделі научно обгрунтованими метриками виміру ефективності результатів.

Наукова новизна та практичне значення: результати цієї роботи дозволять по-новому поглянути на проблему стихійних лих. Отримані результати неймовірно важливі не лише для наукової спільноти в подальших дослідженнях, а і в практичному застосуванні для порятунку життів людей. Розроблена модель може бути імплементована як у приватний програмний продукт, так і для подальших наукових досліджень. Перевага цієї моделі, в першу чергу, полягає в часі, затраченому на прогнозування. Зазвичай прогнозування вулканів роблять висококваліфіковані аналітики та науковці, які витрачають багато часу та ресурсів, які могли би бути застосовані в більш важливих задачах, які не можна виконати автоматично, так як тепер прогнозування вулканів можна зробити за допомогою запропонованої моделі.

Об'єкт дослідження: патерни зміни поведінки сейсмічних сигналів напередодні виверження вулканів.

Предмет дослідження: інформаційні засоби, методи, моделі управління даними при аналізі і прогнозуванні зміни патернів поведінки сейсмічних сигналів.

Завдання: окреслити проблему виверження вулканів, визначити приблизні стратегії дослідження, знайти необхідні історичні дані активних вулканів, виявити патерни зміни їх поведінки та побудувати на їх основі математичну модель для прогнозування часу.

Результати роботи можуть стати дуже впливовим інструментом для завчасного інформування про природну катастрофу, що наближається, попередження негативних наслідків та значного зменшення кінцевих проблем.

З достатнього попередження, райони навколо вулкана можуть бути безпечно евакуйовані до їх знищення. Сейсмічна активність є хорошим індикатором майбутнього виверження, але для покращення довгострокового передбачення необхідно визначити найвпливовіші фактори. Вплив даної роботи неоціненний та відчутний в усьому світі: десятки тисяч життів були би

врятовані завдяки більш передбачуваним виверження вулканів та ранній евакуації.

РОЗДІЛ 1 ПОСТАНОВКА ЗАДАЧІ ТА АНАЛІЗ ЇЇ ВИРІШЕННЯ

1.1 Визначення термінів

Сейсмічність: Поява землетрусів у певній місцевості.

Дегазація: Виділення газів, зазвичай пов’язаних з магмою, що знаходиться нижче, з кратера або інших частин вулкана.

Деформація: Зміна форми; частина вулкана над поверхнею Землі або скелі під ним можуть деформуватися.

Неспокій: Стан вулкана із сейсмічністю, дегазацією та деформацією, які можуть передувати виверженню.

Детермінований: Точне передбачення (так чи ні), чи відбудеться подія (наприклад, виверження).

Імовірнісний: Невизначений прогноз, що визначається як відсоток, що виражає ймовірність того, що подія відбудеться.

Лінійна поведінка: Коли система (наприклад, вулкан) веде себе очікуваним чином.

Нелінійна поведінка: Коли система (наприклад, вулкан) веде себе несподівано.

1.2 Постановка задачі

Основною задачею даної роботи є дослідження регресійної залежності часу до виверження вулкану від активності його сейсмічних сигналів.

Виявлення вивержень вулканів до того, як вони відбулися — це важлива проблема, яка історично здавалася дуже важкою.

У цій роботі, за допомогою напрацьованої експертизи у сфері машинного навчання, буде передбачено, коли відбудеться наступне виверження вулкана. Буде проаналізовано великий набір геофізичних даних, зібраний датчиками, розміщеними на діючих вулканах.

Навчена модель виявлятиме ознаки сейсмічних хвиль, які характеризують розвиток виверження.

Виверження вулканів є вражаючою демонстрацією діяльності нашої планети. Хоча деякі виверження можна безпечно спостерігати на відстані, багато виверженъ, особливо якщо вони вибухонебезпечні, можуть бути небезпечними для населення та навколишнього середовища навколо вулкана, включаючи тварин, рослини та штучні споруди.

Щоб зменшити шкоду, спричинену виверженнями, вчені, викликані вулканологами, намагаються передбачити виверження.

Хоча вулкани зазвичай дають кілька видів попереджень перед виверженням, не існує єдиного попереджувального сигналу, який дозволив би вулканологам точно передбачити кожне виверження.

Натомість дані з різних типів інструментів моніторингу об’єднуються, щоб допомогти вченим прогнозувати виверження принаймні за кілька днів наперед.

У наступні десятиліття, оскільки вулканологи вдосконалюють системи моніторингу та краще розуміють процеси, що відбуваються всередині вулканів, точніші прогнози виверження стануть можливими.

1.3 Аналіз фактичного стану досліджуваної проблеми

Наша планета відчуває кілька типів природних небезпек, включаючи землетруси, виверження вулканів, зсуви, цунамі, урагани та торнадо.

Виверження вулканів можуть бути різного розміру, причому найбільше є найбільш руйнівним з природних небезпек, здатних вплинути на всю планету.

На щастя, ці руйнівні явища рідкісні — зазвичай відбуваються приблизно раз на 100 000 років, тоді як найчастіші виверження, що відбуваються приблизно щотижня, є помірними і зачіпають лише невеликі території. На рис. 1.1 зображено Вулкан Сент-Хеленс, що у США, до і після виверження 1980 року (зображення з USGS).

Вибухове виверження знищило вершину та бічні частини вулкана. Воно також знищило навколишній ліс, включаючи дерева на передньому плані, залишивши ковдру попелу, утворену магмою, що вибухнула [1].

Рисунок 1.1 — Вулкан Сент-Хеленс до і після виверження 1980 року

Рисунок 1.2 демонструє виверження вулкана Усу в Японії у 2000 році. Воно змінило грунт під будинком, покрило дах вулканічним попелом і пронизило дах падаючими бомбами з магми.

Рисунок 1.2 — виверження вулкана Усу в Японії у 2000 році

На Землі близько 600 діючих вулканів, в основному вздовж кордонів рухомих тектонічних плит, які утворюють зовнішню оболонку нашої планети. Підраховано, що близько 800 мільйонів людей — десята частина населення Землі — живуть поблизу діючих вулканів.

Вулканічна діяльність забрала життя майже 300 000 жертв за останні чотири століття [1].

Останньою великою руйнівною подією було виверження Невадо-дель-Руїс в Колумбії в 1985 році. Це виверження було лише помірного розміру, але воно спричинило понад 23 000 жертв за лічені хвилини.

Навіть якщо вони не призводять до смерті людей, виверження можуть впливати на навколишнє середовище, включаючи тварин, рослини та штучні споруди, такі як будівлі, дороги, залізниці, аеропорти, фабрики та електростанції.

Наприклад, під час виверження вулкану Eyjafjallajokull в Ісландії в 2010 році хмара попелу від виверження поширилася по атмосфері і на кілька днів

зупинила повітряний рух над Європою, залишивши 10 мільйонів пасажирів на мілину, що призвело до втрати 5 мільярдів доларів (США).

Щоб зменшити вплив виверженъ вулканів на людей і навколишнє середовище, вчені, яких називають вулканологами, повинні мати можливість передбачати, коли і де збираються вивергатися вулкани.

Прогнозування виверженъ вулканів дозволяє попередити людей, щоб вони могли евакуюватися, зменшуючи негативний вплив виверженъ на населення.

Щоб спрогнозувати виверження, вулканологи повинні спочатку зрозуміти, як працюють вулкани.

Вулкани, як правило, перебувають у безшумному, «спокійному» стані, коли магма під ними не рухається і не накопичується. Іноді магма піднімається до поверхні і, коли вона перестає підніматися, вона накопичується в породі, що складає земну кору. Це скупчення магми в кінцевому підсумку утворює магматичний зал.

Магматична камера може змінювати форму навколишньої породи і створювати тріщини в земній корі, які можуть викликати землетруси. Землетрусна активність називається сейсмічністю. Виникнення землетрусів у певній місцевості.

Розриви також служать шляхами, що дозволяють газам виходити з магми в повітря. Це називається дегазацією. Вивільнення газів, зазвичай пов’язаних з магмою внизу, з кратера або інших частин вулкана. Магма, що накопичується, повільно збільшує вулкан через деформацію, зміну форми; частина вулкана над поверхнею Землі або скелі під нею можуть деформуватися.

Коли вулкан переходить від фази безшумності до фази «збудження», коли відбуваються сейсмічність, дегазація та деформація, це називається неспокійним станом вулкана із сейсмічністю, дегазацією та деформацією, які можуть передувати виверженню.

Фаза хвилювання може тривати від днів до місяців , і це вважається попередженням від вулкана, що він нестабільний і може бути виверження. Більшості вивержень передують хвилювання, але не кожен епізод заворушень закінчується виверженням — іноді після хвилювання вулкан може знову затихнути. Отже, хоча хвилювання не завжди передують виверженням, його поява є головною підказкою для вулканологів для прогнозування вивержень.

Щоб зробити надійний прогноз виверження, вулканологи повинні зрозуміти, що саме відбувається під час хвилювань, аналізуючи сигнали, які посилає вулкан.

Заворушення зазвичай виявляють і вивчають за допомогою точної системи моніторингу, що складається з групи чутливих інструментів, які фіксують активність у реальному часі всередині вулкана, зокрема рух магми. Тому моніторинг сейсмічності, дегазації та деформації, що викликається рухомою магмою, дозволяє вулканологам зрозуміти, що відбувається у вулкані під час хвилювань (рис 1.3).

Математичні моделі використовують цю інформацію моніторингу для визначення розміру, форми та розташування магми, що викликає заворушення [1].

Часто використовувані методи моніторингу вулканів включають GPS, InSAR та нахилу, які є методами вимірювання деформації; зйомка, при якій вулканологи безпосередньо досліджують вулкан; тепловізійна система, яка виявляє коливання температури на вулкані.

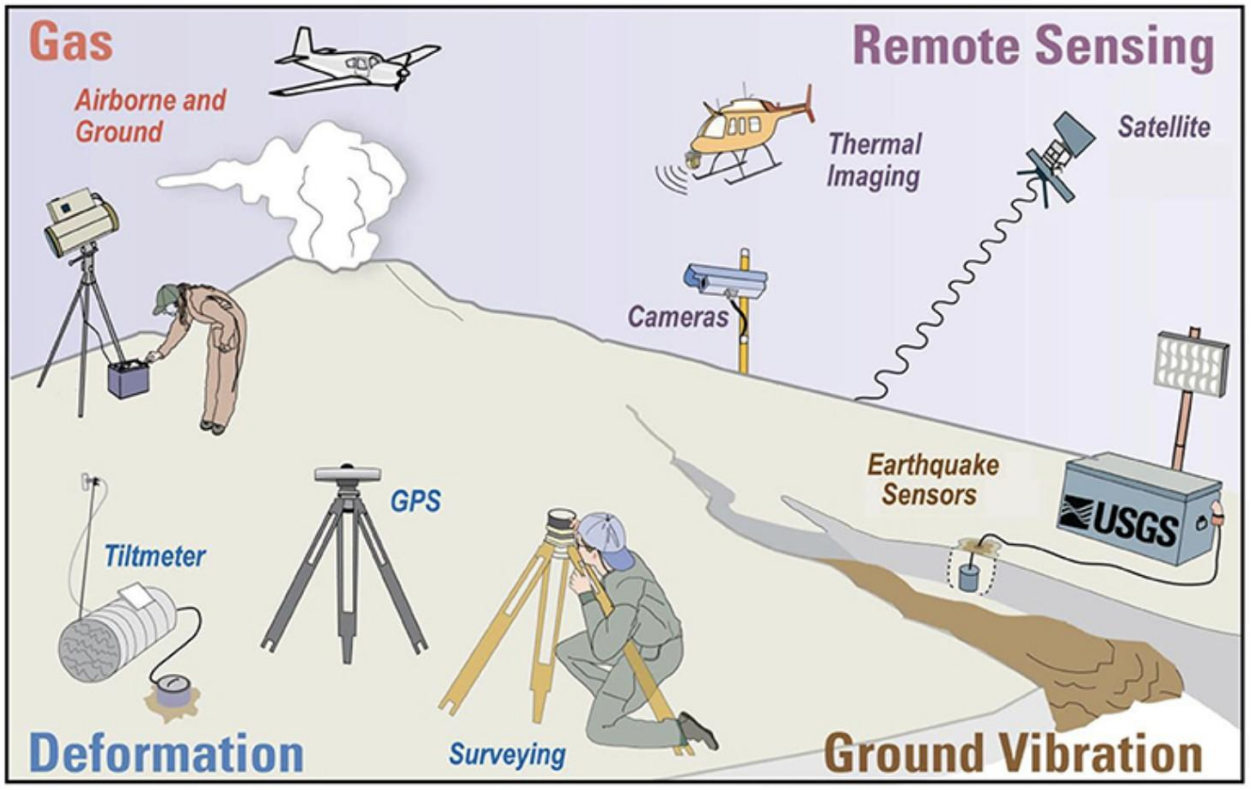

Рисунок 1.3 – Моніторинг вулканів, геологічна служба США

Хоча багато вулканів контролюються, деякі вулкани, особливо у віддалених районах, контролюються погано або взагалі не контролюються. Прогноз вивержень, очевидно, залежить від наявності системи моніторингу, яка може виявити хвилювання.

Вулканологи використовували два види підходів для прогнозування вивержень.

По-перше, вони намагалися точно передбачити, чи буде виверження. Цей підхід називається детермінованим точним прогнозом, чи відбудеться подія.

Зовсім нещодавно вони намагалися спрогнозувати, чи ймовірне виверження, використовуючи відсоток ймовірності. Цей підхід називається ймовірнісним невизначеним прогнозом, який визначається як відсоток, що виражає ймовірність того, що подія відбудеться.

Тільки під час моніторингу даних, зібраних під час хвилювань, спостерігається лінійна поведінка, коли система (наприклад, вулкан) веде себе очікуваним чином.

Виверження можна передбачити також за допомогою детерміністського підходу.

Великі виверження гори Сент-Хеленс (США) у 1980 році та Пінатубо (Філіппіни) у 1991 році були передбачені за кілька днів наперед. Лінійна поведінка означає посилення сигналів вулкана, і обидва епізоди заворушень на горі Сент-Хеленс і Пінатубо показали цю лінійну поведінку.

На жаль, хвилювання, які передують багатьом виверженням, демонструють нелінійну поведінку, коли система веде себе несподівано: збільшення інтенсивності сигналів супроводжується одним або кількома зменшенням і, нарешті, виверженням.

За цих нестабільних умов передбачити виверження детермінованим способом неможливо.

Нелінійна поведінка показує нам складність вулканічної системи, в якій багато процесів, лише деякі з яких можна виявити за допомогою моніторингу, відбуваються одночасно.

Через цю складність прогнозування виверженнь зазвичай робиться за допомогою імовірнісного підходу, в якому можливість виверження виражається у відсотках [2, 3].

Цей підхід, який також використовується в прогнозуванні погоди, точніше пояснює нестабільну поведінку вулканів.

Отже, щоб уникнути помилок, мудрий вулканолог повинен прогнозувати (імовірнісний підхід), а не прогнозувати (детермінований підхід) виверження.

За останні десятиліття деякі виверження були успішно прогнозовані, а інші — ні.

Успішне прогнозування призвело до успішної евакуації, тоді як неточне прогнозування призвело до евакуації, після якої не було виверження, або до пропущеної евакуації, коли виверження відбуваються до евакуації людей. В результаті вулканологія досягла успіхів і впевненості в прогнозуванні вивержень, а також катастроф і розчарувань. Наша здатність прогнозувати виверження все ще обмежена, приблизно 20% вивержень прогнозовано точно.

За останні кілька десятиліть було зроблено лише помірні покращення, незважаючи на широке використання та постійне вдосконалення інструментів моніторингу [4].

Існує два типи вулканів, які впливають на спосіб прогнозування вивержень: вулкани із закритими каналами (рис. 1.4) та вулкани з відкритими каналами (рис. 1.5).

Вулкани із закритими каналами затвердили магму на шляху, по якому магма рухається до поверхні, що відокремлює розплавлену магму від поверхні Землі.

Це призводить до накопичення магми всередині вулкана, що збільшує тиск і розбиває навколишні гірські породи, породжуючи тріщини і землетруси, а також деформує поверхню вулкана.

У вулканах із закритими каналами хвилювання складаються із сейсмічності та деформації поверхні.

І навпаки, вулкани з відкритими каналами, які зустрічаються рідше, мають розплавлену магму, що заповнює вулканічний канал, майже досягаючи поверхні.

Безперервна подача магми зсередини вулкана запобігає застиганню каналу, що викликає часті виверження.

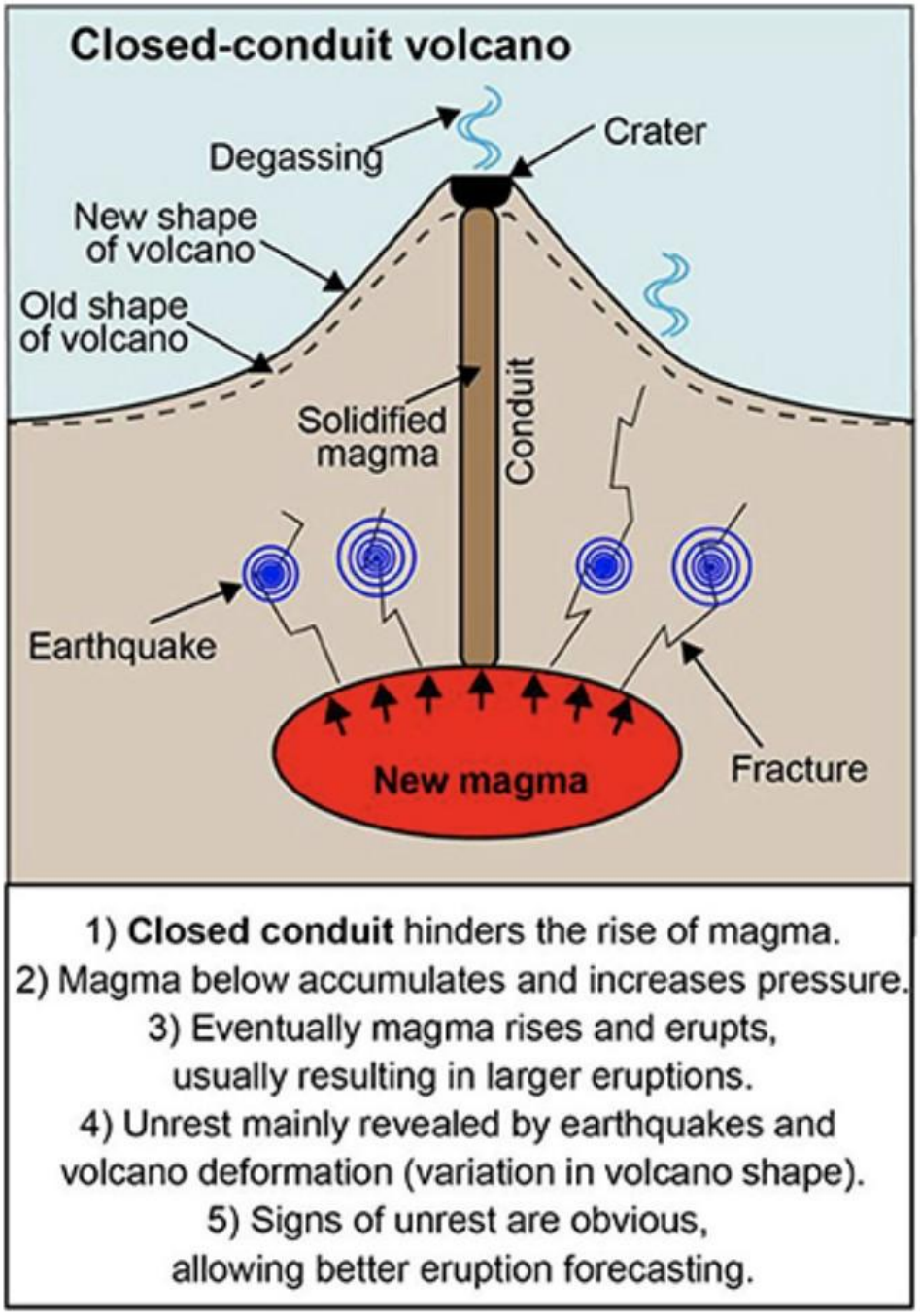

Рисунок 1.4 — Вулкан із закритим каналом

У вулканах з відкритими каналами магма не накопичується всередині вулкана, тому тиск не створюється для руйнування навколишньої породи або деформації вулкана, хоча через відкритий канал часто відбувається дегазація. Це призводить до слабших показників хвилювань. Тому, враховуючи сильніші

сигнали, прогнозувати виверження зазвичай легше для вулканів із закритими каналами.

Тим не менш, деякі добре відстежувані вулкани з відкритими каналами, які часто вивергаються, все ще можуть надати достатньо даних для надійного прогнозу, наприклад, гора Етна (Італія), виверження якої наразі прогнозуються з 97% успіху. Оскільки глобальний рівень успішності прогнозування для всіх вулканів становить ~20%, прогноз гори Етна є надзвичайно успішним.

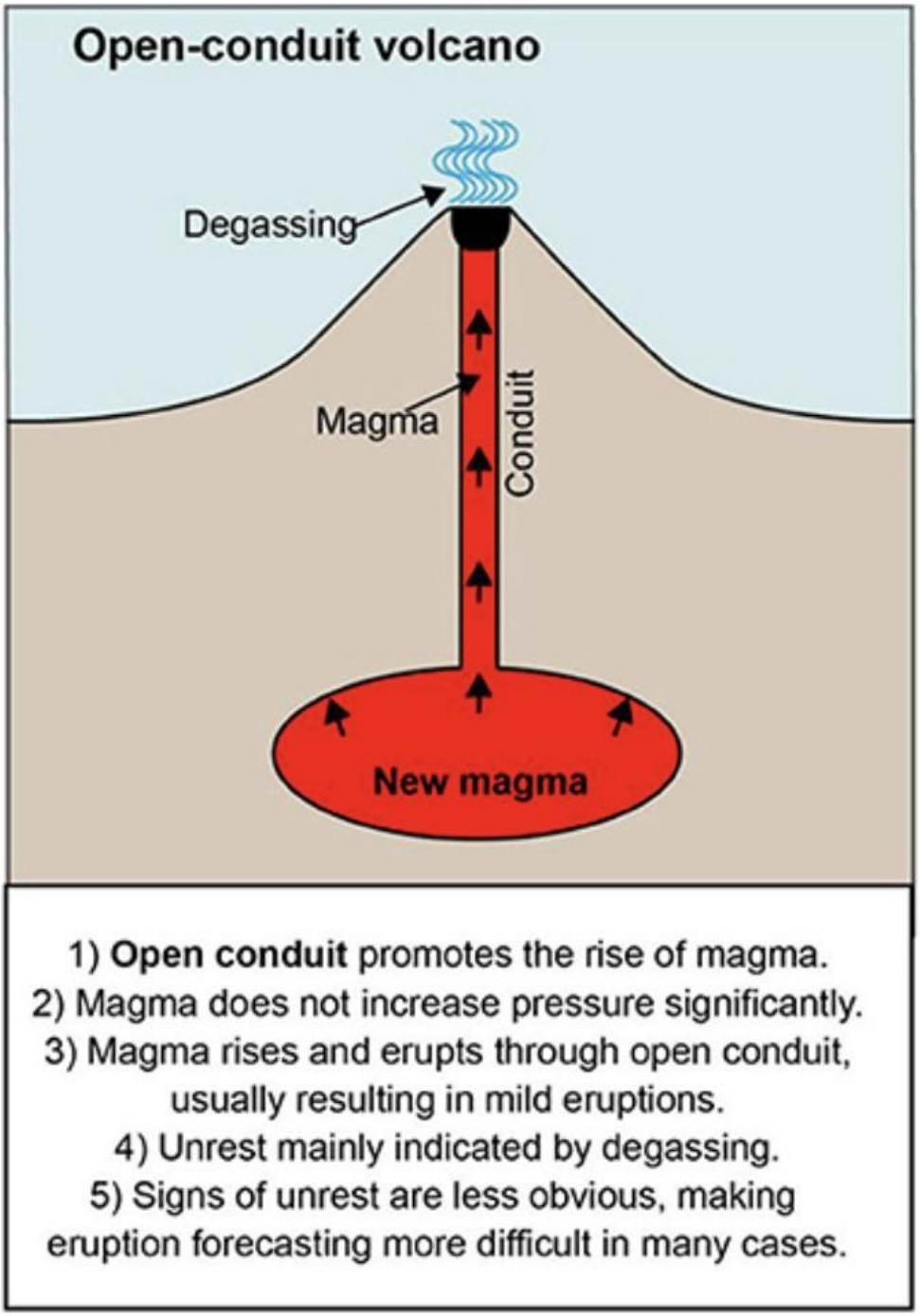

Рисунок 1.5 — Вулкан з відкритим каналом

Під час своєї історії виверження один вулкан може перейти із закритого каналу в відкритий.

Вулканологи постійно шукають нові підходи до покращення прогнозування, поєднуючи дані моніторингу зі знаннями про вулканічні процеси.

Щоб визначити початок виверження, останні дослідження зосереджені на появі конкретних сигналів моніторингу, які безпосередньо пов’язані з підйомом магми, а не з її накопиченням.

Це допоможе дослідникам краще прогнозувати час, коли виверження, ймовірно, відбудеться незабаром.

Оскільки деякі вулкани мають численні кратери, що вивергаються, розкидані на широких просторах, дослідники також намагаються спрогнозувати можливі місця розташування майбутніх кратерів, щоб зменшити вплив вивержень у населених районах [1].

Ці зусилля, безумовно, покращать прогнозування майбутніх вивержень, забезпечуючи більш надійні ймовірнісні оцінки та знижуючи рівень невизначеності чи помилки.

Тим не менш, необхідно пам’ятати, що невизначеність є частиною будь-якого прогнозу, тому прогнозування майбутнього виверження буде схоже на прогноз погоди [5].

Підсумовуючи, хоча прогнозування вивержень наразі є складним, але не нездійсненним завданням, майбутні дослідження та дослідження, а також збільшення доступності даних моніторингу дозволять краще зрозуміти, як працюють вулкани.

Це, в свою чергу, дозволить більш надійно прогнозувати виверження, зменшуючи негативний вплив вулканічної діяльності на населення та навколишнє середовище.

1.4 Аналіз існуючих методів вирішення задачі

Вчені використовують широкий спектр методів для моніторингу вулканів, включаючи сейсмографічне виявлення землетрусів і поштовхів, які майже завжди передують виверженням, точні вимірювання деформації ґрунту, яка часто супроводжує підйом магми, зміни викидів вулканічного газу та зміни сили тяжіння та магнітного поля.

Хоча ці методи не є окремою діагностикою, у поєднанні їх на вулканах, що контролюються, призвело до успішних прогнозів.

На вулкані Пінатубо (Філіппіни) в 1991 році успішний прогноз врятував тисячі життів.

Програма безпеки вулканів USGS зазначає, що ключ до точного короткострокового прогнозу виверження — це можливість розпізнавати, коли такі дані моніторингу показують постійні зміни від нормального фонового рівня активності.

Прогнози на основі моніторингу стають набагато надійнішими, але залишаються недосконалими.

Якщо вченим пощастило, попередники виверження йдуть так само, як і до попередніх вивержень.

Проте шаблони часто змінюються, і спостерігається абсолютно нова поведінка.

Найкращі прогнози будуть засновані на інтеграції геологічної історії, моніторингу в реальному часі та глибокому розумінні внутрішніх водопровідних процесів конкретного вулкана. Навіть за умови найкращого моніторингу та інтерпретації надійні прогнози рідко можливі більш ніж за кілька днів до виверження.

Деякі прогнози вивержень вулканів засновані на інтервалах повторення вивержень, але вони, як відомо, ненадійні з двох причин:

1. Кілька вулканів достатньо добре вивчені, щоб забезпечити точну історію вивержень протягом багатьох сотень або десятків тисяч років, необхідних для встановлення надійний інтервал повторень;
2. Деякі вулкани зберігають ту саму поведінку протягом тривалого часу (частіше, як тільки стає очевидним повторюваний шаблон, вулкан змінює поведінку).

Обсерваторії вулканів роблять прогнози з великою обережністю, оскільки вони можуть мати величезний вплив на постраждале населення, у деяких випадках змушуючи людей залишати будинки, ферми та худобу. Неточні прогнози можуть призвести до непотрібних зобов'язань щодо дефіцитних ресурсів та/або підірвати довіру мешканців до майбутніх прогнозів.

Надійні прогнози, однак, можуть робити співробітники обсерваторії вулканів, які мають досвід інтерпретації їх моніторингу, який виявляє попередники виверження.

Більшість країн, де є вулкани, доручили створену обсерваторію, яку керує уряд або університет, надавати громадськості прогнози виверження. Усі ці обсерваторії є членами Всесвітньої організації вулканічних обсерваторій (WOVO).

Що, якби вчені могли передбачати виверження вулканів так само, як вони передбачають погоду? Хоча визначити дощ або яскравість днів наперед непроста задача, прогнози погоди стають точнішими в коротких часових масштабах. Подібний підхід до проблеми виверження вулканів може мати великий вплив.

Лише одне непередбачене виверження може призвести до десятків тисяч загиблих.

Якби вчені могли достовірно передбачити, коли наступне виверження вулкана, евакуація могла б бути більш вчасною, а збитки від катастрофи значно б зменшилися.

В даний час вчені часто визначають «час до виверження», досліджуючи вулканічні поштовхи за сейсмічними сигналами.

У деяких вулканах вони посилюються, коли вулкани прокидаються і готуються до виверження.

На жаль, закономірності сейсмічності важко інтерпретувати. У дуже активних вулканах сучасні підходи передбачають виверження на кілька хвилин наперед, але вони зазвичай невдалі при довгострокових прогнозах.

У цій роботі, використовуючи напрацьовану експертизу у сфері машинного навчання, буде передбачено, коли відбудеться наступне виверження вулкана.

Був проаналізований великий набір геофізичних даних, зібраний датчиками, розміщеними на діючих вулканах.

Навчена модель виявляє ознаки сейсмічних хвиль, які характеризують розвиток виверження.

Якщо інформувати про наступаюче виверження завчасно, райони навколо вулкана можуть бути безпечно евакуйовані перед їх знищенням. Сейсмічна активність є хорошим індикатором майбутнього виверження, але для покращення довгострокової передбачуваності необхідно визначити попередні закономірності.

Вплив даної роботи можна було б відчути в усьому світі: десятки тисяч життів були врятовані завдяки більш передбачуваним розривам вулканів та більш ранній евакуації.

1.5 Збір даних для вирішення задачі

Для подальшого аналізу та побудови моделей на базі математичного апарату, потрібно використати набір даних, що відображує історію виверження вулканів.

Набір даних має складатися екземплярів вивержених вулканів та їх відповідних сейсмічних прекурсорів.

За їх допомогою та використовуючи останні дослідження в сфері алгоритмів машинного навчання можна буде оцінити, скільки часу залишиться до наступного виверження.

Дані представлятимуть класичну систему обробки сигналів, яка протистояла традиційним методам.

Було проаналізовано велику кількість академічних, приватних та державних баз даних, останнім часом були зроблені зусилля, щоб зібрати та зробити загальнодоступними набори даних, що складаються з сигналів і пов’язаних з ними метаданих.

Проаналізувавши різні набори даних та спрогнозувавши можливі перспективи, в кінцевому рахунку було обрано одне з досліджень [6] Національного інституту геофізики та вулканології [7] “Італійський сейсмічний набір даних для машинного навчання”.

Детально набір даних, використаний у роботах [8] та [9] можна завантажити з центру даних землетрусів Південної Каліфорнії на веб-порталі [10]. Цей набір даних включає 4,8 мільйона часових рядів, записаних майже 700 приймачами від понад 270 000 землетрусів у південній Каліфорнії. Набір даних STEAD [11], зібраний включає 1,2 мільйона слідів ЗС, що включають 450 000 локальних землетрусів і 100 000 шумових вікон, зафіксованих більш ніж 2600 станціями в глобальному масштабі.

Набір даних LEN-DB [34] також є глобальним набором даних локальних землетрусів і включає 1,2 мільйона слідів форми 3C, половина з яких належить до землетрусів, а половина — до шуму.

Набір даних NEIC (Yeck and Patton, 2020) включає глобальні дані і був використаний Yeck et al. (2020), щоб навчити 1,3 мільйона приходів сейсмічної фази, використовуючи три окремі моделі згорткової нейронної мережі, щоб передбачити час прибуття, тип фази та відстань.

Результати, отримані Ross et al. (2018b), L. Zhu et al. (2019), W. Zhu et al. (2019), Мусаві та ін. (2020), а також Мусаві та Бероза (2020) є чудовими прикладами успішних застосувань машинного навчання, які можуть істотно покращити рівень виявлення землетрусів щодо більшості традиційних методів, що приведе до локалізації крихітних і раніше невиявлених землетрусів, покращуючи наші знання про неоднорідність. зняття стресу від відомих і невідомих несправностей.

Ця розширена інформація має вирішальне значення для більш ретельної оцінки поточної сейсмотектоніки та сейсмічної небезпеки. Методи машинного навчання, ймовірно, стануть незамінним інструментом у сейсмології, щоб отримати якомога більше інформації з великої кількості даних, які вже зберігаються в архівах.

Серед непрямих переваг покращене виявлення може певною мірою також регулювати ущільнення мережі з значним скороченням інвестицій у обладнання та витрат на технічне обслуговування.

РОЗДІЛ 2 МАТЕМАТИЧНІ МОДЕЛІ ТА МЕТОДИ ДОСЛІДЖЕННЯ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ

2.1 Методи збору даних для дослідження

Проаналізувавши різні набори даних та спрогнозувавши можливі перспективи, в кінцевому рахунку було обрано одне з досліджень [6] Національного інституту геофізики та вулканології [7] “Італійський сейсмічний набір даних для машинного навчання”.

Цей інститут фінансується Міністерством освіти, університетів і досліджень Італії [12].

Його основними обов’язками в рамках системи цивільного захисту Італії є обслуговування та моніторинг національних мереж щодо сейсмічних та вулканічних явищ, а також інформаційно-просвітницька діяльність для італійського населення.

В інституті працює близько 2000 людей, розподілених між штаб-квартирою в Римі та іншими відділами в Мілані, Болоньї, Пізі, Неаполі, Катанії та Палермо.

До складу національного інституту геофізики та вулканології входить до 20 найкращих науково-дослідних установ за обсягом виробництва наукових публікацій [13].

Він бере участь та координує декілька дослідницьких проектів ЄС та організовує міжнародні наукові зустрічі у співпраці з іншими установами[14][15][16].

Дані про форму хвилі землетрусу в Італії зібрані у набір даних, який підходить для програм аналізу машинного навчання.

Набір даних складається з майже 1,2 мільйона трикомпонентних осцилограм приблизно 50 000 землетрусів і понад 130 000 слідів сигналів трикомпонентного шуму, загалом приблизно 43 000 год даних і в середньому 21 трикомпонентний слід за 1 подію.

Список землетрусів базується на Італійському сейсмічному бюлетені [17] національного інституту геофізики та вулканології з січня 2005 року по січень 2020 року і включає події з магнітудою від 0,0 до 6,5.

Дані форми хвилі були записані переважно Італійською національною сейсмічною мережею (код мережі IV) і включають записи як слабких (канали HH, EH), так і записів сильного руху (канали HN).

Усі сліди форми сигналу мають довжину 120 с, відбираються на частоті 100 Гц і надаються як у підрахунках, так і в фізичних одиницях руху землі після деконволюції передавальних функцій приладу.

Набір даних [18] форми хвилі супроводжується метаданими, що складаються з більш ніж 100 параметрів, що надають вичерпну інформацію про джерело землетрусу, станції запису, особливості трасування та інші отримані величини.

Цей багатий набір метаданих дозволяє користувачам обирати дані для власних цілей.

Більшу частину цих метаданих можна використовувати як позначки в аналізі машинного навчання або для інших досліджень.

2.2 Аналіз методів вирішення задачі

Важливих проривів у розумінні явищ землетрусів можна досягти шляхом аналізу дуже великої кількості безперервних записів сигналів, що зберігаються в існуючих сейсмічних архівах.

З цією метою може бути важливим зробити доступними добре організовані репрезентаційні підмножини архівів разом із пов’язаною з ними інформацією метаданими.

Останні розробки програмних платформ машинного навчання, таких як TensorFlow [19] , PyTorch [20] , Keras [21] , Caffe [22], наявність високопродуктивного обчислювального обладнання, наприклад, графічних процесорів і доступ до ретельно відібраних контрольних наборів даних, таких як STEAD [23] і LEN-DB [24] пропонують нові можливості застосування методологій машинного навчання до сейсмологічних проблем та землетрусів.

Зокрема, використання складних та оптимізованих алгоритмів машинного навчання для аналізу великої кількості сейсмічних даних може призвести до значних удосконалень для автоматизованих завдань, таких як вибір форми сейсмічної хвилі, прогнозування руху землі та раннє попередження про землетруси; для виявлення прихованих сигналів, які в даний час розпізнаються як шум; або для нових стратегій моделювання та інверсії [25] [26] [27].

Зокрема, поява машинного навчання у сфері сейсмології підкреслила важливість контрольних наборів даних для порівняння розроблених методологій, а також посприяла більш ретельним і статистично обґрунтованим схемам для аналізу даних, таких як поділ усіх доступних даних на навчання, перевірку, і тестові набори.

Застосування методів машинного навчання до даних сейсмологічної форми сигналу може бути досить простим. Дійсно, велика кількість маркованих даних уже доступна завдяки аналізу, який протягом багатьох десятиліть проводили експертні аналітики, які склали та переглянули каталоги землетрусів (які включають показання фази початку, місце землетрусу та оцінки розмірів) або які зібрали параметри руху ґрунту у спеціальних плоских файлах і картах сильного руху ґрунту серед найпоширеніших завдань.

Їхня робота ефективно забезпечує метадані, які можна пов’язати із записаними формами сигналів і які можна використовувати як мітки під час виконання аналізу машинного навчання.

Однак основним вузьким місцем у широкомасштабній реалізації машинного навчання є швидкий доступ до сигналів і пов’язаних з ними метаданих.

Архіви з відкритим доступом, доступні сейсмологічній спільноті, наприклад, EIDA [28] або IRIS [29] , були в основному розроблені для збереження безперервних даних і надання їх науковому співтовариству. На практиці однією з головних цілей сейсмологічних центрів обробки даних було безперебійне отримання безперервних даних з мереж і збереження, зберігання та архівування всього запису безперервних сигналів. У цьому контексті користувачі мають повну гнучкість у виборі даних для завантаження, але доступ до великих обсягів даних може зайняти дуже багато часу.

Таким чином, незважаючи на досягнення, досягнуті за останні десятиліття з впровадженням добре перевірених та ефективних веб-сервісів, доступність віддалених серверів все ще залишається громіздкою [30].

Звідси випливає, що для залучення ширшої аудиторії користувачів і розробників існує гостра потреба в збиранні та публікації контрольних наборів даних, які можна легко використовувати з існуючими програмними платформами [11].

У практичному плані справа полягає в збиранні перевірених якісних даних і метаданих відповідно до обсягу та форматів, готових до використання в програмах машинного навчання.

Загалом, вражаюча продуктивність програм машинного навчання тісно пов’язана з доступністю великих обсягів даних із належним чином позначеними метаданими. Великі обсяги даних мають вирішальне значення для правильного навчання та уникнення переобладнання даних.

Збір даних під назвою INSTANCE збирає дані форми сейсмічної хвилі від станцій слабкого та сильного руху, які були вилучені з італійського вузла EIDA [31].

Метадані, пов’язані з формами сигналів, витягуються з каталогу землетрусів італійського національного інституту геофізики та вулканології і з самих слідів сигналів.

2.3 Регресія часових рядів

Регресія часових рядів — це статистичний метод для прогнозування інформації на основі історичних даних (авторегресійна динаміка) та передачі динаміки від відповідних предикторів.

Регресія часових рядів може допомогти зрозуміти та передбачити поведінку динамічних систем на основі експериментальних даних або даних спостережень.

Загальні види використання регресії часових рядів включають моделювання та прогнозування економічних, фінансових, біологічних та інженерних систем.

Основна концепція полягає в прогнозуванні часових рядів, що представляють інтерес y, припускаючи, що він має лінійну залежність з іншими часовими рядами x.

Наприклад, можна спрогнозувати місячні продажі y, використовуючи загальні витрати на рекламу x як провісник. Або можна спрогнозувати щоденну потребу в електроенергії y, використовуючи температуру x1і день тижня x2 як провісники.

Змінну прогнозу y іноді також називають регресантом, залежною або пояснюваною змінною.

Змінні-провісники x іноді також називають регресорами, незалежними або пояснювальними змінними.

2.3.1 Лінійна модель

У найпростішому випадку регресійна модель допускає лінійну залежність між залежною змінною y та незалежною змінною x:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$, (2.1)

де коефіцієнти β0 і β1 позначають перетин і нахил прямої відповідно;

перехоплення β0 представляє прогнозоване значення y, коли x=0;

схил β1 представляє середню прогнозну зміну в y в результаті збільшення на одиницю x.

Штучний приклад даних такої моделі наведено на рисунку 2.1.

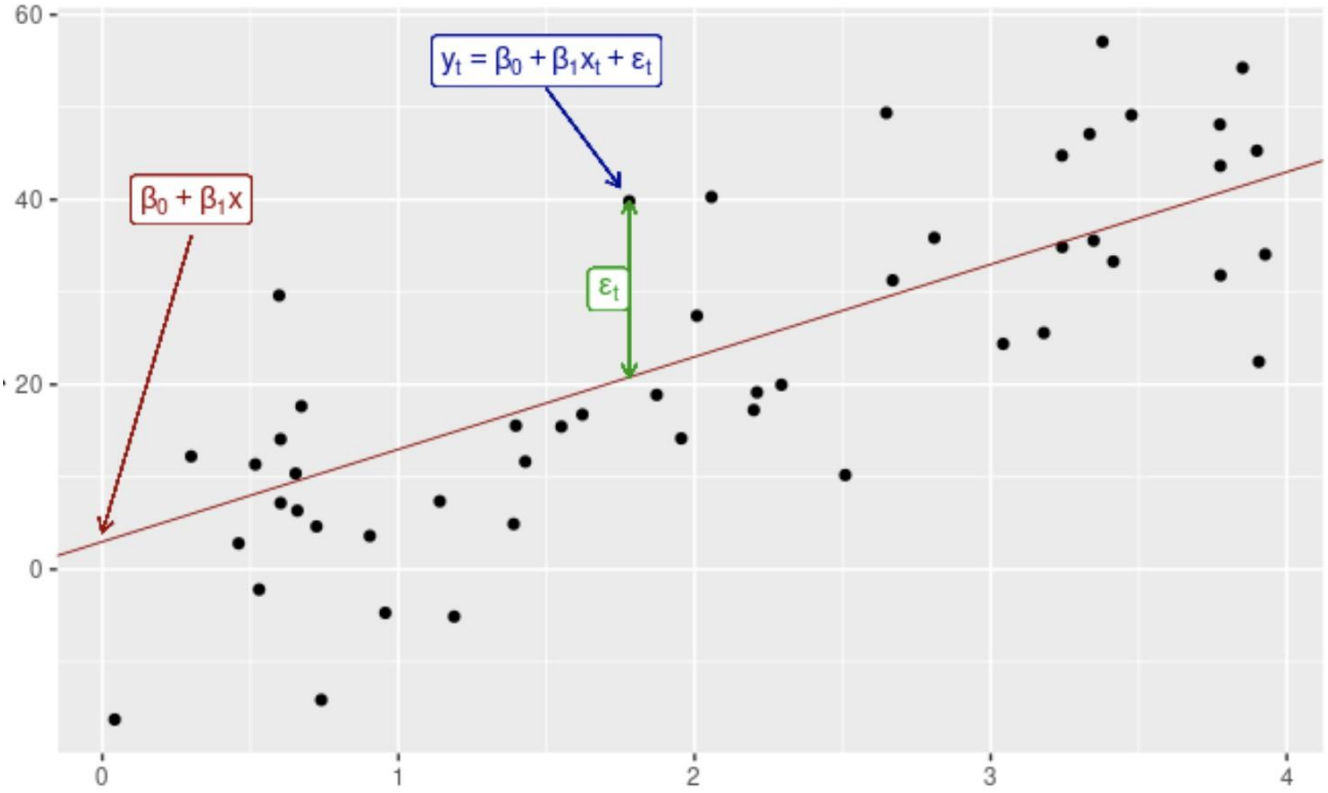

Рисунок 2.1 — Приклад даних із простої моделі лінійної регресії

Необхідно зауважити, що спостереження не лежать на прямій, а розкидані навколо неї. Можна думати про кожне спостереження yt як складову із систематичної або поясненої частини моделі, β0+β1xt і випадкова «помилка», εt.

Термін «помилка» означає не помилку, а відхилення від базової моделі прямої лінії. Він фіксує все, що може вплинути yt окрім xt. Якщо є дві або більше змінних-провісників, модель називається моделлю множинної регресії.

Загальна форма моделі множинної регресії наведена у формулі (2.2):

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t, \quad (2.2)$$

де y є змінною, яку потрібно прогнозувати;

x1, …, xk є k прогнозні змінні;

коефіцієнти β1, …, βk вимірюють ефект кожного предиктора після врахування ефектів усіх інших предикторів у моделі.

Кожна зі змінних-провісників має бути числовою. Таким чином, коефіцієнти вимірюють граничні ефекти провісних змінних.

Коли ми використовуємо модель лінійної регресії, ми неявно робимо деякі припущення щодо змінних на рис. 2.1. По-перше, ми припускаємо, що модель є розумним наближенням до реальності; тобто зв'язок між змінною прогнозу та змінними-провісником задовольняє цьому лінійному рівнянню.

По-друге, ми робимо наступні припущення щодо помилок (ε1, …, εt):

- вони мають нульове середнє; інакше прогнози будуть систематично необ'єктивними;
- вони не автокорельовані, інакше прогнози будуть неефективними, оскільки в даних є більше інформації, яку можна використати;
- вони не пов'язані зі змінними-провісниками, інакше було б більше інформації, яка повинна бути включена в систематичну частину моделі.

Також корисно, щоб помилки розподілялися нормально з постійною дисперсією σ2, щоб можна було легко створювати інтервали передбачення.

Інше важливе припущення в моделі лінійної регресії полягає в тому, що кожен предиктор x не є випадковою величиною. Якби ми проводили контрольований експеримент у лабораторії, ми могли б контролювати значення кожного x (щоб вони не були випадковими) і спостерігали за отриманими значеннями y.

За допомогою даних спостережень (включаючи більшість даних у бізнесі та економіці) неможливо контролювати значення x, тому ми робимо це припущенням.

2.3.2 Оцінка за найменшими квадратами

На практиці, звичайно, ми маємо сукупність спостережень, але ми не знаємо значень коефіцієнтів β0, β1, …, βk . Їх необхідно оцінити на основі даних.

Принцип найменших квадратів забезпечує спосіб ефективного вибору коефіцієнтів шляхом мінімізації суми квадратів помилок.

Тобто ми обираємо значення β0, β1, …, βk, які мінімізують оцінку за найменшими квадратами на формулі (2.3).

$$\sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 x_{1,t} - \beta_2 x_{2,t} - \cdots - \beta_k x_{k,t})^2. \tag{2.3}$$

Оцінка за найменшими квадратами дає найменше значення для суми квадратів помилок.

Знаходження найкращих оцінок коефіцієнтів часто називають «підгонкою» моделі до даних або іноді «навчанням» моделі. Таким чином була отримана лінія, яка показана на рисунку 2.2.

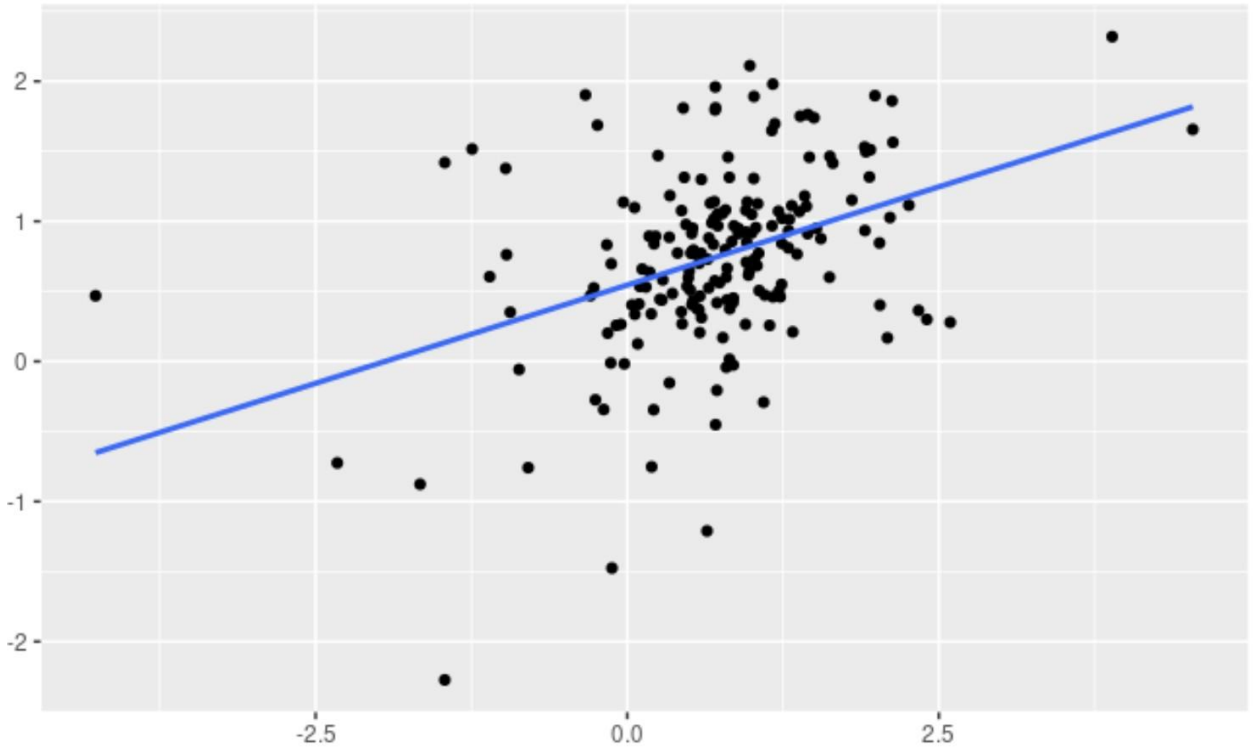

Рисунок 2.2 — Лінія регресії, отримана за допомогою оцінки найменших квадратів

Функція tslm() відповідає моделі лінійної регресії до даних часових рядів.

Вона схожа на функцію lm(), яка широко використовується для лінійних моделей, але tslm() надає додаткові можливості для обробки часових рядів.

Придатність. Поширеним способом підсумувати, наскільки добре модель лінійної регресії відповідає даним, є коефіцієнт детермінації або R2. Його можна розрахувати як квадрат кореляції між спостережуваними y та прогнозованими ^y значеннями. Крім того, його також можна розрахувати як показано у формулі (2.4).

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2},$$

(2.4)

Формула (2.4) зображує спосіб розрахунку коефіцієнту детермінації, де підсумовуються всі спостереження. Таким чином, він відображає частку варіації змінної прогнозу, яка враховується (або пояснюється) регресійною моделлю.

У простій лінійній регресії значення R2 також дорівнює квадрату кореляції між y та x (за умови, що було включено перехоплення).

Якщо прогнози близькі до фактичних значень, ми б очікували, щоб R2 був близьким до 1.

З іншого боку, якщо передбачення не пов’язані з реальними значеннями, то R2=0 (знову ж таки, припускаючи, що є перехоплення). У всіх випадках, R2 лежить між 0 і 1.

Значення R2 використовується часто, хоча часто неправильно, у прогнозуванні.

Значення R2 ніколи не зменшиться при додаванні додаткового предиктора до моделі, і це може призвести до перепідгонки. Немає встановлених правил щодо того, що є хорошим R2 значенням та типові значення R2 залежать від типу використовуваних даних.

Перевірка ефективності прогнозування моделі за тестовими даними набагато краще, ніж вимірювання R2 значення на даних навчання.

2.4 Регресор Catboost

Як проводиться навчання. Метою навчання є вибір моделі y, залежно від набору функцій x, що найкраще вирішує задану проблему для будь-якого вхідного об’єкта. Ця модель знайдена за допомогою навчального набору даних, який являє собою набір об’єктів з відомими ознаками та значеннями міток. Точність перевіряється на наборі даних перевірки, який містить дані в тому ж форматі, що й у наборі навчальних даних. Він використовується лише для оцінки якості навчання, але не для самого навчання.

CatBoost базується на деревах рішень із посиленням градієнта. Під час навчання послідовно будується набір дерев рішень. Кожне наступне дерево будується з меншими втратами порівняно з попередніми деревами.

Кількість дерев регулюється стартовими параметрами. Щоб запобігти перенавчанню, необхідно використовувати детектор виявлення перенавчання. Коли він спрацьовує, дерева перестають будуватися.

2.4.1 Детектор перенавчання

Якщо відбувається перенавчання, CatBoost може зупинити навчання раніше, ніж це диктують параметри навчання. Наприклад, його можна зупинити до побудови вказаної кількості дерев. Ця опція встановлюється в стартових параметрах.

Підтримуються такі методи виявлення перенавчання:

- IncToDec;
- Перевірка ітераціями.

2.4.2 IncToDec

Перед побудовою кожного нового дерева CatBoost перевіряє результуючу зміну втрат у наборі даних перевірки. Детектор переобладнання

спрацьовує, якщо Порогове значення, встановлене в початкових параметрах, більше ніж CurrentPValue формула (2.5).

$$CurrentPValue < Threshold \tag{2.5}$$

Як CurrentPValue розраховується з набору значень для максимального показника:

1. Спочатку розраховується ExpectedInc за формулою (2.6)

$$ExpectedInc = max_{i_1 \leq i_2 \leq i} 0.99^{i-i_1} \cdot (score[i_2] - score[i_1]) \tag{2.6}$$

2. Потім розраховується *x* за формулою (2.7)

$$x = \frac{ExpectedInc[i]}{max_{j \leq i} score[j] - score[i]} \tag{2.7}$$

3. І, нарешті, рахується CurrentPValue за формулою 2.8

$$CurrentPValue = exp\left(-\frac{0.5}{x}\right) \tag{2.8}$$

2.4.3 Перевірка ітераціями

Перед побудовою кожного нового дерева CatBoost перевіряє кількість ітерацій після ітерації з оптимальним значенням функції втрат. Модель вважається перенавченою, якщо кількість ітерацій перевищує значення, зазначене в параметрах навчання.

2.4.4 Етапи побудови одного дерева.

2.4.4.1 Попередній розрахунок сегментів.

2.4.4.1.1 Квантування

Перед навчанням можливі значення об’єктів поділяються на непересікаючі діапазони (секції), розмежовані пороговими значеннями (розділи). Розмір квантування (кількість розщеплень) визначається вихідними параметрами (окремо для числових ознак і чисел, отриманих в результаті перетворення категоріальних ознак у числові).

Квантування також використовується для поділу значень міток під час роботи з категоріальними ознаками. Для цього на великих наборах даних використовується випадкова підмножина набору даних. У таблиці 2.1 показано режими квантування, надані в CatBoost.

Таблиця 2.1 — РЕЖИМИ КВАНТУВАННЯ, НАДАНІ В CATBOOST [32]

Mode	How splits are chosen
Median	Include an approximately equal number of objects in every bucket.
Uniform	Generate splits by dividing the `[min_feature_value, max_feature_value]` segment into subsegments of equal length. Absolute values of the feature are used in this case.
UniformAndQuantiles	Combine the splits obtained in the following modes, after first halving the quantization size provided by the starting parameters for each of them: - Median. - Uniform.
MaxLogSum	Maximize the value of the following expression inside each bucket: $\sum_{i=1}^{n} \log(weight), where$ - n — The number of distinct objects in the bucket. - $weight$ — The number of times an object in the bucket is repeated.
MinEntropy	Minimize the value of the following expression inside each bucket: $\sum_{i=1}^{n} weight \cdot log(weight), where$ - n — The number of distinct objects in the bucket. - $weight$ — The number of times an object in the bucket is repeated.
GreedyLogSum	Maximize the greedy approximation of the following expression inside every bucket: $\sum_{i=1}^{n} \log(weight), where$ - n — The number of distinct objects in the bucket. - $weight$ — The number of times an object in the bucket is repeated.

Квантування виконується для кожної числової ознаки, щоб визначити можливі способи поділу даних на сегменти. Отримана інформація використовується для вибору структури дерева.

Спосіб квантування та кількість відерів задаються у стартових параметрах.

2.4.4.2 Вибір структури дерева

Елементи вибираються в порядку разом з їх розщепленнями для заміни в кожному аркуші. Відбір кандидатів здійснюється на основі даних попереднього розрахунку розщеплень. Глибина дерева та інші правила вибору конструкції задаються у стартових параметрах.

Як вибирається пара функція-розщеплення для листа:

1. Утворюється список можливих кандидатів (пар функція-розщеплення), які будуть призначені аркушу як розділенню.
2. Для кожного об’єкта розраховується ряд штрафних функцій (за умови, що всі кандидати, отримані на етапі 1, були призначені до аркуша).
3. Вибирається розкол з найменшим штрафом.
4. Отримане значення присвоюється листу.

Цю процедуру повторюють для всіх наступних листків (кількість листя повинна відповідати глибині дерева).

Перед побудовою кожного нового дерева виконується випадкова перестановка об’єктів класифікації. Для вибору структури наступного дерева використовується метрика, яка визначає напрямок подальшого вдосконалення функції.

Значення обчислюється послідовно для кожного об’єкта. У розрахунку використовується перестановка, отримана перед побудовою дерева – дані для об’єктів використовуються в тому порядку, в якому вони були розміщені до процедури.

3 РЕАЛІЗАЦІЯ ПРОГНОЗУЮЧОЇ МОДЕЛІ

3. 1 Підготовка та імпорт даних

3.1.1 Підготовка даних

Збір даних проводиться за основними етапами, переліченими нижче:

- відбір землетрусів;
- вибір станції;
- вибір і завантаження даних форми сигналу;
- перехресна перевірка між вибором станції на основі фази та завантаженими даними сигналу;
- обробка сигналів підрахунку даних;
- застосування інструментальної передавальної функції до сигналів.

3.1.1.1 Вибір землетрусу

Перший крок полягав у отриманні всіх землетрусів $M \geq 0$ з 1 січня 2005 р. по 31 січня 2020 р. у збільшеній зоні в межах кутів широти та довготи (35,0, 5,0) і (49,0, 19,0). Всього було виявлено 315 225 землетрусів. Початок запиту приблизно відповідає оновленням, оновленням та збільшенням кількості станцій національної сейсмічної мережі [33] [31] [34]. Приблизно в 2005 році мережа INGV зазнала серйозного оновлення, при цьому існуючі, переважно аналогові, прилади були замінені високоякісними цифровими реєстраторами сейсмічних даних і новими, переважно широкосмуговими (і деякий тривалий короткий період), трикомпонентними (3C) датчики. Вибрані станції також були доповнені додатковими датчиками сильного руху 3C. В результаті модернізації кількість станцій IV мережі збільшилася більш ніж у 2 рази. Крім того, з 2005

року було багато тимчасових розгортань сейсмічних станцій, що збігалися з сейсмічними послідовностями та конкретними експериментами, дані яких також доступні через вузол EIDA INGV [31]. Загальна кількість станцій також зросла завдяки внеску мереж, що належать іншим італійським установам (наприклад, Університету Генуї, Національному інституту океанографії та експериментальної геофізики (OGS) та Університету Неаполя та інших). Це призвело до значного покращення виявлення землетрусів низької магнітуди. У регіональному масштабі Італії повнота бюлетеня INGV становить приблизно ~M 1,7–M 1,8, хоча значні відмінності виникають залежно від регіону. У цьому відношенні переважною величиною каталогу INGV є локальна величина Ml (Richter, 1935), але іноді також Mw і Md.

Важливим аспектом під час складання цього набору даних, який можна буде використовувати для прогнозування, є збір збалансованого розподілу даних. У сейсмології, коли для класифікації використовується магнітуда землетрусів, неможливо досягти збалансованого представлення, оскільки землетруси невеликого розміру, відповідно до магнітуди Гутенберга-Ріхтера в порівнянні з законом кількості землетрусів [35], переважають більші землетруси. Щоб вирішити цю проблему, або, принаймні, пом’якшити її вплив, встановимо обмеження:

- переважна більшість землетрусів з $M \geq 4{,}0$ – землетруси, які були відкинуті (30) усі (крім 5) відбулися за межами кордонів італійської країни та переважно на Балканах (землетруси в Італії, всі з $M<5$, буде включено до майбутнього оновлення набору даних);
- землетруси з часами виникнення, що відрізняються більш ніж на 120 с в діапазоні $2{,}0 \leq M<4{,}0$;
- додаткові 20 000 землетрусів, вибраних випадковим чином, з часами виникнення, що відрізняються більш ніж на 120 с для $M<2{,}0$.

Отриманий розподіл землетрусів за їх магнітудою детально наведено в таблиці 3.1, де All — це загальна кількість землетрусів у бюлетені INGV за період з 1 січня 2005 року по 31 січня 2020 року, Selected та Percent kept відносяться до землетрусів, а Nb. 3C records відноситься до слідів осциллограми, включених до набору даних.

Таблиця 3.1 — ОСТАТОЧНИЙ ВИБІР ДАНИХ

$\geq M_{\min}$	$< M_{\max}$	All	Selected	Percent kept	Nb. 3C records
0	1	57 746	4462	7.73	39 794
1	2	209 652	15 249	7.27	202 572
2	3	43 109	30 845	71.55	757 129
3	4	4342	3106	71.53	139 338
4	5	342	315	92.11	18 659
5	6	31	28	90.32	1593
6	7	3	3	100.0	164
0	7	315 225	54 008	17.13	1 159 249

Самі землетруси наведені на рисунку 3.1.

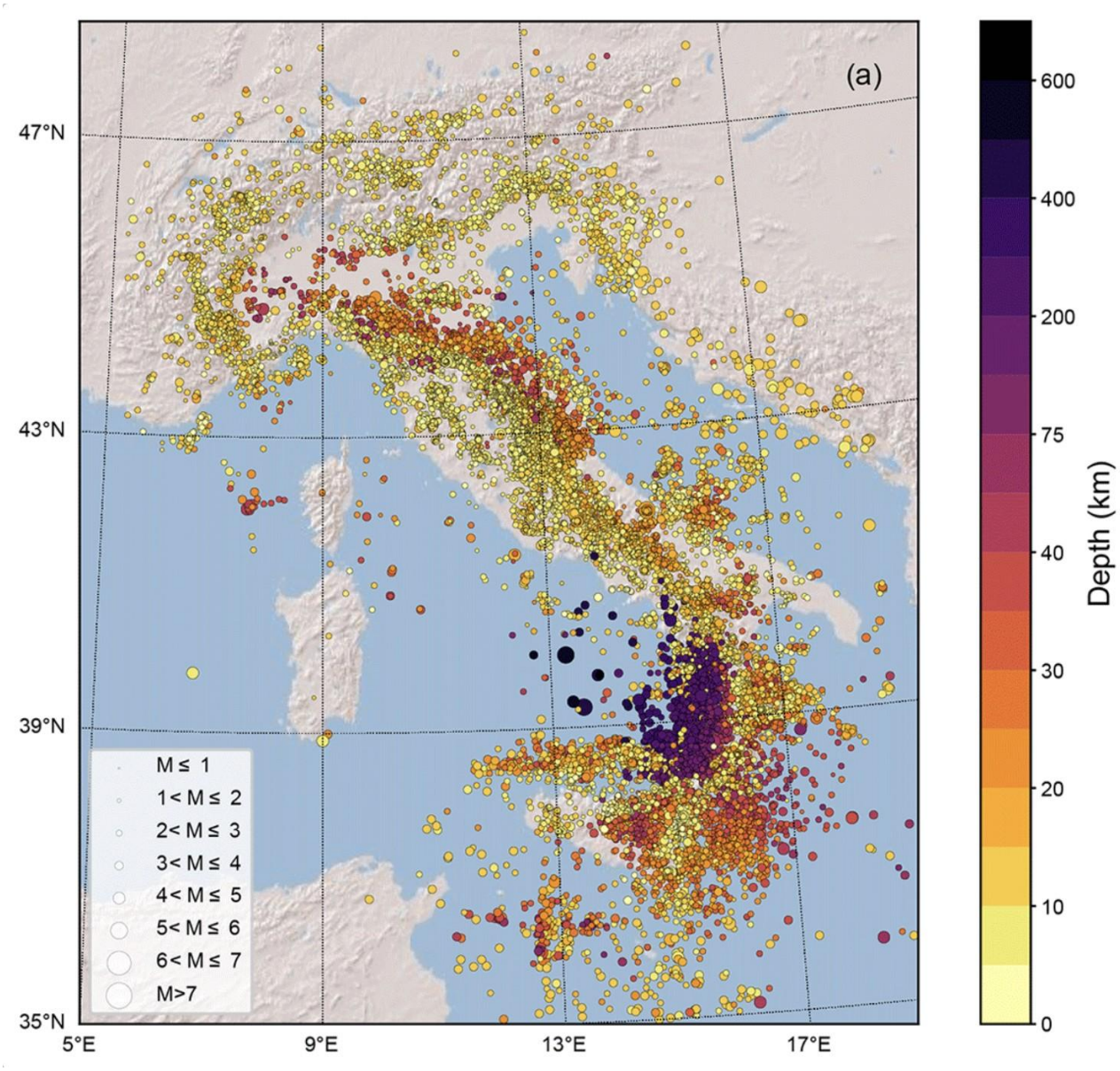

Рисунок 3.1 — Карта землетрусів, що входять до набору даних, показана у вигляді суцільних кіл з кольорами, вибраними відповідно до глибини. Розмір символу пропорційний магнітуді землетрусу.

3.1.1.2 Вибір станції

Щоб зібрати високоякісні сигнали землетрусу, необхідно звернути увагу на найбільш точно підібрані фази початку P- і S-хвиль, опублікованих у бюлетені INGV. У зв'язку з цим, ручний вибір етапів прибуття зазвичай

виконується групою з близько 20 висококваліфікованих співробітників INGV, які також перевіряють розташування гіпоцентрів і визначення величини перед публікацією бюлетеня. Ці місця, перевірені вручну, вказані як бажані рішення у бюлетені INGV. На практиці були обрані лише ті станції, які мали P- і, якщо доступні, S-хвилі вибірки, пов'язані з бажаним розташуванням бюлетеня INGV. Дані про сильний рух, надані національною мережею сильного руху, не входять до складу землетрусів і локації, які виконує персонал INGV, і ті самі дані не є доступні через EIDA. Однак вони можуть бути включені в майбутні випуски набору даних.

Підсумовуючи, було прийнято наступні критерії для визначення записів форми сигналу, які будуть включені в набір даних після застосування вибраного вище землетрусу:

- всі станції, які мають фази початку хвилі P (і фази початку хвилі S, якщо є), які використовуються для бажаного місця землетрусу (не робиться різниця між Pg та Pn і не вибираються вторинні фази, такі як PmP);
- всі станції з даними про форму сигналу, доступними через італійський вузол EIDA (див. мережі, що сприяють набору даних на круговій діаграмі рис. 3.2);

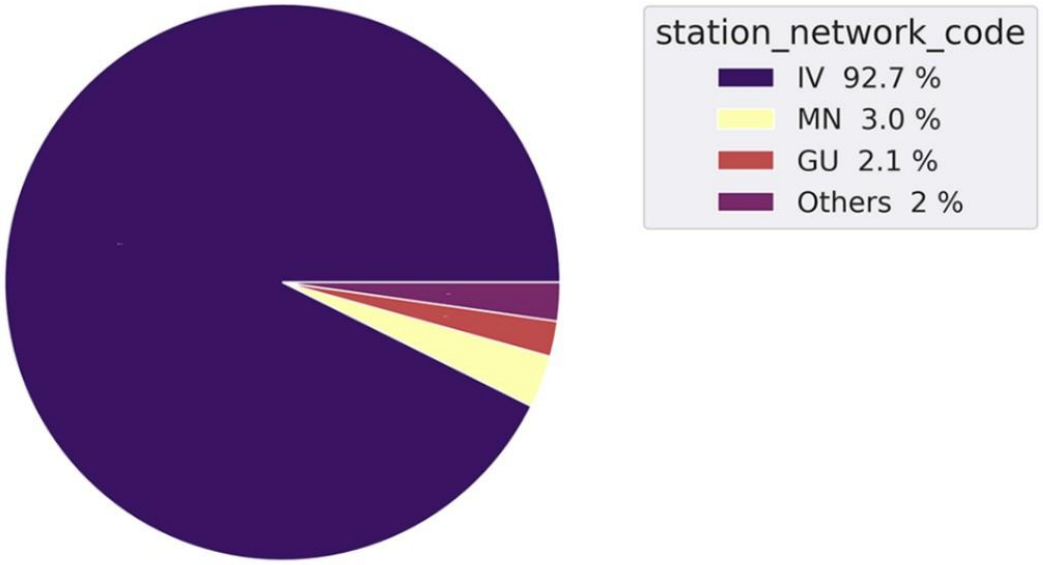

Рисунок 3.2 — Мережі, що сприяють набору даних

- залишковий час розташування P- і S-хвиль менше 1,0 с;
- фази P- і S-хвиль, які внесли вклад у локалізацію з вагою більше 10 %.

Ця процедура відбору зменшила кількість фаз P- і S-хвиль з ~1,9 до ~1,2 і з ~1,1 до ~0,7 мільйонів відповідно.

3.1.1.3 Вибір і завантаження даних форми сигналу

Процедура відбору, описана в підпункті 3.1.1.2, призвела до компіляції списку часових вікон даних сигналів, які потрібно завантажити з архіву безперервних сигналів EIDA. Обираємо часове вікно 120 с, щоб включати хвилі P і S від станцій, відстань яких становить до ~600 км від гіпоцентру. Дійсно, у цих випадках різниця в часі $S-P$ становить приблизно 75–80 с. Додавши приблизно 20 s сигналу до часу P-хвилі та приблизно 20 s після хвилі S, отримуємо вибір вікна 120 s, що забезпечує найбільш значущі сигнали

землетрусу або для найвіддаленіших станцій, у випадку глибини земної кори. землетруси, або ближчі станції, у разі глибоких землетрусів субдукції Калабрійської дуги.

Більш технічно, тимчасові вікна, встановлені для завантаження даних, були визначені шляхом вставки випадково вибраного часу буфера в діапазоні від 15 до 20 с до фази початку приходу зубця P і збільшення часового вікна до 125 с. Використання довгих вікон 125 s на етапі завантаження даних є довільним, оскільки після обробки даних тимчасові вікна були встановлені на 120 s. Цей критерій гарантував, що переважна більшість завантажених слідів осциллограми містила буферний час початку перед хвилі P між 15 і 20 с. Однак було виявлено, що при роботі з такою великою кількістю сигналів, отриманих різноманітними інструментами, налаштованими по-різному, можуть виникнути деякі розбіжності. На практиці, оскільки дані архівуються у стисненому форматі miniSEED, який містить різні розміри логічних записів, і оскільки веб-служба витягує повний логічний запис, що містить попередньо визначений час початку трасування, час початку трасування може бути раніше, ніж попередньо визначений мінімальний час 20 с (тобто в цьому випадку є більший інтервал часу між P-прибуттям і фактичним часом початку трасування). Навпаки, коли дані відсутні до часу початку зубця P (тобто за 15–20 секунд буферного часу перед початком P), час початку виділеного вікна може бути відкладено, і коротший інтервал часу відокремить слід. час початку вікна від часу приходу P-хвилі (тобто <15 с).

На рисунках 3.3 та 3.4 зображені розподіли вибірок часу приходу фаз P- і S-хвиль відповідно.

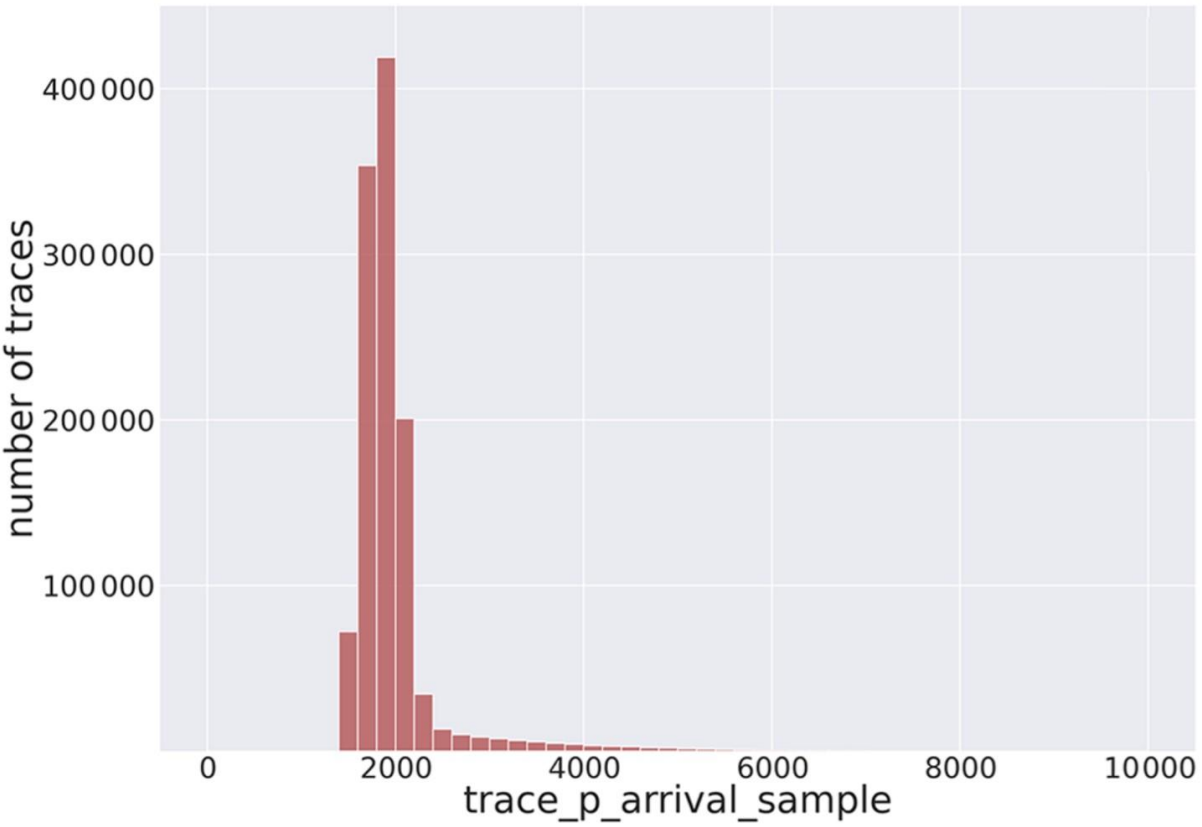

Рисунок 3.3 — Розподіл вибірки часу приходу фаз P-хвиль

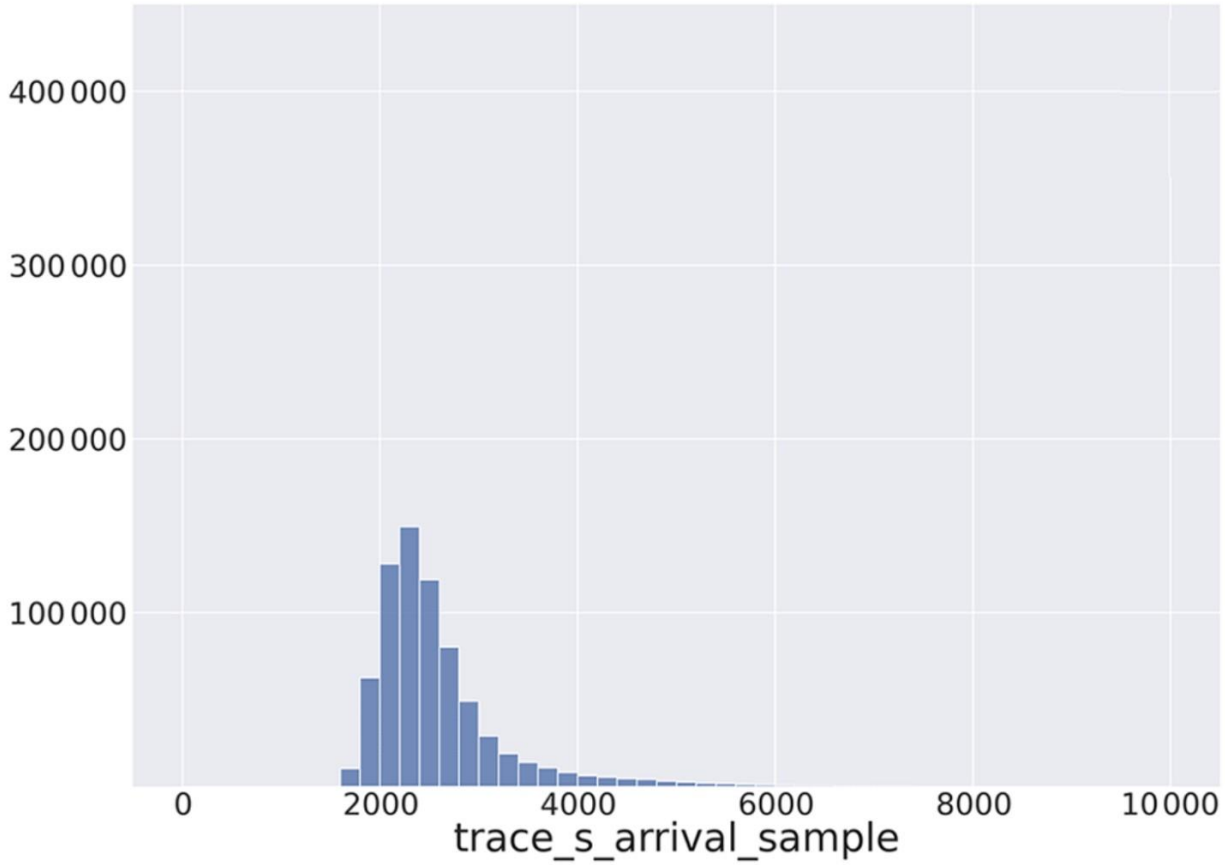

Рисунок 3.3 — Розподіл вибірки часу приходу фаз S-хвиль

Дані (формат miniSEED) були завантажені за допомогою веб-сервісів FDSN dataselect, наданих INGV [36]. Використовуючи набір із 14 процедур запитів на основі контейнерів, що виконуються паралельно, на цьому етапі знадобилося близько 7 днів для завершення завантаження ~4 мільйонів слідів форми сигналу (тобто ~1,3 мільйона слідів 3C) з потребою в пам'яті ~80 GB (стиснення miniSEED STEIM1).

3.1.1.4 Перехресна перевірка між метаданими на основі фаз і завантаженими даними форми сигналу

Після того, як масове завантаження даних було завершено, був створений список усіх завантажених файлів. Цей список перетинався з початково вибраними метаданими (підпункт 3.1.1.2), щоб мати відповідність один до одного між даними miniSEED і метаданими (тобто кожен запис сигналу 3C – три файли miniSEED – повинен відповідати рядку файл метаданих).

3.1.1.5 Підготовка оброблених сигналів у цифрових одиницях

Ця частина процедури збирання даних спрямована на підготовку сигналів цифрових відліків. Він включає в себе наступні кроки:

- видалення слідів, що містять прогалини в даних (тобто відсутні дані);
- обрізка траси сигналу до найближчого зразка до часу початку;
- вікно трасування 120 s;
- видалення середніх і лінійних тенденцій з даних;
- повторна дискретизація на 100 Гц;
- розрахунок відношення сигнал/шум;
- вилучення показників якості даних.

Обертання горизонтальної складової вздовж напрямків N–S та E–W не було необхідним, оскільки всі використані датчики орієнтовані відповідно. Для кожного сліду форми сигналу (тобто кожного компонента) максимальне значення відношення сигнал/шум (SNR) було виділено та збережено як метадані. SNR було розраховано як показано у формулі (3.1).

$$\mathrm{SNR} = 20 \log_{10} \frac{|S_{95}|}{|N_{95}|},$$

(3.1)

де |S95| та |N95| – це 95-й процентиль абсолютних значень даних у вікні 5 s відразу після початку зубця S і безпосередньо перед часом прибуття зубця P.

Якщо початок S-хвилі був недоступний, вікно S-хвилі визначали після обчислення прогнозованого приходу S-хвилі з використанням середньої швидкості 3,0 км с−1 та гіпоцентральної відстані.

Під час цього етапу підготовки даних також розраховуємо деякі параметри якості, витягнуті зі слідів форми сигналу, з метою подальшого включення в інформацію метаданих. Ці додаткові параметри, що забезпечують розподіл значень слідів, були обчислені за допомогою класу MSEEDMetadata програмного забезпечення ObsPy python [37] [38] [39]. З цією ж метою було визначено кількість піків за допомогою фільтра Hampel на ковзному вікні з 161 вибіркою, щоб знайти викиди в слідах.

Остаточний набір даних складається із загалом 1 159 249 записів сигналів ЗС від 54 008 землетрусів у одиницях підрахунку, зібраних у файлі формату HDF5. У таблиці 3.1 наведено кількість слідів у кожному інтервалі величини остаточного зібраного набору даних.

3.1.1.6 Застосування передавальної функції приладу до сигналів

Щоб зробити набір даних більш загальним, було створено також набір даних у одиницях фізичного руху землі після розгортання реакції приладу. З цією метою було завантажено файли відповідей станцій для всіх використовуваних станцій і застосовано функції передачі до окремих трас з

кутами частотної фільтрації 0,01, 0,04, 25 і 40 Гц за допомогою косинусного фланку частотного звуження і застосування 5 % косинуса звуження на обох кінцях сигналу трасування.

Після видалення реакції приладу було виділено показники інтенсивності:

- IMs, тобто пікове прискорення землі, PGA;
- пікову швидкість землі, PGV;
- спектральні прискорення за період 0,3, 1,0 і 3,0 с для кожного компонента, щоб вони могли включатися серед параметрів метаданих.

Пікові переміщення ґрунту не враховуються, оскільки вони отримані від одноразового або подвійного інтегрування записів швидкості та прискорення відповідно, і їх визначення може бути неточним, якщо виконується автоматично.

3.1.3 Опис метаданих

115 метаданих, пов’язаних з кожною трасою 3С сигналу колекції, наведено в таблиці 3.2. Одиниці вимірювання вказані в дужках у стовпці «Опис». Тільки підмножина метаданих може бути пов’язана зі слідами шуму (зірочка в стовпці «Шум»).

Метадані надають різну інформацію, яку можна розділити на чотири основні типи – метадані джерела, станції, трасування та шляху. Одиниця кожної метаданих надається в її найменуванні.

Таблиця 3.2 — СПИСОК МЕТАДАНИХ ДЛЯ ПОДІЙ І СИГНАЛІВ ШУМУ

Metadata parameter name	Noise	Description
source_id	⋆	Earthquake and noise ID (INGV and UTC time, respectively)
source_origin_time		Location preferred origin time (YYYY-MM-DDTHH:MM:SS.SSZ)
source_latitude_deg		Location preferred latitude (°)
source_longitude_deg		Location preferred longitude (°)
source_depth_km		Location preferred depth (km)
source_origin_uncertainty_s		Location preferred origin time uncertainty (s)
source_latitude_uncertainty_deg		Location preferred latitude uncertainty (°)
source_longitude_uncertainty_deg		Location preferred longitude uncertainty (°)
source_depth_uncertainty_km		Location preferred depth uncertainty (km)
source_stderror_s		Preferred earthquake location standard deviation (s)
source_gap_deg		Location preferred location gap (°)
source_horizontal_uncertainty_km		Location preferred horizontal uncertainty (km)
source_magnitude		Preferred magnitude
source_magnitude_type		Preferred magnitude type
source_mt_eval_mode		Moment tensor evaluation mode (e.g., manual)
source_mt_status		Status of the evaluation (“reviewed” or “final”)
source_mt_scalar_moment_Nm		Scalar moment (N m)
source_mechanism_strike_dip_rake		Strike, dip, rake of the two planes (two tuples)
source_mechanism_moment_tensor		Six components of the moment tensor (m_rr, m_tt, m_pp, m_rt, m_rp, m_tp)
source_type		Earthquake or other sources (quarry_blast, controlled explosion, experimental explosion, etc.)
station_network_code	⋆	Two characters FDSN network code (e.g., IV)
station_code	⋆	Station name (International Registry of Seismograph Stations, IR)
station_location_code	⋆	Location name identifier (Buland, 2006)
station_channels	⋆	Two characters identifying the sampling and the instrument gain (HN, HH, EH, etc.)
station_latitude_deg	⋆	Station latitude (°)
station_longitude_deg	⋆	Station longitude (°)
station_elevation_m	⋆	Station elevation (m)
station_vs30_mps	⋆	$V_{S,30}$ ($m\,s^{-1}$)
station_vs30_detail	⋆	$V_{S,30}$ information
path_ep_distance_km		Epicentral distance
path_hyp_distance_km		Hypocentral distance
path_azimuth_deg		Direction from event location to station (°)
path_backazimuth_deg		Direction from station location to event epicenter (°)
path_residual_[P,S]_s		P- or S-arrival time residual between picked arrival time and traveltime using preferred location (s)
path_weight_phase_location_[P,S]		P- or S-phase location weight resulting from preferred location (range 0–100)
path_travel_time_[P,S]_s		P- or S-wave traveltime (s)

Metadata parameter name	Noise	Description
trace_name	⋆	Waveform name within the HDF5 file
trace_start_time	⋆	Waveform trace UTC start time (YYYY-MM-DDTHH:MM:SS.SSZ)
trace_dt_s	⋆	Sampling interval (s)
trace_npts	⋆	Number of samples in waveform trace (integer)
trace_[P,S]_uncertainty_s		Assigned P- or S-onset arrival time uncertainty (s)
trace_eval_[P,S]		P- or S-type of picking (currently only “manual”)
trace_[P,S]_arrival_time		P- or S-arrival UTC start time (YYYY-MM-DDTHH:MM:SS.SSZ)
trace_polarity		P onset polarity (“negative”, “positive”, “undecidable”)
trace_[P,S]_arrival_sample		P- and S-onset sample number on waveform trace (integer)
trace_[E,N,Z]_median_counts	⋆	*E*-, *N*-, or Z-component sample median (counts, integer)
trace_[E,N,Z]_mean_counts	⋆	*E*-, *N*-, or Z-component sample mean (counts, integer)
trace_[E,N,Z]_min_counts	⋆	*E*-, *N*-, or Z-component sample minimum (counts, integer)
trace_[E,N,Z]_max_counts	⋆	*E*-, *N*-, or Z-component sample maximum (counts, integer)
trace_[E,N,Z]_rms_counts	⋆	*E*-, *N*-, or Z-component sample root mean squared
trace_[E,N,Z]_lower_quartile_counts	⋆	*E*-, *N*-, or Z-component sample lower quartile (counts, integer)
trace_[E,N,Z]_upper_quartile_counts	⋆	*E*-, *N*-, or Z-component sample upper quartile (counts, integer)
trace_[E,N,Z]_snr_db		*E*-, *N*-, or Z-component signal-to-noise ratio
trace_[E,N,Z]_spikes	⋆	*E*-, *N*-, or Z-component number of spikes (integer)
trace_GPD_[P,S]_number	⋆	P and S number of picks retrieved with GPD
trace_EQT_[P,S]_number	⋆	P and S number of picks retrieved with EQT
trace_EQT_number_detections	⋆	Number of detections retrieved with EQT
trace_[E,N,Z]_pga_cmps2		*E*-, *N*-, or Z-component PGA ($cm\,s^{-2}$)
trace_[E,N,Z]_pgv_cmps		*E*-, *N*-, or Z-component PGV ($cm\,s^{-1}$)
trace_[E,N,Z]_pga_perc		*E*-, *N*-, or Z-component PGA (% *g*)
trace_[E,N,Z]_pga_time		*E*-, *N*-, or Z-component PGA UTC time (YYYY-MM-DDTHH:MM:SS.SSZ)
trace_[E,N,Z]_pgv_time		*E*-, *N*-, or Z-component PGV UTC time (YYYY-MM-DDTHH:MM:SS.SSZ)
trace_[E,N,Z]_sa03_cmps2		*E*-, *N*-, or Z-component spectral acceleration at $t = 0.3$ ($cm\,s^{-2}$)
trace_[E,N,Z]_sa10_cmps2		*E*-, *N*-, or Z-component spectral acceleration at $t = 1.0$ ($cm\,s^{-2}$)
trace_[E,N,Z]_sa30_cmps2		*E*-, *N*-, or Z-component spectral acceleration at $t = 3.0$ ($cm\,s^{-2}$)
trace_pga_cmps2		Max. horizontal components PGA value ($cm\,s^{-2}$)
trace_pgv_cmps		Max. horizontal components PGV value ($cm\,s^{-2}$)
trace_pga_perc		Max. horizontal components PGA value (% *g*)
trace_sa03_cmps2		Max. horizontal components spectral acceleration ($t = 0.3$) ($cm\,s^{-2}$)
trace_sa10_cmps2		Max. horizontal components spectral acceleration ($t = 1.0$) ($cm\,s^{-2}$)
trace_sa30_cmps2		Max. horizontal components spectral acceleration ($t = 3.0$) ($cm\,s^{-2}$)
trace_deconvolved_units		Ground motion units of the traces in the HDF5 volume (e.g., mps and mps2 for $m\,s^{-1}$ and $m\,s^{-2}$, respectively)

Метадані джерела надають інформацію про землетрус з описом часу виникнення джерела; Розташування; розмір; і, якщо є, фокальний механізм, тензор моменту та кінцеву несправність.

Метадані станції надають інформацію про характеристики станції запису, які включають:

- станцію, канал, мережу та розташування (SCNL) [40];
- географічні координати;
- середня швидкість зсувної хвилі верхньої 30 м Землі, VS,30, що є важливим параметром для класифікації об'єктів у сейсмічних інженерних додатках [41] і витягується з карти, яка використовується в реалізації INGV програмного забезпечення USGS ShakeMap в Італії [42].

Метадані трасування складаються з параметрів, які витягуються з осциллограм, таких як максимальні та мінімальні амплітуди, середньоквадратичні значення трас і, після застосування функції передачі, вимірювання інтенсивності (IMs) руху грунту. До цього класу метаданих включено зубець P (і S), наданий бюлетенем INGV, і, крім того, кількість вибірок P і S, отриманих шляхом обробки сигналів з двома методами глибокого навчання, фазового вибору та виявлення подій (алгоритми GPD і EQTransformer; [43]; [11]), щоб звернути увагу користувача на те, що сигнал форми, що використовується, може включати більше одного землетрусу.

Метадані шляху випливають із обчислення параметрів, які пов'язують типи метаданих вище (наприклад, час подорожі, гіпоцентральна та епіцентральна відстані).

Обгрунтування вибору метаданих відображає намір надати користувачам вичерпну інформацію про дані. Здається, це важлива проблема, оскільки дані, які записуються автоматично, можуть страждати від багатьох

різноманітних проблем, які виникають через несправність реєстраторів даних і датчиків або через погану передачу даних. Оскільки мета зібрати набір даних, який також можна використовувати для аналізу потоків даних у реальному часі за допомогою машинного навчання, автоматична обробка, узагальнена вище, суттєво не відрізняється від тієї, яка зазвичай застосовується до потокових даних.

Однією з альтернатив до комплексного підходу метаданих було б «очищення» набору даних шляхом повного видалення несправних слідів із набору даних. Але тмиакий спосіб не підходить, оскільки в цьому випадку набір даних не буде репрезентативним для «справжніх» даних, які збираються в реальному часі мережами моніторингу. Таким чином, основна ідея критерію полягає в тому, щоб користувачі могли робити власний вибір, використовуючи відповідні фільтри для використання даних у своїх цілях.

Наприклад, якщо користувач шукає найчистіші дані, цього можна досягти шляхом відповідної фільтрації метаданих (наприклад, насичені швидкісні дані, отримані широкосмуговими датчиками, оснащеними 24 розрядними реєстраторами даних, можуть бути видалені консервативним способом, просто вибравши лише ці сліди з відліками в межах ±0,8×223). На відміну від цього, користувач може також залишити модель машинного навчання, щоб дізнатися про проблеми з даними, щоб їх можна було виявити під час використання реальних даних. Такий підхід був використаний [44] за відсутні дані. У [44] набір даних, який використовується для машинного навчання, складається з фіксованої кількості станцій, і коли дані з однієї або кількох станцій відсутні (або вся траса, або її частини), траса сигналу встановлюється як масив нулів. Було виявлено, що застосована там модель машинного навчання виявляла та вивчала проблемні значення та компенсувала їх, маючи подібну точність передбачення на цих станціях, як точність на станціях, які мали доступні вхідні дані. На практиці надання широкого набору

описових метаданих форми сигналу важливо не тільки для використання розширеного набору міток, які можна використовувати для різних цілей, але також для виявлення проблем із даними сигналу та включення або відфільтрування їх.

Метадані включають початок P- і S-хвиль, які вручну вибирають аналітики INGV, як зазначено в бюлетені INGV. Нагадаємо, що траси були обрані таким чином, щоб включати лише один час приходу P-хвилі і, можливо, один час прибуття S-хвилі, оскільки передбачалося зібрати один землетрус на трасу вікна. Цей критерій був обраний з метою полегшення навчання моделей машинного навчання із використанням трас, що містять лише один землетрус (для фазового вибору, пікових рухів ґрунту тощо). Однак, незважаючи на те, що було докладено значних зусиль, щоб ізолювати лише один землетрус за часове вікно, протягом одного і того ж періоду часу може бути присутнім більше одного (наприклад, аналітик не бачив або просто проігнорував інші події з меншою амплітудою).

Оскільки наявність додаткових, неідентифікованих землетрусів ускладнює етап навчання машинного навчання, було дотримано того ж підходу, який використовували [11], щоб запустити алгоритми автоматичного вибору на наборі даних форми хвилі та включити як метадані також кількість фаз P- і S-хвиль, які автоматично вибираються за допомогою узагальненого визначення фази, GPD, методики, запропонованої [43] та техніка EQTransformer від [45]. В аналізі було використовано як поріг виявлення 0,99 для P- і S-фази для GPD і 0,2, 0,1 і 0,1 для землетрусів, виявлення P-фази та виявлення S-фази, відповідно, для EQTransformer. I GPD, і EQTransformer були запущені лише на каналах з високим коефіцієнтом посилення (тобто HH, EH).

Як зазначено вище, метадані є важливою складовою колекції даних. Їх можна використовувати для ідентифікації даних, які підлягають аналізу, або як

мітки в програмах машинного навчання. На додаток до того факту, що не вся інформація метаданих в INSTANCE завжди доступна (наприклад, тензори моментів, як правило, доступні лише для подій з магнітудою ~M≥3,5 або вибір початку хвилі S, отриманий з бюлетеня INGV, може бути відсутнім), було виявлено, що автоматично оброблені дані трасування руху землі можуть страждати від помилок, оскільки вихідні траси містили вже невиявлені проблеми з несправністю (наприклад, стрибки, аномальні тенденції), які після застосування функції передачі приладу відображаються в помилкові траси руху землі. та значення IM. Аналогічно, могло статися, що в окремих випадках коефіцієнти передавальних функцій приладу були неправильними, створюючи також у цьому випадку неправильні сліди та значення IM. Щоб вирішити ці проблеми, було застосовано 2 способи. По-перше, було вирішено виявити максимальні та мінімальні значення слідів, що виходять за межі допустимого фізичного діапазону, і замінити їх на NumPy nan у файлі метаданих. Цей прийнятний діапазон був заснований на IM, зазначених у «плоскому» файлі ESM DB [46] [47], який включає всі IM (отримані в результаті аналітичної обробки) усіх доступних записів землетрусів з M≥4,0 в Європі. По-друге, була перевірена процедура обробки функції передачі інструментів шляхом перехресної перевірки всіх значень IM із тими, які вказано у плоскому файлі ESM DB. У зв’язку з цим було виявлено дуже хорошу відповідність між IM, отриманими за допомогою двох методологій, що дає впевненість у якості застосовуваної обробки даних та наданих метаданих IM.

3.1.4 Опис набору даних

На рисунку 3.1 показано землетруси, включені в набір даних. Розмір символу пропорційний магнітуді землетрусу. Обрані землетруси 54 008, що складають набір даних, можна вважати репрезентативною підмножиною всієї сейсмічності в Італії, а для більших подій також для тих землетрусів, що відбуваються поблизу національних кордонів.

На рис. 3.4 зображено 527 тензорів моментів, включених до метаданих. Розмір символу тензора моменту пропорційний source_magnitude, тоді як кольори визначаються відповідно до поширеного режиму деформації: індиго, лаванда та темно-оранжевий для сдвигу, нормального та насувного розломів відповідно. Поширений режим деформації визначається відповідно до викривлення розлому, отриманого з source_mechanisms_strike_dip_rake.

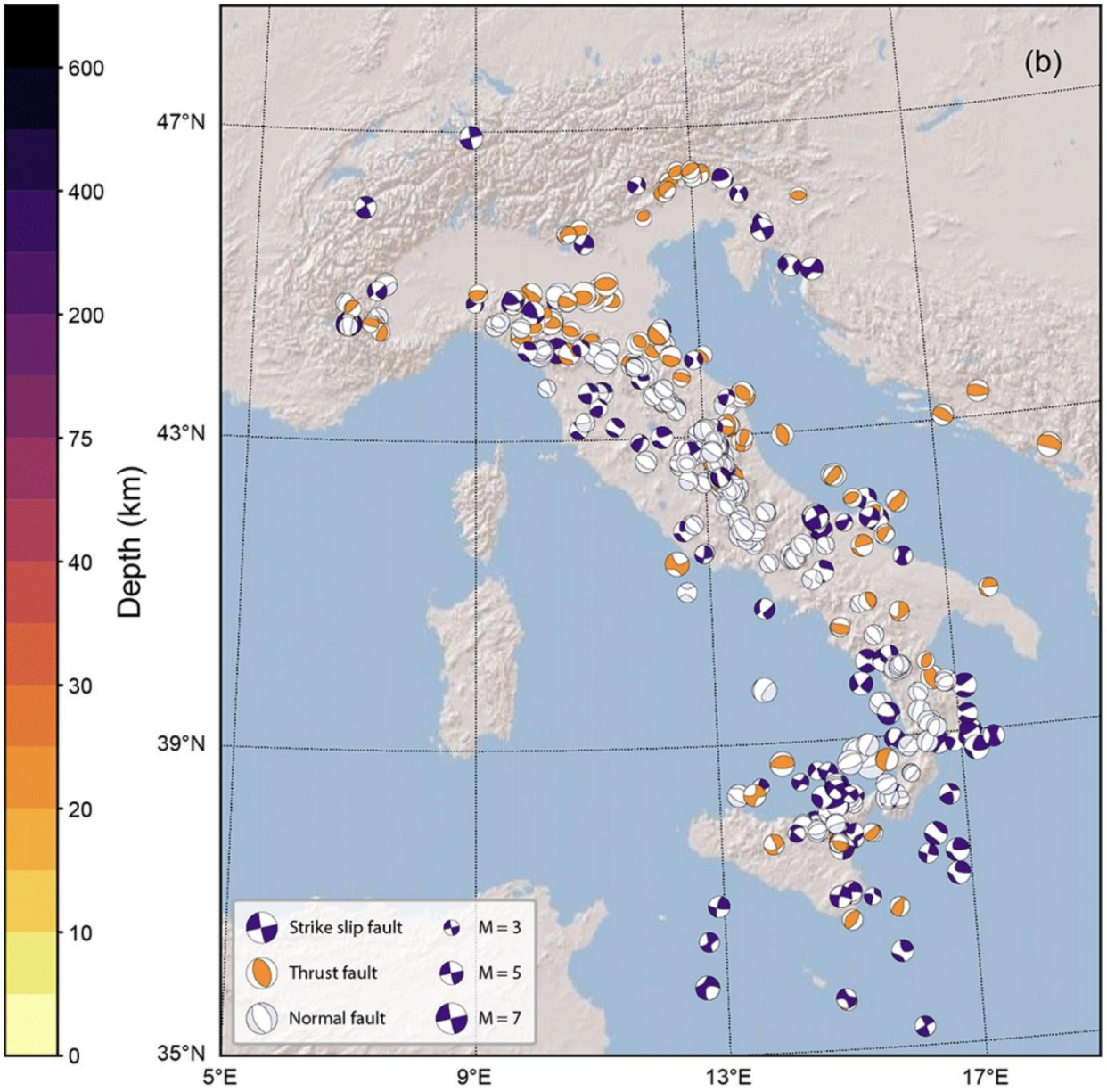

Рисунок 3.4 — Карта доступних тензорів моментів з кольорами, призначеними залежно від фокального механізму. Розмір символу на пропорційний магнітуді землетрусу.

На рис. 3.5 та 3.6 показані карти станцій, включених у набори даних про події та шум відповідно. Розмір символу на рис. 3.5 пропорційний кількості повідомлених фазових надходжень кожною станцією, тоді як на панелі 3.6 він пропорційний кількості сигналів, включених у набір даних для кожної станції.

На рис. 3.5 показано, що станції, включені в набір даних подій, повідомляють про різну кількість фаз.

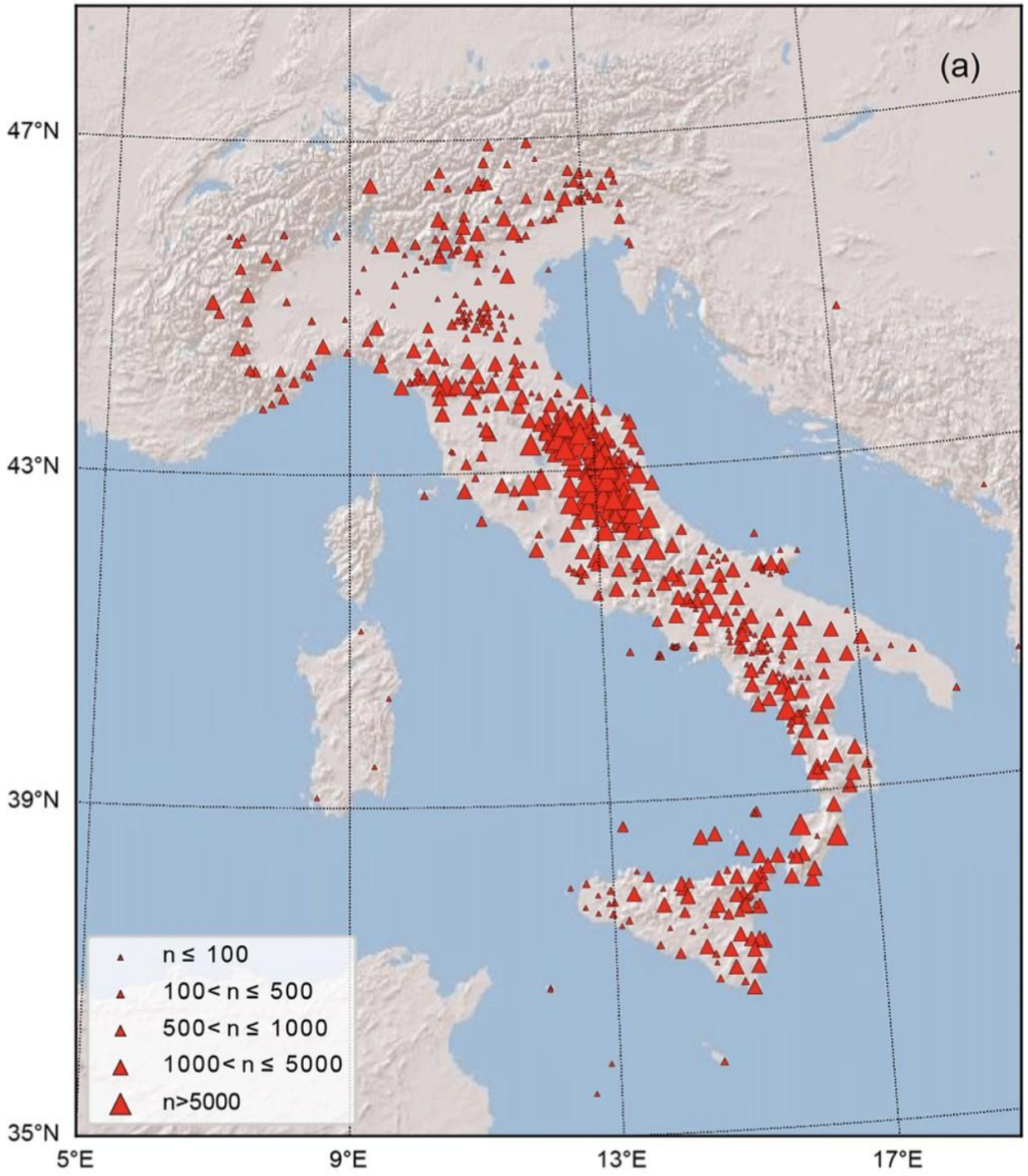

Рисунок 3.5 — Карта станцій, які використовуються для збирання наборів даних про події

Ці відмінності залежать від кількох факторів, наприклад від того, чи є станції постійними або тимчасовими, тривалості зйомки, рівня шуму та рівня сейсмічності району, де були розгорнуті станції. Наприклад, очевидно, що багато станцій у центральній Італії відображають багато фаз (і пов'язані з ними записи слідів) головним чином тому, що ця область була вражена послідовностями землетрусів 2009 та 2016 років. Навпаки, станції, які розташовані на рівнині По, зазвичай мають невелику кількість фаз, головним чином через високий рівень шуму, що ускладнює вибір фази. Така сама диверсифікація в кількості доступних слідів не спостерігається для набору даних шуму, показаного на рис. 3.6.

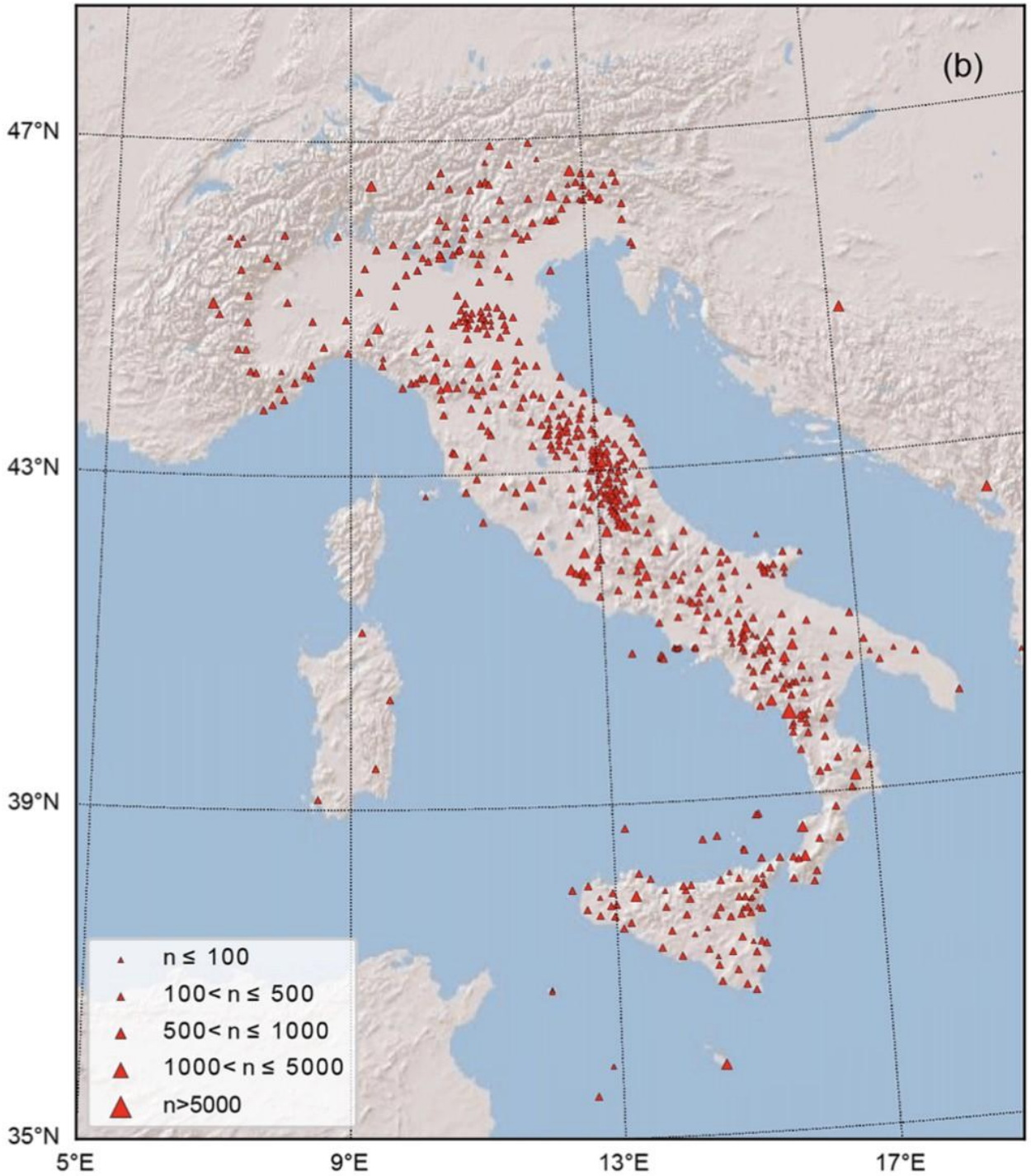

Рисунок 3.6 — Карта станцій, які використовуються для збирання наборів даних про шум

Це відбувається тому, що це був навмисний вибір, щоб вибрати більш-менш парну кількість трас для всіх каналів станції.

3.1.5 Імпорт даних

В датасеті можна знайти файли train.csv та метадані для них.

Segment_id: ідентифікаційний код сегмента даних. Відповідає імені пов’язаного файлу даних.

Time_to_eruption: цільове значення, час до наступного виверження.

[train|test]/*.csv: файли даних. Кожен файл містить десять хвилин журналів від десяти різних датчиків, розташованих навколо вулкана.

Показання були нормалізовані в межах кожного сегмента, частково для того, щоб показання потрапляли в діапазон значень int16.

Перш за все, імпортуємо необхідні бібліотеки. Рис. 3.7 демонструє імпорт необхідних для виконання аналізу бібліотек, таких як numpy, pandas, tqdm, sklearn та catboost.

```
import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as plt
import os
from tqdm import tqdm
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from catboost import CatBoostRegressor, Pool
from sklearn.metrics import mean_absolute_error, mean_squared_error
```

Рисунок 3.7 — Імпорт необхідних бібліотек

Імпортуємо дані для аналізу:

```
PATH = '../input/predict-volcanic-eruptions-ingv-oe/'

train_list = os.listdir('../input/predict-volcanic-eruptions-ingv-oe/train')
test_list = os.listdir("../input/predict-volcanic-eruptions-ingv-oe/test")
train_time = pd.read_csv(PATH + 'train.csv')
```

Рисунок 3.8 — Імпорт даних

Перевіримо, які масштаби має датасет:

```
print('Number of train files: {}'.format(len(train_list)))
print('Number of test files: {}'.format(len(test_list )))
```

```
Number of train files: 4431
Number of test files: 4520
```

Рисунок 3.9 — Розмір вхідних файлів

З рис. 3.9 можна побачити, що тека з тренувальним датасетом містить 4431 файл, а тека з тестовим — 4520 файлів.

Маємо 2 теки: трейн та тест, в кожній з яких по декілька csv файлів, які треба почергово передобробити.

Подивимося, як виглядають вхідні дані. Для цього зчитаємо нульовий файл з теки з тренувальними даними та присвоїмо йому змінну example:

```
example = pd.read_csv(PATH + 'train/' + train_list[0])
```

Рисунок 3.10 — Імпорт нульового файлу тренувального датасету

3.2 Визначення підходу до вирішення задачі

По суті метої цієї роботи стоїть вирішення задачі регрессії. Дослідимо нульовий файл із теки з тренувальними даними детальніше:

```
example[:5]
```

	sensor_1	sensor_2	sensor_3	sensor_4	sensor_5	sensor_6	sensor_7	sensor_8	sensor_9	sensor_10
0	-560.0	-508.0	NaN	-261.0	-348.0	1681.0	-764.0	-1193.0	NaN	-516.0
1	-508.0	-460.0	NaN	-276.0	-252.0	1934.0	-774.0	-1276.0	NaN	-537.0
2	-630.0	-260.0	NaN	-310.0	-174.0	2229.0	-785.0	-1298.0	NaN	-535.0
3	-587.0	1.0	NaN	-352.0	-69.0	2069.0	-788.0	-1249.0	NaN	-507.0
4	-778.0	240.0	NaN	-390.0	71.0	1850.0	-825.0	-1402.0	NaN	-437.0

Рисунок 3.11 – Дослідження нульового тренувального файлу

З рис. 3.11 можна побачити показники десяти сенсорів перших п’яти сутностей, серед яких є пропущені значення. Аналогічна ситуація з тестовим набором:

```
example_test = pd.read_csv(PATH + 'test/' + test_list[0])
```

```
example_test[:5]
```

	sensor_1	sensor_2	sensor_3	sensor_4	sensor_5	sensor_6	sensor_7	sensor_8	sensor_9	sensor_10
0	-511.0	NaN	-131.0	-457.0	47.0	-35.0	185.0	367.0	858.0	-492.0
1	-556.0	NaN	-105.0	-534.0	-7.0	-84.0	190.0	195.0	881.0	-368.0
2	-615.0	NaN	-97.0	-473.0	-50.0	8.0	219.0	327.0	937.0	-260.0
3	-682.0	NaN	-75.0	-388.0	-58.0	28.0	255.0	-249.0	995.0	-187.0
4	-763.0	NaN	-18.0	-358.0	-53.0	-104.0	271.0	-162.0	1032.0	-160.0

Рисунок 3.12 – Дослідження нульового тестового файлу

Також ми маємо набір даних з відповідними результатами до тренувального набору — цільовою інформацією про час до наступного виверження, яку ми будемо передбачати для тестового набору.

```
train_time
```

	segment_id	time_to_eruption
0	1136037770	12262005
1	1969647810	32739612
2	1895879680	14965999
3	2068207140	26469720
4	192955606	31072429
...	...	...
4426	873340274	15695097
4427	1297437712	35659379
4428	694853998	31206935
4429	1886987043	9598270
4430	1100632800	20128938

4431 rows × 2 columns

Рисунок 3.13 – Час до наступного виверження

Візуалізуємо перший файл:

```
example.plot(figsize=(15,15), subplots=True);
```

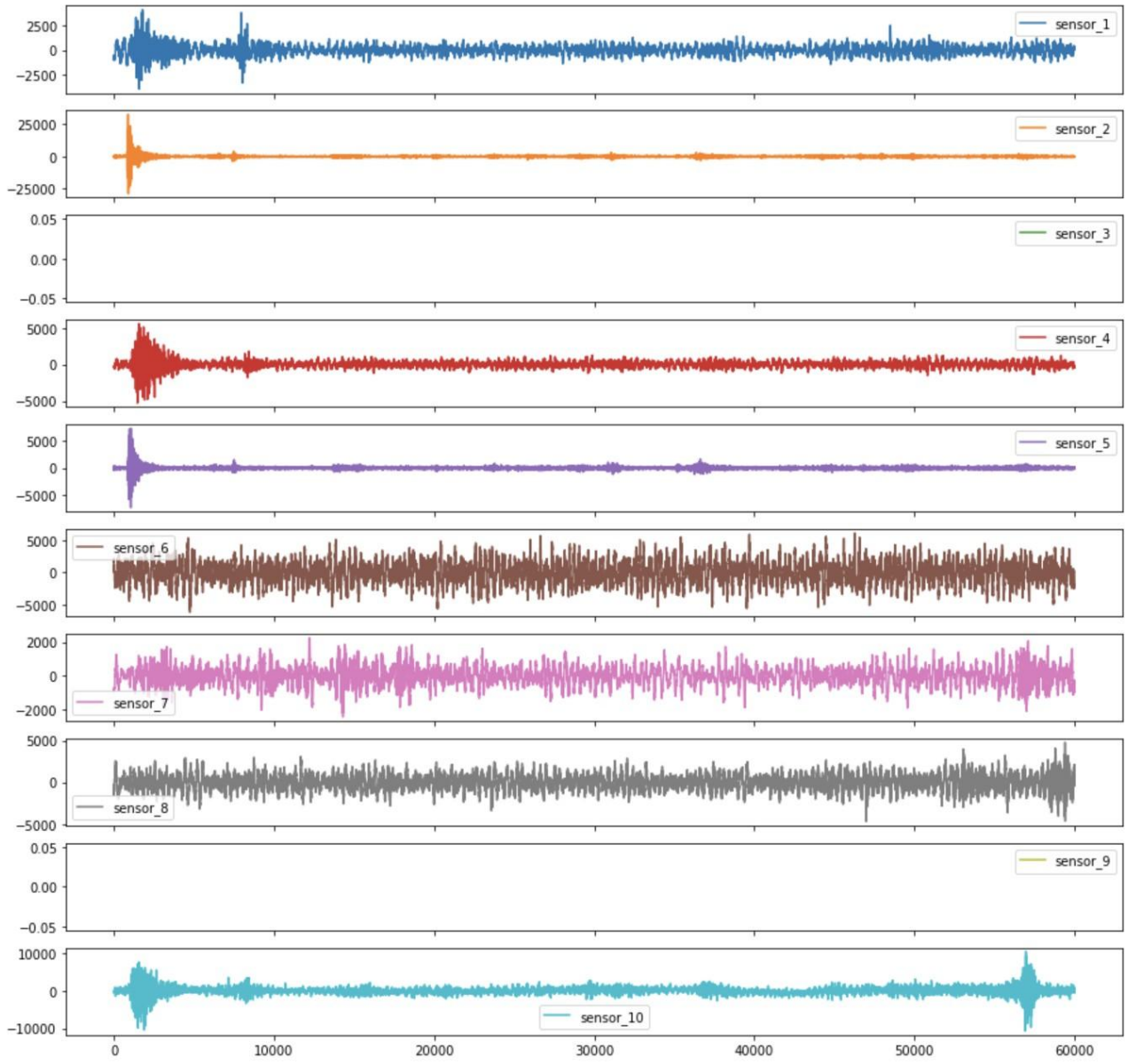

Рисунок 3.14 – Візуалізація сенсорів

З графіку можна побачити, як відрізняються показники різних сенсорів один від одного. Припустимо, що ми маємо деякий часовий ряд, який має свої характеристики: мінімуми, максимуми і тд. Наприклад, для порівняння: сенсори 4 і 5 мають по 1 піку, сенсор 10 має 2 піки, а сенсори 6, 7 і 8 відрізняються своєю високочастотністю, при цьому, деякі сенсори характеризуються наявністю дисперсії. Пропонуємий підхід до вирішення

наявної задачі: візьмемо набір визначених параметрів, і для кожного з сенсорів сформуємо значення за кожним з показників. Втілимо припущення на практиці. Для цього, перш за все, необхідно визначити точний час:

```
train_time[train_time.segment_id == int(train_list[0].split('.')[0])]
```

	segment_id	time_to_eruption
699	800654756	16818516

Рисунок 3.15 – Пропонуємий підхід до вирішення задачі

Далі візьмемо файл з різними сенсорними показниками і заповнимо нулями, припускаючи, що на цей момент там нічого немає, після чого виведемо горизонтальну матрицю з основними значеннями описових статистик. Для кращої наочності використаємо unstack(), щоб перевернути матрицю:

```
pd.DataFrame(example.fillna(0).describe().iloc[1:, :].unstack()).reset_index()
```

	level_0	level_1	0
0	sensor_1	mean	4.114265
1	sensor_1	std	470.544946
2	sensor_1	min	-3891.000000
3	sensor_1	25%	-273.000000
4	sensor_1	50%	0.000000
...	...	...	...
65	sensor_10	min	-10709.000000
66	sensor_10	25%	-473.000000
67	sensor_10	50%	0.000000
68	sensor_10	75%	487.000000
69	sensor_10	max	10389.000000

70 rows × 3 columns

Рисунок 3.16 – Створення описових статистик

Отриману таблицю запишемо у змінну process в більш наочному вигляді, присвоївши стовпцю зі значеннями сенсорів назву ‘value’ та створивши колонку з більш ємними індексами:

```
process = pd.DataFrame(example.fillna(0).describe().iloc[1:, :].unstack()).reset_index()
process = process.rename(columns={0: 'value'})
process['feature'] = process['level_0'] + '_' + process['level_1']
```

```
process
```

	level_0	level_1	value	feature
0	sensor_1	mean	4.114265	sensor_1_mean
1	sensor_1	std	470.544946	sensor_1_std
2	sensor_1	min	-3891.000000	sensor_1_min
3	sensor_1	25%	-273.000000	sensor_1_25%
4	sensor_1	50%	0.000000	sensor_1_50%
...	...	...	...	...
65	sensor_10	min	-10709.000000	sensor_10_min
66	sensor_10	25%	-473.000000	sensor_10_25%
67	sensor_10	50%	0.000000	sensor_10_50%
68	sensor_10	75%	487.000000	sensor_10_75%
69	sensor_10	max	10389.000000	sensor_10_max

70 rows × 4 columns

Рисунок 3.17 – Трансформація датасету 3.2.6

Позбавимося зайвих стовпців та перевернемо матрицю назад, використовуючи транспонування, щоб отримати однострокові дані, оскільки припускається, що кожне зі значень відповідатиме єдиниці часу (як з рис. 3.2.8):

```
process = process.drop(['level_0', 'level_1'], axis=1).set_index('feature').T
```

```
process
```

feature	sensor_1_mean	sensor_1_std	sensor_1_min	sensor_1_25%	sensor_1_50%	sensor_1_75%	sensor_1_
value	4.114265	470.544946	-3891.0	-273.0	0.0	286.0	40

1 rows × 70 columns

Рисунок 3.18 – Створення вектору для опису вулкану

На рис. 3.18 виведені описові статистики по кожному з сенсорів. Додамо до process колонку з часом, надавши їй назву time:

```
train_time[train_time.segment_id == int(train_list[0].split('.')[0])].time_to_eruption.values[0]
```

```
process
```

feature	sensor_1_mean	sensor_1_std	sensor_1_min	sensor_1_25%	sensor_1_50%	sensor_1_75%	sensor_1_max	sensor_2_mean
value	4.114265	470.544946	-3891.0	-273.0	0.0	286.0	4065.0	3.18048

1 rows × 71 columns

Рисунок 3.19 – Створення колонки з часом

3.3 Обробка даних

Таким чином, яким було визначено ознаки для всіх сенсорів одного з елементів вихідного датасету, необхідно обробити всі інші елементи. Для спрощення подальшого аналізу та економії часу створимо відповідну функцію:

```python
def create_frame(data, data_time=None, type_data='train'):
    data = data.fillna(0)

    # основні статистики
    data_transform = data.describe().iloc[1:, :]

    # додаткові параметри
    # коефіцієнт асиметрії
    data_transform.loc['skew'] = data.skew().tolist()

    #середнє абсолютне відхилення
    data_transform.loc['mad'] = data.mad().tolist()

    #коефіцієнт ексцесу - міра гостроти піку розподілу випадкової величини
    data_transform.loc['kurtosis'] = data.kurtosis().tolist()

    # додавання квантилів
    for i in range(0, 100, 5):
        if ((i!=25) & (i!=50)):
                str_col = f"{i}%"
                int_col = float(i)/100
                data_transform.loc[str_col] = data_transform.quantile(int_col).tolist()
        else:
            continue

    data_transform = pd.DataFrame(data_transform.unstack()).reset_index()
    data_transform = data_transform.rename(columns={0: 'value'})
    data_transform['feature'] = data_transform['level_0'] + '_' + data_transform['level_1']
    data_transform = data_transform.drop(['level_0', 'level_1'], axis=1).set_index('feature').T

    if type_data=='train':
        data_transform['time'] = data_time
    return data_transform
```

Рисунок 3.20 – Функція для обробки даних

Рис. 3.20 демонструє функцію, яка приймає тренувальний (за замовчуванням) або тестовий набір даних в якості вхідного параметру, визначає основні статистики, попередньо заповнивши поля нулями.

Далі отримуємо додаткові параметри, крім основних статистик: коефіцієнт асиметрії, середнє абсолютне відхилення, коефіцієнт ексцесу — міра гостроти піку розподілу випадкової величини, щоб якнайкраще описати даний числовий ряд.

Після цього додаємо квантилі, які корисні в якісному дослідженні ряду і дають розуміння, що відбувається на кожному з проміжків набору даних. Після цього зберемо всі пораховані значення у спільну таблицю, при чому, якщо на вхід буде поданий тренувальний датасет, до таблиці додається додатковий параметр часу (в іншому ж випадку вихідна таблиця цю колонку мати не буде).

Після визначення функції пройдемося за допомогою циклу по файлам вихідного тренувального набору даних, паралельно зберігаючи в кожному з них значення часу та обробляючи нещодавно створеною функцією create_frame, передаючи в неї необхідні параметри, та записуємо всі отримані значення в загальний масив all_train:

```
all_train = pd.DataFrame()

for file in tqdm(train_list):
    df = pd.read_csv(PATH + 'train/' + file)
    data_time = train_time[train_time.segment_id == int(file.split('.')[0])].time_to_eruption.values[0]
    df = create_frame(df, data_time, type_data='train')
    all_train = all_train.append(df)

all_train = all_train.reset_index(drop=True)
```

```
100%|██████████| 4431/4431 [20:03<00:00,  3.68it/s]
```

Рисунок 3.21 – Обробка тренувального набору

Подібним чином обробляємо тестовий набір:

```
all_test = pd.DataFrame()

for file in tqdm(test_list):
    df = pd.read_csv(PATH + 'test/' + file)
    df = create_frame(df, data_time=None, type_data='test')
    all_test = all_test.append(df)

all_test = all_test.reset_index(drop=True)
```

```
100%|██████████| 4520/4520 [19:59<00:00,  3.77it/s]
```

Рисунок 3.22 – Обробка тестового набору

І отримаємо такі результати для тренувального (рис. 3.23) та тестового (рис. 3.24) набору даних:

```
all_train[:5]
```

feature	sensor_1_mean	sensor_1_std	sensor_1_min	sensor_1_25%	sensor_1_50%	sensor_1_75%	sensor_1_max	sensor_1_skew	sensor_1_mad	se
0	4.114265	470.544946	-3891.0	-273.0	0.0	2.162283	4065.0	0.210301	348.802605	
1	-0.145998	584.061135	-2165.0	-382.0	0.0	0.003277	2610.0	0.006554	460.779468	
2	0.825053	262.683950	-1242.0	-163.0	0.0	0.449688	1435.0	0.074322	202.104232	
3	-1.490725	598.164998	-2310.0	-428.0	0.0	-0.095267	2370.0	-0.009542	483.291742	
4	-0.737071	399.197472	-2717.0	-213.0	0.0	0.175911	3812.0	0.351822	279.638276	

5 rows × 271 columns

Рисунок 3.23 – Результати після обробки тренувального набору

```
all_test[:5]
```

feature	sensor_1_mean	sensor_1_std	sensor_1_min	sensor_1_25%	sensor_1_50%	sensor_1_75%	sensor_1_max	sensor_1_skew	sensor_1_mad	se
0	2.934168	469.553705	-2035.0	-313.0	0.0	0.089316	1768.0	0.008376	371.762610	
1	-2.721021	324.254991	-2101.0	-207.0	0.0	-0.037527	1884.0	-0.075053	250.818417	
2	-1.021166	187.078093	-830.0	-122.0	0.0	-0.024409	811.0	-0.048817	147.487040	
3	0.394010	220.053325	-1078.0	-146.0	0.0	0.105554	904.0	0.022154	174.231944	
4	-3.621906	257.255088	-2004.0	-146.0	0.0	-0.097519	1681.0	-0.195039	189.118104	

5 rows × 270 columns

Рисунок 3.24 – Результати після обробки тестового набору

В цій роботі реалізується один з підходів до вирішення поставленої задачі, але, в теорії, додавши більшу кількість параметрів, точність роботи алгоритма збільшиться.

3.4 Моделювання математичної моделі

Першим кроком розіб’ємо дані на train, test та validation:

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.25, shuffle=True, random_state=10)

X_train, X_val, y_train, y_val = train_test_split(
    X_train, y_train, test_size=0.25, shuffle=True, random_state=10)
```

Рисунок 3.25 – Поділ даних на train, test та validation

Перш за все, перед побудовою найточніше можливої моделі, задано деякий базис — результат роботи моделі на стандартних параметрах, від якого будемо намагатися покращити результат.

Для цього визначимо метрику MAPE — нормовану сума помилок на тестове значення, подібну до метрики accuracy в задачах регресії, яку можна висловити як долю помилки у відсотках.

```
def mape(y_true, y_pred):
    return np.mean(np.abs((y_pred-y_true)/y_true))
```

Рисунок 3.26 – Визначення метрики MAPE

Використаємо Cat Boost Regressor, визначивши параметр функції витрат як MAPE:

```
clf = CatBoostRegressor(loss_function='MAPE')
train_dataset = Pool(data=X_train,
                     label=y_train,
                     )

eval_dataset = Pool(data=X_val,
                    label=y_val,
                    )

clf.fit(train_dataset,
          use_best_model=True,
          verbose = 0,
          eval_set=eval_dataset)
```

```
<catboost.core.CatBoostRegressor at 0x7f1f4fb2dbd0>
```

Рисунок 3.27 – Використання Cat Boost Regressor

Подивимося на результати роботи моделі:

```
y_pred = clf.predict(Pool(data=X_test))

print(f"MAPE: {mape(y_test, y_pred)}")
print(f"MAE: {mean_absolute_error(y_test, y_pred)}")
print(f"RMSE: {np.sqrt(mean_squared_error(y_test, y_pred))}")
```

```
MAPE: 0.8902223116424627
MAE: 16691321.883976629
RMSE: 20466697.72996543
```

Рисунок 3.28 – Проміжкові результати роботи моделі

С рис. 3.4.4 бачимо, що метрика MAPE дала результат близько 0.89, MAE приблизно 16601322, а RMSE майже 20466698.

3.5 KFold

Оберемо найкращі параметри моделі, використовуючи перехресну перевірку KFold з деякими параметрами. Для цього перемішуємо вибірку, підбираючи розмір відповідно до розміру тренувальних даних, визначаємо основні метрики, на основі яких порівнюватимемо результати, розбиваємо дані на KFolds та навчаємо модель, рахуючи значення метрик на кожному фолді та усереднюємо показники:

```
n_fold = 5
cv = KFold(n_splits=n_fold, shuffle=True, random_state=10)
prediction = np.zeros(len(test))
mape_, mae, rmse = [], [], []
params = {  'iterations':1000,
            'learning_rate':0.1,
            'depth':6,
            'eval_metric':'MAPE' }

for fold, (train_index, val_index) in enumerate(cv.split(X)):
    X_train = X.iloc[train_index,:]
    X_val = X.iloc[val_index,:]
    y_train = y.iloc[train_index]
    y_val = y.iloc[val_index]
    clf = CatBoostRegressor(**params)
    train_dataset = Pool(data=X_train,
                     label=y_train, )
    eval_dataset = Pool(data=X_val,
                    label=y_val, )
    clf.fit(train_dataset,
              use_best_model=True,
              verbose = 0,
              eval_set=eval_dataset)
    y_pred = clf.predict(Pool(data=X_test))

    mape_.append(mape(y_test, y_pred))
    mae.append(mean_absolute_error(y_test, y_pred))
    rmse.append(np.sqrt(mean_squared_error(y_test, y_pred)))
    print(f"fold: {fold}, MAPE: {mape(y_test, y_pred)}")
    print(f"fold: {fold}, MAE: {mean_absolute_error(y_test, y_pred)}")
    print(f"fold: {fold}, RMSE: {np.sqrt(mean_squared_error(y_test, y_pred))}")
    prediction += clf.predict(Pool(data=test))
    prediction /= n_fold

print('CV mean MAPE:  {0:.4f}, std: {1:.4f}.'.format(np.mean(mape_), np.std(mape_)))
print('CV mean MAE: {0:.4f}, std: {1:.4f}.'.format(np.mean(mae), np.std(mae)))
print('CV mean RMSE: {0:.4f}, std: {1:.4f}.'.format(np.mean(rmse), np.std(rmse)))
```

Рисунок 3.29 – Побудова моделі

```
fold: 0, MAPE: 0.6363779782532855
fold: 0, MAE: 3766619.120144325
fold: 0, RMSE: 5308180.127327653
fold: 1, MAPE: 0.44833988593256924
fold: 1, MAE: 1796125.994581451
fold: 1, RMSE: 2935063.2999856514
fold: 2, MAPE: 0.1151356312127801
fold: 2, MAE: 635795.8949968457
fold: 2, RMSE: 839119.0300593303
fold: 3, MAPE: 0.13120921266815777
fold: 3, MAE: 605853.062263884
fold: 3, RMSE: 806425.250295085
fold: 4, MAPE: 0.2541742260581759
fold: 4, MAE: 1137252.072717069
fold: 4, RMSE: 1494958.3100079847
CV mean MAPE:  0.3170, std: 0.1992.
CV mean MAE: 1588329.2289, std: 1171680.5125.
CV mean RMSE: 2276749.2035, std: 1700751.6538.
```

Рисунок 3.30 – Результати роботи оптимізованої моделі

З рис. 3.30 можна побачити остаточні метрики моделі після її оптимізації за допомогою KFold. Без оптимізації, базова модель показувала MAPE 0.9, в той час як в останньому варіанті MAPE дорівнює 0.3. Беручи до уваги, що для оптимізації моделі було використано дуже небагато ресурсів та навчання проводиться досі дуже швидко, це дуже гарні показники. Отже, можна зробити висновок про ефективне покращення моделі та кращі результати, які демонструють збільшення точності майже на 62%.

Повний текст роботи програми наведений у додатку А.

4 ПРАКТИЧНЕ ЗАСТОСУВАННЯ РОЗРОБЛЕНОЇ МОДЕЛІ

Розроблена модель має багато цілей використання, починаючи від покращення існуючих конфігурацій сейсмічного моніторингу до розробки та тестування нових методів виявлення землетрусів та оцінки руху грунту.

4.1 Загроза тихих землетрусів

Відсутність гулу не обов’язково робить землетрус нешкідливим. Деякі з тихих типів можуть передвіщати руйнівні цунамі або більші поштовхи, що стрясають землю.На початку листопада 2000 року на Великому острові Гаваї стався найбільший землетрус за більше ніж десятиліття. Близько 2000 кубічних кілометрів південного схилу вулкана Кілауеа хилилося до океану, вивільняючи енергію магнітудою 5,7. Частина цього руху відбулася під місцевістю, де тисячі людей щодня зупиняються, щоб побачити один з найбільш вражаючих потоків лави на острові. Але коли стався землетрус, ніхто не помітив — навіть сейсмологи. Як можна було не помітити таку знаменну подію?

Як виявилося, землетрус не є невід'ємною частиною всіх землетрусів. Подія на Кілауеа була одним із перших однозначних записів так званого тихого землетрусу, типу масивного руху землі, невідомого науці лише кілька років тому. Під час типового поштовху протилежні сторони розлому пролітають один одного за лічені секунди — досить швидко, щоб створити сейсмічні хвилі, які викликають гуркіт і тремтіння землі.

Але те, що землетрус відбувається повільно і тихо, не робить його незначним. Якби той самий великий масив каменів і уламків набрав обертів і набув форми гігантського зсуву, який відокремився від решти вулкана і швидко сповз у море, наслідки були б руйнівними. Матеріал, що руйнується, штовхне

морську воду у височенні хвилі цунамі, які можуть загрожувати прибережним містам уздовж усього Тихоокеанського краю. Такий катастрофічний зрив флангу, як це називають геологи, є потенційною загрозою навколо багатьох острівних вулканів у всьому світі.

4.1.1 Несподіване переміщення

На щастя, відкриття тихих землетрусів відкриває більше хороших новин, ніж поганих. Імовірність катастрофічного провалу флангу незначна, і прилади, які фіксують тихі землетруси, можуть зробити можливими раннє попередження. Нові докази умов, які можуть викликати тихе ковзання, пропонують сміливі стратегії для запобігання обвалу флангу. Повідомляються також про випадки тихих землетрусів у районах, де прорив флангу не є проблемою. Там тихі землетруси є надихаючими способами покращити прогнози їхніх побратимів, що потрясають землю.

Відкриття тихих землетрусів та їх зв'язок із катастрофічним обваленням флангу було побічним продуктом зусиль з вивчення інших потенційних природних небезпек. Руйнівні землетруси та вулкани викликають занепокоєння в Японії та на північному заході Тихого океану США, де тектонічні плити постійно занурюються глибоко в землю вздовж так званих зон субдукції. Починаючи з початку 1990-х років, геологи почали розгортати великі мережі безперервних приймачів глобальної системи позиціонування (GPS) у цих регіонах і вздовж схилів діючих вулканів, таких як Кілауеа. Отримуючи сигнали від сузір’я з більш ніж 30 навігаційних супутників, ці прилади можуть вимірювати своє власне положення на поверхні планети в будь-який момент часу з точністю до кількох міліметрів.

Вчені, які розгорнули ці GPS-приймачі, очікували побачити як повільний, невблаганний рух оболонки планети з тектонічних плит, так і

відносно швидкі рухи, які викликають землетруси та вулкани. Було сюрпризом, коли ці інструменти виявили невеликі рухи землі, які не були пов'язані з жодним відомим землетрусом чи виверженням. Коли дослідники наносили рух землі на карту, отриманий малюнок дуже нагадував одну характеристику руху розломів. Іншими словами, всі GPS-станції з одного боку даного розлому перемістилися на кілька сантиметрів в одному загальному напрямку. Цей шаблон не був би дивним, якби на формування знадобився рік чи більше. У цьому випадку вчені знали б, що відповідальним є повільний і постійний процес, який називається повзучістю відбою. Але зі швидкістю до сантиметрів на день таємничі події відбувалися в сотні разів швидше. Крім відносної швидкості, ці безшумні землетруси мають ще одну характеристику зі своїми шумними аналогами, що відрізняє їх від повзучості розломів: вони не є постійними процесами, а є дискретними подіями, які починаються і закінчуються раптово.

Цей раптовий початок, коли це відбувається на схилах вулканічного острова, викликає занепокоєння щодо можливої катастрофічної події на фланзі. Більшість типових землетрусів відбуваються вздовж розломів, які мають вбудовані гальма: рух припиняється, як тільки напруга знята між двома шматками землі, які намагаються пройти один повз одного. Але активність може не припинитися, якщо сила тяжіння стане основним рушієм. У найгіршому випадку ділянка вулкана, що лежить над розломом, стає настільки нестабільною, що коли починається зсув, сила тяжіння тягне весь схил гори вниз, поки він не розпадеться на купу сміття на дні океану.

Схили вулканів, таких як Кілауеа, стають крутими і вразливими до такого роду обвалів, коли лава від повторних вивержень нарощує їх швидше, ніж вони можуть розмиватися. Виявлення тихого землетрусу на Кілауеа свідчить про те, що південний фланг вулкана рухається — можливо, на шляху до остаточного знищення.

Наразі тертя по розлому діє як екстрене гальмо. Але гравітація перемогла в багатьох інших випадках у минулому. Вчені вже давно бачать докази стародавніх обвалів на гідролокаційних знімках гігантських полів сміття на мілководді, що оточує вулканічні острови по всьому світу, включаючи Майорку в Середземному морі та Канарські острови в Атлантичному океані. На Гавайських островах геологи знайшли понад 25 окремих обвалів, які відбулися за останні п’ять мільйонів років — миття ока в геологічному часі.

У типовому слайді обсяг матеріалу, який потрапляє в океан, у сотні разів перевищує ділянку гори Сент-Хеленс, що розірвався під час виверження 1980 року — більш ніж достатньо, щоб викликати величезні цунамі. На гавайському острові Ланаї, наприклад, геологи виявили докази дії хвиль, у тому числі рясні уламки морських раковин, на висоті 325 метрів. Гарі М. МакМертрі з Гавайського університету в Маноа та його колеги приходять до висновку, що найбільш вірогідний спосіб, яким снаряди могли досягти такого високого місця, був у межах хвиль цунамі, які досягли дивовижної висоти 300 метрів уздовж деяких гавайських берегів. Більшість найвищих хвиль, зареєстрованих у наш час, були не більше однієї десятої цього розміру.

4.1.2 Попередження катастрофи

Як би страшно не звучала така подія, цю небезпеку слід розуміти у належному контексті. Катастрофічний руйнування вулканічних схилів є дуже рідкісним явищем у людському масштабі, хоча набагато частіше, ніж ймовірність зіткнення великого астероїда чи комети із Землею. Десь на Гавайських островах колапси, достатньо великі, щоб спричинити цунамі, трапляються приблизно раз на 100 000 років. Деякі вчені підрахували, що такі події відбуваються у всьому світі раз на 10 000 років. Оскільки небезпека

надзвичайно руйнівна, коли вона трапляється, багато вчених погоджуються, що до неї варто підготуватися.

Щоб виявити деформацію в межах нестабільних вулканічних островів, мережі безперервних приймачів GPS починають розгортати на острові Реюньон в Індійському океані, на Фого на островах Зеленого Мису та на всьому архіпелазі Галапагоських, серед інших. Наприклад, мережа Кілауеа з понад 20 станцій GPS вже показала, що вулкан відчуває повзучі, тихі землетруси, а також великі, руйнівні типові землетруси. Однак деякі вчені припускають, що Кілауеа в даний час може бути захищена від катастрофічного обвалу кількома підводними купами бруду та каменів — імовірно, уламками від старих обвалів флангів, — які укріплюють його південний фланг. Нові відкриття про те, як Кілауеа сповзає, можна легко узагальнити на інші острівні вулкани, які можуть не мати подібних опорних структур.

Якими б не були конкретні обставини для острова, перехід від тихого ковзання до різкого колапсу включатиме раптове прискорення рухомого схилу. У гіршому випадку це прискорення миттєво досягне шаленої швидкості, не залишаючи шансів на раннє виявлення та попередження. У кращому випадку прискорення відбувалося б поривками, у каскаді тихих землетрусів, які повільно переростали б у звичайні землетруси, а потім до катастрофи. Безперервна мережа GPS могла легко виявити це поривчасте прискорення задовго до того, як почалися землетруси, і, якщо пощастило, за достатньо часу для корисного попередження про цунамі.

Однак, якби обвал був достатньо великим, попередження за кілька годин або навіть днів могло б не втішити, тому що в цей момент було б дуже важко евакуювати всіх. Ця проблема породжує питання про те, чи зможе влада коли-небудь запровадити превентивні заходи. Проблема стабілізації балансування флангів океанічних вулканів в принципі розв’язна. На практиці, однак, потрібні були б величезні зусилля. Розглянемо просту грубу силу. Якби

з верхньої течії нестабільного вулканічного флангу було вилучено достатньо гірських порід, то гравітаційна потенційна енергія, яка веде систему до колапсу, зникла б принаймні на кілька сотень тисяч років. Другий можливий метод — повільне опускання нестабільного флангу через серію невеликих землетрусів — був би набагато дешевшим, але загрожував геологічними невідомими та потенційними небезпеками. Для цього вчені могли б використовувати як інструмент для запобігання колапсу саме те, що зараз може викликати тихі землетруси на Кілауеа.

За дев’ять днів до останнього тихого землетрусу на Кілауеа проливний дощ скинув на вулкан майже метр води менш ніж за 36 годин. Геологам давно відомо, що вода, яка просочується в розломи, може спровокувати землетруси, і дев’ять днів – це приблизно той самий час, який, за їхніми підрахунками, потрібно воді, щоб пройти через тріщини та пори у тріщинах базальтової породи Кілауеа на глибину до п’яти кілометрів. де стався тихий землетрус. Тягар скелі, що лежить вище, створив тиск на дощову воду, роз’єднавши сторони розлому і полегшив їм прослизання один повз одного.

Це відкриття підтверджує суперечливу ідею примусового нагнітання води або пари в розломи біля основи нестабільного флангу, щоб викликати землетруси, що знімають стрес, необхідні для повільного його опускання. Цей вид спричиненого людиною ковзання відбувається в дуже малих масштабах постійно на геотермальних станціях та інших місцях, де вода закачується в землю.

Але коли справа доходить до вулканів, надзвичайна складність полягає в тому, щоб помістити потрібну кількість рідини в потрібне місце, щоб ненавмисно не спричинити саме обвал, якого слід уникнути. Деякі геофізики розглядали цю стратегію як спосіб зняти стрес уздовж сумнозвісного розлому Каліфорнії Сан-Андреас, але в кінцевому підсумку вони відмовилися від цієї ідеї, побоюючись, що це створить більше проблем, ніж вирішить.

Окрім звернення уваги до явища катастрофічного обвалення флангу вулкана, відкриття тихих землетрусів змушує вчених переглянути різні аспекти руху розломів, включаючи оцінку сейсмічної небезпеки. У північно-західній частині Тихого океану США дослідники спостерігали багато тихих землетрусів уздовж величезної зони розлому Каскадія між Північноамериканською плитою та плитою Хуан де Фука, що занурюється. Однією цікавою особливістю цих тихих землетрусів є те, що вони відбуваються через регулярні проміжки часу — насправді настільки регулярні, що вчені тепер успішно прогнозують їх виникнення.

Ця передбачуваність, швидше за все, пов’язана з тим фактом, що вода, яка тече з нижньої зони субдукції, може мати значний контроль над тим, коли і де ці розломи безшумно прослизають. Оскільки субдукційна плита занурюється глибше в землю, вона стикається з все вищими температурами і тиском, які вивільняють значну кількість води, що затримується в багатих водою мінералах, які існують у плиті. Тихі землетруси можуть мати місце, коли порція рідини з плити рухається вгору — коли рідина проходить, вона трохи розтискає зону розлому, можливо, дозволяючи повільне ковзання.

Більше того, Гаррі Роджерс і Герб Драгерт з Геологічної служби Канади повідомили в червні минулого року, що ці безшумні поштовхи можуть навіть служити попередниками деяких великих поштовхів у регіоні. Оскільки повільні ковзання відбуваються глибоко і через дискретні інтервали, вони регулюють швидкість, з якою напруга накопичується на більш мілкій частині зони розлому, яка рухається поривами і початками. У цьому неглибокому замкненому сегменті розлому зазвичай потрібні роки або навіть століття, щоб накопичити стрес, необхідний для початку сильного потрясіння. Однак Роджерс і Драгерт припускають, що тихе ковзання може різко прискорити наростання стресу, тим самим підвищуючи ризик регулярного землетрусу протягом тижнів і місяців після тихого землетрусу.

Безшумні землетруси змушують вчених переглянути сейсмічні прогнози і в інших частинах світу. Вважається, що регіони Японії поблизу кількох так званих сейсмічних розривів — районів, де регулярних землетрусів менше, ніж очікувалося, відбуваються в сейсмічно активному регіоні — запізнилися на руйнівний поштовх. Але якщо тихе ковзання знімало напругу на цих розломах, а вчені цього не усвідомлювали, то ступінь небезпеки насправді може бути меншим, ніж вони думають. Аналогічно, якщо безшумне ковзання буде виявлено вздовж розломів, які до цього часу вважалися неактивними, ці структури потребують ретельної оцінки, щоб визначити, чи вони також здатні до руйнівних землетрусів.

Якщо майбутнє дослідження покаже, що тихі землетруси є загальною ознакою більшості великих розломів, тоді вчені будуть змушені переглянути давні доктрини про всі землетруси. Спостереження за різними швидкостями ковзання розломів ставить справжню проблему для теоретиків, які намагаються пояснити процес розлому, наприклад, за допомогою фундаментальних фізичних законів. Зараз вважається, що кількість і розміри спостережуваних землетрусів можна пояснити за допомогою досить простого закону тертя. Але чи може цей закон також враховувати тихі землетруси? Поки що остаточної відповіді не знайдено, але дослідження тривають.

Тихі землетруси тільки починають входити в суспільний лексикон. Ці тонкі події передвіщають експоненціальне збільшення нашого розуміння того, як і чому відбувається збій. Важко переоцінити важливість розшифровки несправностей, оскільки коли помилки швидко усуваються, вони можуть завдати величезної шкоди, іноді на великій відстані від джерела. Існування тихих землетрусів дає вченим абсолютно новий погляд на процес ковзання, дозволяючи детально вивчати зони розломів на кожному етапі їх руху.

Основна мета моніторингу – дізнатися, коли в вулкані піднімається нова магма, яка може призвести до виверження.

Це надзвичайно важливо. Є очевидні небезпеки для мешканців поблизу. Крім безпеки людей, існують великі економічні проблеми. В цьому контексті більше йде мова про завчасне інформування, як у випадку з ураганами, ніж про можливість запобігання виверження.

Моніторинг пов’язаний із дослідженням прилеглої території, щоб побачити, куди поділися попередні потоки лави, і побачити, де відбувався попередній випадання попелу. Таким чином, ви отримаєте певне уявлення про історію вулкана та типи вивержень, які він зазвичай має. Кожен вулкан відрізняється, тому доводиться проводити індивідуальне дослідження та індивідуальний моніторинг.

У повітрі величезна небезпека через шлейфи виверження. Вулканічний попіл не схожий на попіл з каміна. В основному це подрібнені породи та частинки скла. Поміщати скло в реактивний двигун не добре. Тому моніторинг на Алясці надзвичайно важливий для авіаційної промисловості.

У США більшість вулканів них знаходиться на Алясці. Тільки цього літа було три виверження одночасно. Рідко трапляється, що вулкан на Алясці не вивергається. Гора Сент-Хеленс [у штаті Вашингтон] нещодавно припинила виверження, а Кілауеа на Гаваях тривають виверження.

4.1.3 За якими вулканами слід спостерігати найбільш уважно

У Йеллоустоні в минулому були величезні виверження, але це надзвичайно рідкісні події. Звичайно, з точки зору безпосереднього впливу людини, найбільше занепокоєння викликало б виверження ще одного з вулканів Каскадного хребта на західному узбережжі США. Найгіршим сценарієм було б, якби на горі Сент-Хеленс відбулося ще одне виверження такого розміру, як у 1980 році [в результаті якого загинули 57 людей і завдало збитків приблизно в 1,1 мільярда доларів] — або гора Реньєр поблизу Сіетла

або гора Худ поблизу Портленда, штат Орегон. Вони, швидше за все, вибухнуть раніше, ніж Єллоустоун. Чудово говорити про Аляску, але там небагато людей. Це велика авіаційна небезпека, але якщо на західному узбережжі вибухне вулкан, людський вплив буде набагато сильніше.

4.1.4 Очевидна користь моніторингу вулканів

Гора Сент-Хеленс була чудовим прикладом. Ідеальний приклад був не в США, а на Філіппінах з гори Пінатубо в 1991 році. Програма допомоги при катастрофі вулканів (VDAP) USGS відреагувала на це. Там, з бази ВМС США, чиновники VDAP зайшли при перших ознаках активності та встановили багато обладнання для моніторингу та провели швидке дослідження в надзвичайних ситуаціях.

Розроблена модель має неоціненний потенціал в попередженні стихійного лиха та збереженні багатьох життів. Багато людей висміюють фінансування програм спостереження за вулканами, але насправді вулкани потенційно смертельні, і за ними слід спостерігати, щоб попередити можливе виверження.

4.2 Необхідність моніторингу вулканів

Чи то буріння на природному газі, що вивільняє грязьовий вулкан, який поглинув 12 індонезійських сіл, чи виверження Кракатау в 1883 році, яке покрило світ достатньою кількістю частинок, щоб заблокувати сонячне світло та знизити температуру більш ніж на 1 градус Цельсія, вулкан є одними з найбільш руйнівних природних явищ Землі.

Ці отвори або вентиляційні простори в земній корі дозволяють вивергатися гарячому попелу, парі або навіть магмі. Потоки лави потім можуть

побудувати нові землі в океані — як у випадку Гаваїв — або поховати цілі міста, як у випадку Помпеї в 79 році нашої ери.

Вчені можуть передбачити, коли відбудеться таке виверження, вимірюючи ряд індикаторів, включаючи землетруси та викиди газу на вулкані. Але ці методи не є надійними, і все одно трапляються сюрпризи. Але в той же час просування гарячих порід глибоко під поверхнею може стати сильним і поновлюваним джерелом енергії.

Геологічне товариство відповідає за спостереження за вулканами та їх територіями. Проблема виверження вулканів не обмежується лише смертельно небезпечними вибухами лави голлівудського калібру. Інші загрози включають потенційно смертельні зсуви, падіння кам'янистого попелу та затоплення токсичними газами, які можуть бути викликані виверженнями вулканів.

Що таке моніторинг вулканів? Існує багато різних методів, але в основному це дослідження того, що саме відбувається на вулкані. Це може включати сейсмічну активність, невеликі землетруси, викиди газу, деформацію (вибухання вулкана або занурення) тощо.

Багато з вулканічних виверженъ відбулися в недалекому минулому і вивергатимуться знову в осяжному майбутньому. Зі збільшенням популяції розбудовуються території поблизу вулканів і збільшуються маршрути авіації. В результаті вулканічної активності під загрозою все більше людей та нерухомості.

Виверження вулканів є одним із найбільш драматичних і жорстоких факторів змін на землі. Мало того, що потужні вибухові виверження можуть різко змінити географічний стан території на десятки кілометрів навколо вулкана, але крихітні рідкі крапельки сірчаної кислоти, що вивергаються в стратосферу, можуть змінювати екологію нашої планети.

Вибухове виверження — енергетичне виверження, яке утворює переважно попіл, пемзу та осколкові балістичні уламки (на відміну від ефузійного виверження).

На рис. 4.1 можна побачити виверження гори Сент-Хеленс 22 липня 1980 року, підняло пемзу та попіл на висоту від 10 до 18 км у повітря і було видно в Сіетлі, штат Вашингтон (100 миль/160 км на північ).

Рисунок 4.1 — Виверження гори Сент-Хеленс. Дукас, Майк, 22 липня 1980 року

Зола — дрібні уламки (менше 2-4 мм в діаметрі) вулканічної породи, утворені в результаті вулканічного вибуху або викиду з вулканічного жерла. Золу можна побачити на рис. 4.2.

Рисунок 4.2 — Попіл від виверження гори Сент-Хеленс 18 травня 1980 року зібраний 39 км за вітром у Рендлі, штат Вашингтон, Вейпрехт, Д.

Вентиляційний отвір — будь-який отвір на поверхні Землі, через який вивергається магма або викидаються вулканічні гази.

Магма — розплавлена порода під поверхнею Землі. На рисунку можна побачити розташування утворення, накопичення та зберігання магми під горою Сент-Хеленс (розташування визначено з наукових даних).

Магма утворюється над субдукційною плитою океанічної кори і накопичується біля основи твердої земної кори, перш ніж збирається в зоні зберігання в 13 км (8 миль) під вулканом до виверження.

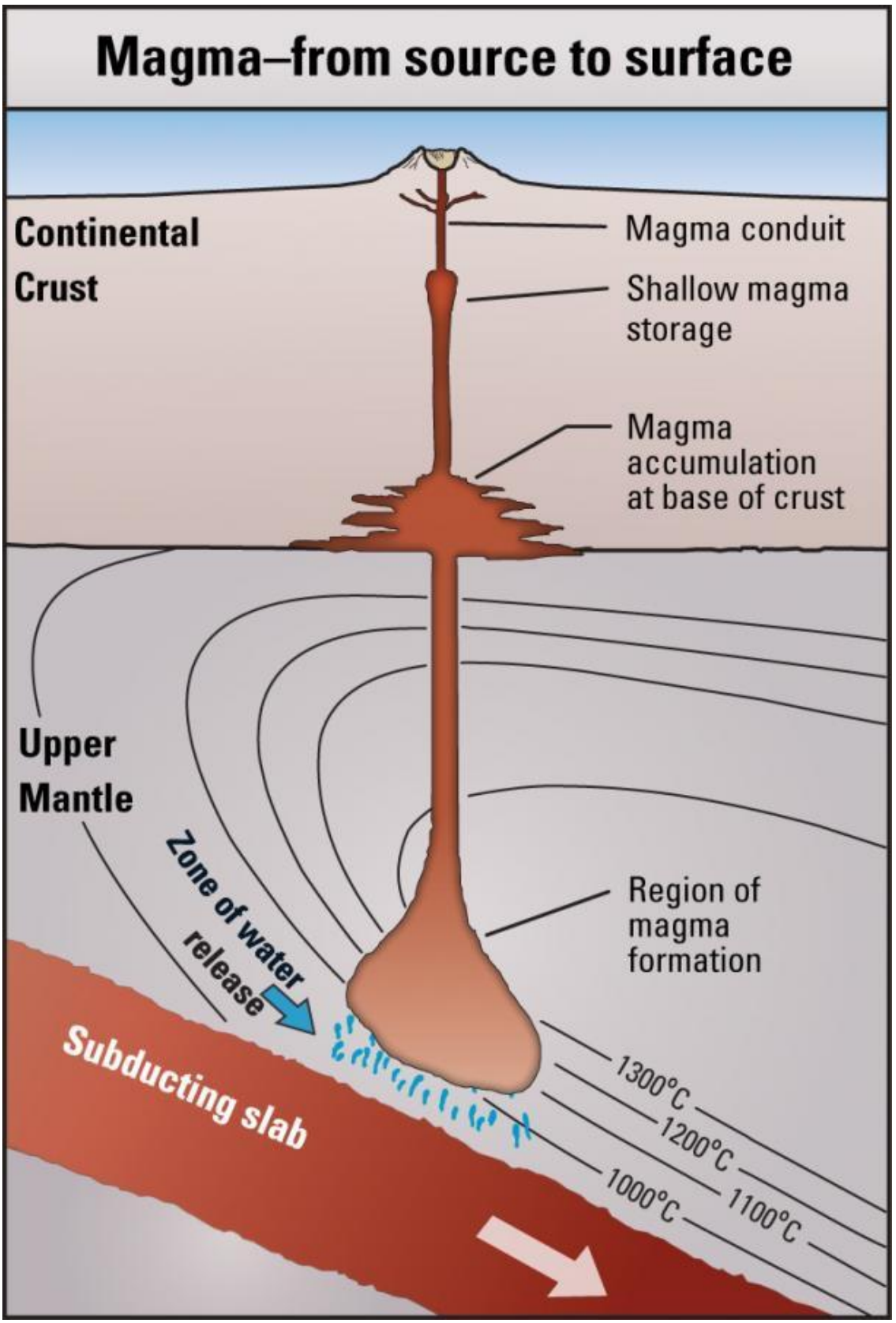

Рисунок 4.3 — Дзурісін та ін. 2012 рік. Розташування утворення, накопичення та зберігання магми під горою Сент-Хеленс

Пемза — сильно везикулярний вулканічний викид, типово кремнієвий за складом. По суті, це магма, яка була спінена газами, що виділяються, а потім охолоджена та затверділа під час виверження. Риолітова пемза зазвичай має досить низьку щільність, щоб вона плавала на воді. Біля вентиляційного отвору гаряча пемза може накопичуватися і утворювати пемзовий конус.

Викиди — матеріали, які викидаються з вулкану шляхом вибуху.

Кремній описує магму, яка містить понад ~63% кремнезему і, як правило, в'язка, багата газом і має тенденцію до вибухового виверження. Включає ріоліт і дацит.

Кремнезем — діоксид кремнію, найбільш поширена породоутворююча сполука на Землі та переважна молекулярна складова вулканічних порід і магми. Він має властивість полімеризуватися в молекулярні ланцюги, збільшуючи в'язкість магми. Базальтова магма, що має нижчий рівень SiO_2, досить текуча, але зі збільшенням вмісту SiO_2 андезитові, дацитові та ріолітові магми поступово стають більш в'язкими. Оскільки розчиненому газу важче виходити з більш в'язкої магми, магма з вищим кремнеземом, як правило, вивергається більш вибухово.

Базальт — Вулканічна порода (або лава), яка характерно темного кольору (від сірого до чорного), містить від 45 до 53 відсотків кремнезему і багата залізом і магнієм. Базальтові лави більш текучі, ніж андезити або дацити, які містять більше кремнію.

Базальт є найпоширенішим типом гірських порід у земній корі (зовнішня частина від 10 до 50 км). Насправді більша частина дна океану складається з базальту.

Величезні виливи лави під назвою «базальти повені» зустрічаються на багатьох континентах. Базальти річки Колумбія, вивержені від 15 до 17 мільйонів років тому, покривають більшу частину південно-східного Вашингтона та регіони прилеглих штатів Орегон та Айдахо.

Базальтова магма зазвичай утворюється шляхом прямого плавлення мантії Землі, області Землі під зовнішньою корою. На материках мантія починається на глибинах від 30 до 50 км.

Щитові вулкани, такі як ті, що складають острови Гаваї, майже повністю складаються з базальту. На рисунку 4.4 можна побачити гарячу базальтову лаву, що тече по поверхні охолодженого потоку базальтової лави.

Базальт — це тверда чорна вулканічна порода з вмістом кремнезему (SiO2) менше 52 вагових відсотків. Через низький вміст кремнезему в базальті він має низьку в’язкість (стійкість до течії). Тому базальтова лава може швидко та легко переміщатися >20 км від жерла. Низька в’язкість зазвичай дозволяє вулканічним газам виходити назовні, не утворюючи величезних стовпів виверження. Однак фонтани базальтової лави та виверження тріщин все ще утворюють вибухові фонтани висотою в сотні метрів. Поширені мінерали в базальті включають олівін, піроксен і плагіоклаз. Базальт вивергається при температурах від 1100 до 1250 °C.

Рисунок 4.4 — Свонсон, Дон А. Гаряча базальтова лава, що тече по поверхні охолодженого потоку базальтової лави.

Лава — Загальний термін для магми (розплавленої породи), яка була вивержена на поверхню Землі і зберігає свою цілісність у вигляді рідини або в’язкої маси, а не вибухає на уламки.

Рисунок 4.5 — Гріггс, Дж.Д. Лава рухається по землі у вигляді потоку пахоехо, вулкан Кілауеа, Гаваї, 13.11.1985

Андезит — це вулканічна порода від сірого до чорного кольору з вмістом кремнезему (SiO2) приблизно від 52 до 63 вагових відсотків. Андезити містять кристали, що складаються в основному з плагіоклазу польового шпату та одного або кількох мінералів піроксену (клінопіроксену та ортопироксену) та меншої кількості рогової обманки. У нижній частині діапазону кремнезему андезитова лава також може містити олівін. Андезитова магма зазвичай вивергається зі стратовулканів у вигляді потоків густої лави, деякі з яких

досягають кількох кілометрів у довжину. Андезитова магма також може генерувати сильні вибухові виверження, утворюючи пірокластичні потоки та сплески, а також величезні колони виверження. Андезити вивергаються при температурах від 900 до 1100 °C.

Факти:

1. Слово андезит походить від гір Анд, розташованих уздовж західного краю Південної Америки, де поширені андезитові породи.
2. Андезит був основним типом порід, вивержених під час великого виверження Кракатау 1883 року.

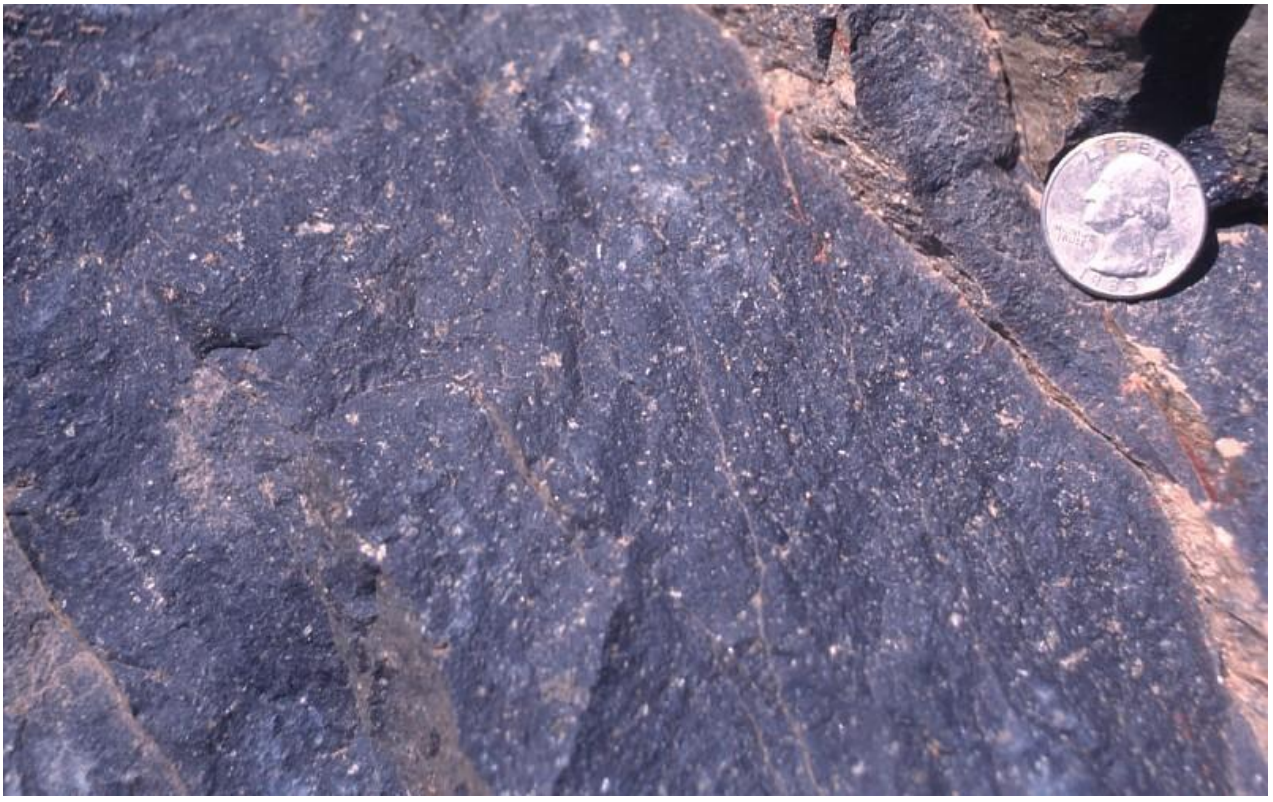

Рисунок 4.6 — Андезитовий потік лави через виверження вулкану. Близький вид на потік андезитової лави вулкана Brokeoff, Каліфорнія

Пірокластичний потік — гаряча (зазвичай >800 °C), хаотична суміш уламків гірських порід, газу та попелу, яка швидко (десятки метрів за секунду) рухається від вулканічного жерла або фронту потоку, що руйнується.

Базовий сплеск — кільцеподібна хмара газу та зваженого твердого сміття, що рухається радіально назовні з високою швидкістю від основи вертикального стовпа виверження. Може супроводжувати фреатомагматичні висипання.

Стовп виверження — висхідна, вертикальна частина маси уламків виверження та вулканічного газу, що піднімається безпосередньо над вулканічним жерлом. Нижче в атмосфері колони зазвичай поширюються збоку в шлейфи або парасолькові хмари.

Хмара тефри та газів, що утворюється за вітром від вулкана, що вивергається, називається хмарою виверження. Вертикальний стовп з тефри та газів, що піднімається безпосередньо над вентиляційним отвором, є колоною виверження. Хмари виверження можуть дрейфувати на тисячі кілометрів за вітром і часто стають все більш поширеними на більшій території зі збільшенням відстані від виверження (зверніть увагу на віялоподібну хмару виверження на фотографіях ліворуч). Великі виверження хмари можуть оточити Землю за кілька днів.

Хмара виверження часто використовується як хмара шлейфу або попелу.

На рисунку 4.7 можна побачити хмару попелу від виверження вулкана Гуагуа-Пічінча в 10 км на захід від Кіто, Еквадор, 7 жовтня 1999 року.

Рисунок 4.7 — Хмара попелу від виверження Гуагуа-Пічінча, Еквадор, USGS.

Тефра — будь-який тип і розмір уламків породи, які примусово викидаються з вулкана і рухаються повітряним шляхом під час виверження (включаючи попіл, бомби та окалини).

Окалина — везикулярний вулканічний викид, по суті магма, яка була спінена газами, що вириваються. Це текстурний варіант пемзи, з шкірою, як правило, менш везикулярною, щільнішою і зазвичай андезитною або базальтовою.

Фреатомагматичний висип — Виверження, яке включає як магму, так і воду, які зазвичай вибухово взаємодіють, що призводить до одночасного викиду пари та пірокластичних уламків.

На рисунку 4.8 можна побачити фреатомагматичну колону виверження, що піднімається зі сходу кратера Укінрек Маар, Аляска. Фото зроблено близько 17:00, вид на південний схід.

Рисунок 4.8 — Департамент риби та дичини Аляски, 1977-04-06

Дацит — Вулканічна порода (або лава), яка характерно світлого кольору і містить від 62 до 69 відсотків кремнезему та помірну кількість натрію та калію. Дацитові лави є в’язкими і мають тенденцію утворювати товсті блокові

потоки лави або круті купи лави, які називаються лавовими куполами. Дацитові магми мають тенденцію до вибухового виверження, таким чином також викидаючи велику кількість попелу та пемзи.

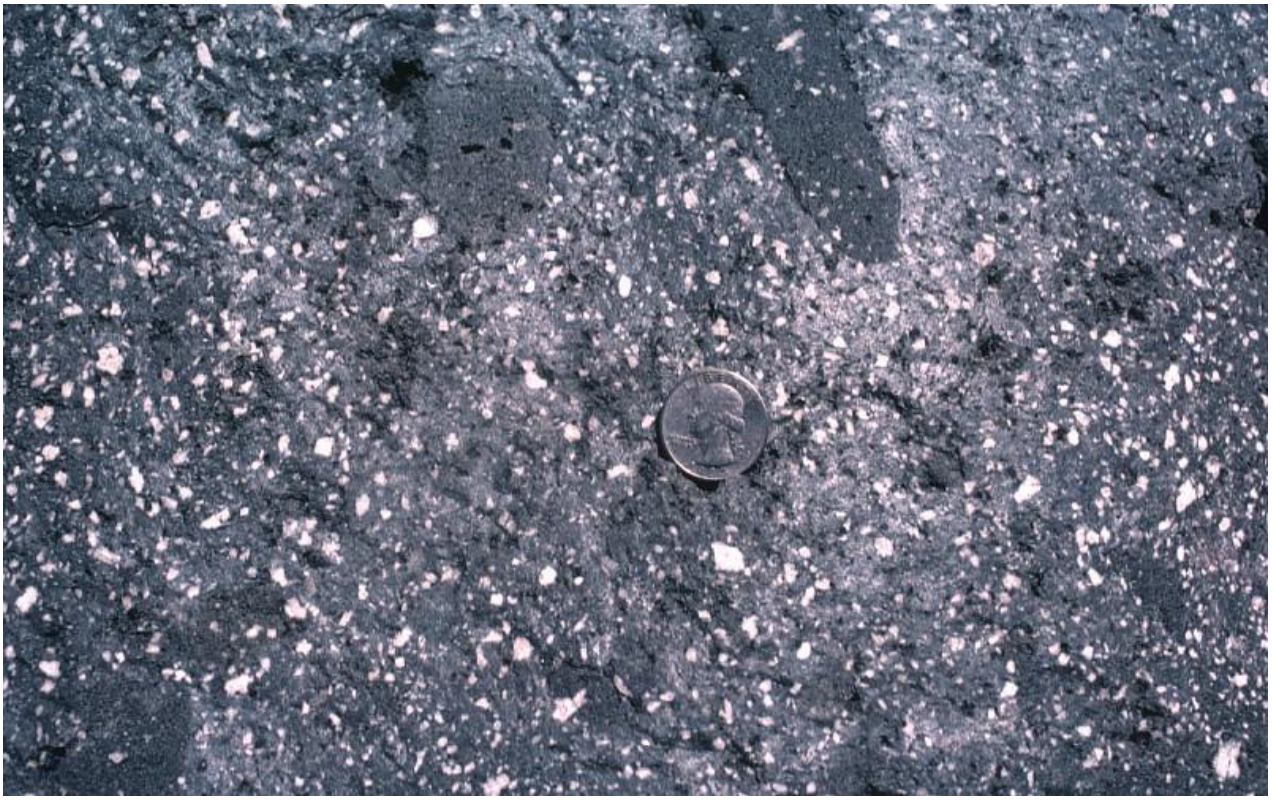

Рисунок 4.9 — Близький вид на дацитову лаву під час виверження Лассен-Пік у травні 1915 року, Каліфорнія

Купол — крутостороння маса в'язкої і часто блокованої лави, видавлена з отвору; зазвичай має округлу верхівку і охоплює приблизно круглу область. Можуть бути ізольованими (наприклад, скляний потік Medicine Lake) або, як альтернатива, пов’язані з частками або потоками лави з одного отвору. Зазвичай кремній (ріоліт або дацит) за складом.

Вулканічний купол у Каскадах — пік Лассен на півночі Каліфорнії. Пік Лассен є частиною більшого вулканічного центру, і його останнє виверження між 1914 і 1917 роками додало до його вершини новий потік лави.

Вулканічний ланцюг озера Моно-Іньо-Кратери в північно-центральній Каліфорнії складається з 13 куполів; чотири вивергалися лише близько 600 років тому (кратер Панум, Південний Дедмен, Обсидіановий потік і куполи Гласс-Крік).

Три найактивніших стратовулкани в Каскадах – гора Сент-Хеленс, гора Шаста та пік Льодовика – вивергали лавові куполи на своїх вершинах. До виверження гори Сент-Хеленс у 1980-1986 роках її симетричний верховий конус був укритий лавовим куполом, виверженим між 1500-ми і кінцем 1700-х років.

Хоча вулканічні куполи будуються в результаті невибухових вивержень в'язкої лави, куполи можуть генерувати смертельні пірокластичні потоки. Сторони купола або вивергається лавовий потік на куполі можуть обрушитися вниз по крутому схилу, утворюючи гарячу лавіну з уламків гарячої лави та газу (пірокластичний потік). Нещодавні виверження лавових куполів на вулкані Унзен в Японії та на пагорбах Суфрієр на Монтсерраті змусили тисячі людей покинути свої домівки.

Деякі куполи вивергають обсидіан, який являє собою вулканічне скло, яке може утворюватися в потоках лави з ріоліту або дациту. Більшість обсидіанів чорні, але відомі червоні, зелені та коричневі обсидіани. Обсидіан утворюється, коли магма охолоджується настільки швидко, що окремі мінерали не можуть кристалізуватися. На рисунку можна побачити вид з повітря на кратери Іньо з потоком обсидіана на передньому плані.

Вид на обсидіановий потік на північ; Лавовий купол Вілсона Батта (курган у центрі ліворуч), а також потоки лави та куполи Моно-кратерів видно за потоком Обсидіана.

Рисунок 4.10 — Брантлі, С.Р. 1998-07-29, вида на кратери Іньо з потоком обсидіана

Ріоліт — вулканічна порода (або лава), яка характерно світлого кольору, містить 69 або більше відсотків кремнезему, багата калієм і натрієм. Риоліт з низьким вмістом кремнію містить від 69 до 74 відсотків кремнезему. Висококремнеземний ріоліт містить від 75 до 80 відсотків кремнезему. Ріолітові лави є в'язкими і мають тенденцію утворювати товсті блокові потоки лави або круті купи лави, які називаються лавовими куполами. Ріолітові магми мають тенденцію до вибухового виверження, зазвичай також утворюючи рясні попіл і пемзу.

Потоки лави — це маси розплавлених порід, які виливаються на поверхню Землі під час ефузивного виверження. Як рухома лава, так і утворився затверділий осад називають лавовими потоками. Через широкий діапазон (1) в'язкості різних типів лави (базальту, андезиту, дациту та ріоліту); (2) викид лави під час виверженъ; і (3) характеристики джерела виверження і

топографії, по якій рухається лава, потоки лави бувають найрізноманітніших форм і розмірів.

На рисунку 4.11 можна побачити виверження отворів у північно-східній рифтовій зоні Мауна-Лоа поблизу Червоного пагорбу 25 березня 1984 року, яке спричинило потоки масивної лави вниз по розлому до Кулані.

Рисунок 4.11 — Виверження отворів у північно-східній рифтовій зоні Мауна-Лоа поблизу Червоного пагорбу, USGS, 2012-02-06

Обсидіан — це щільне вулканічне скло, зазвичай за складом ріоліту і зазвичай чорного кольору. Обсидіан утворюється в лавових потоках, де лава охолоджується так швидко, що кристали не встигають рости.

Рисунок 4.12 — USGS, зразок обсидіану, щільне вулканічне скло, склад ріоліту.

Льодовик — маса льоду, яка свідчить про течію протягом багатьох років, про що свідчить наявність ліній течії, тріщин та інших геологічних ознак.

Рисунок 4.13 — Скурлок, Джон, 09.10.2009. Кратер Кармело на горі Бейкер на вершині прорваний льодовиком Рузвельта, вид на південний схід, Вашингтон.

Кора є найвіддаленішим основним шаром землі, товщина якого в усьому світі становить від 10 до 65 км. Найвищі 15-35 км земної кори досить крихкі, щоб викликати землетруси.

На рисунку можна побачити блок-схему, що показує субдукцію плити Хуан де Фука під Північноамериканську плиту вздовж Каскадської западини, яка є західним краєм зони субдукції Каскадії. Океанічна кора утворюється в результаті виверженьуздовж хребта Хуан-де-Фука. Коли плита Хуан де Фука дрейфує на схід, вона охолоджується, стає більш щільною і врешті-решт занурюється під менш щільну Північноамериканську плиту в западині Каскадії. Вода, що виділяється з субдукційної плити, спричиняє часткове танення верхньої мантії, утворюючи магму, яка підтримує Каскадний хребет вулканів (чорні трикутники).

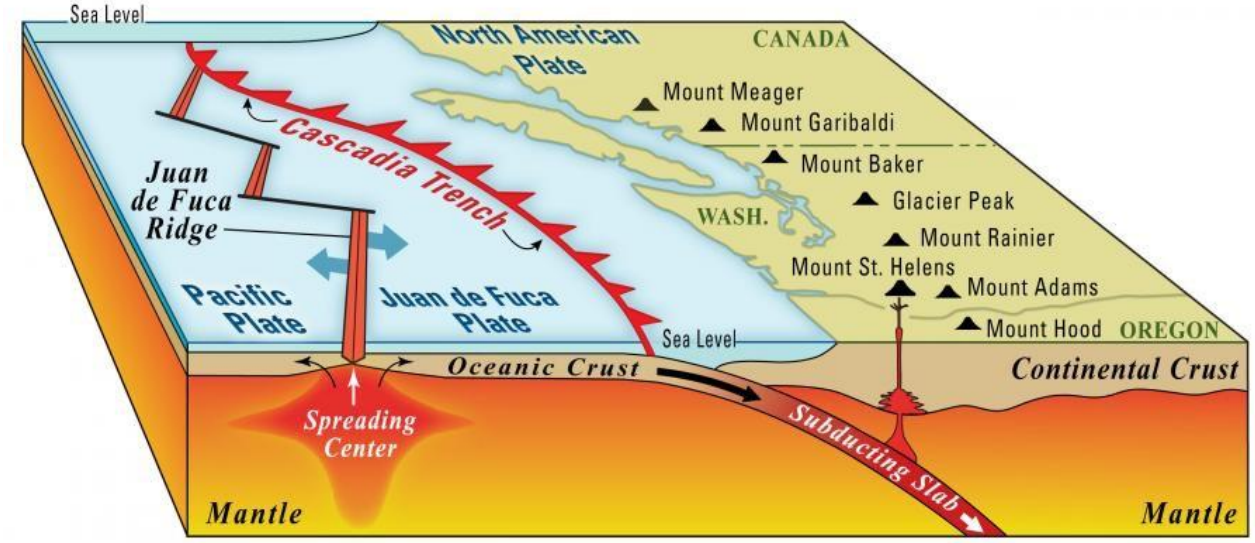

Рисунок 4.14 — Субдукція плити Хуан-де-Фука під Північноамериканську плиту вздовж Каскадної западини є відповідальним за вулканізм Каскадного хребта, Дзурісін та ін., 2013.

Сейсмічність — явище землетрусів, викликане крихким розтріскуванням гірських порід у земній корі. Синонім сейсмічної активності.

Мантія — це частина надр Землі між металевим зовнішнім ядром (під мантією) і корою (над мантією).

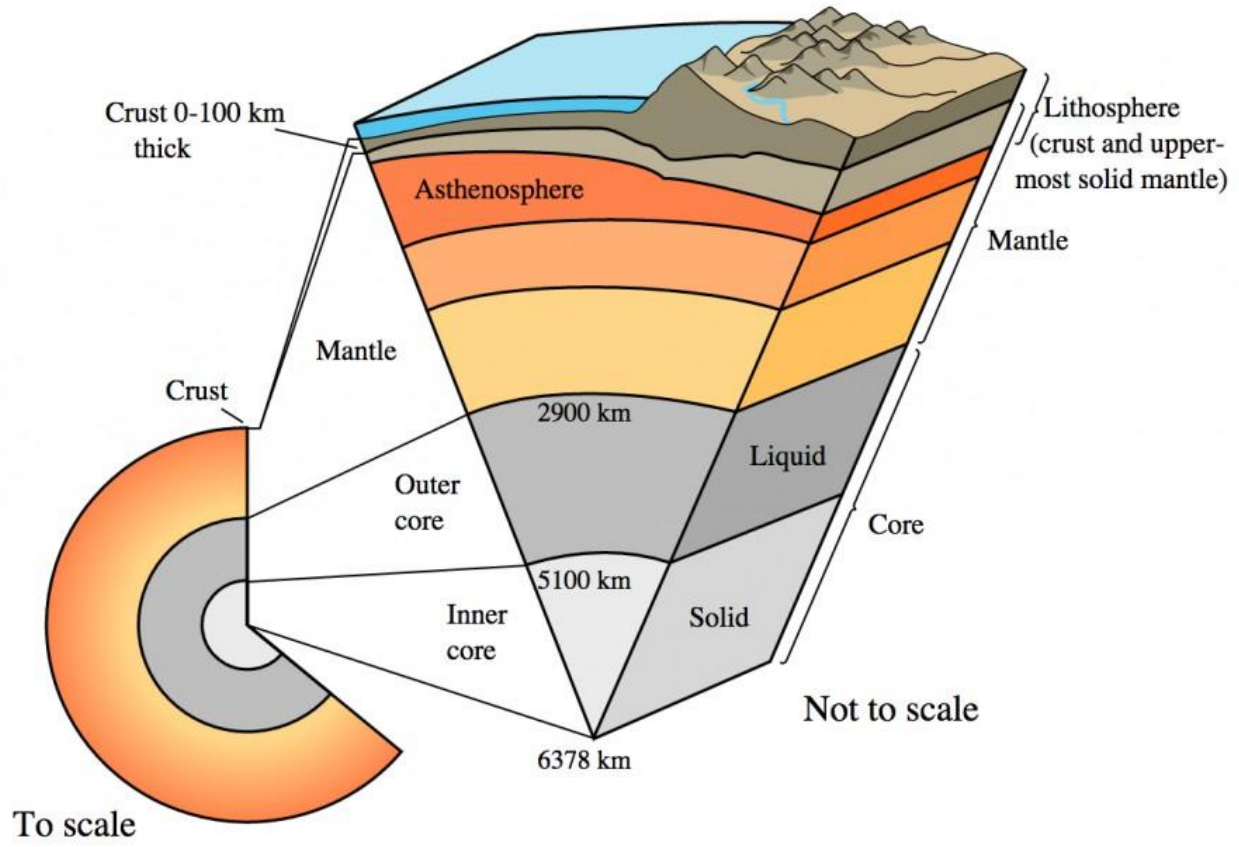

Рисунок 4.15 — Розрізи, що показують внутрішню структуру Землі. Ліворуч: масштабний малюнок показує, що земна кора дуже тонка. Праворуч: не в масштабі, більше деталей трьох основних шарів (кора, мантія, ядро), USGS.

Ефузивне виверження — виверження, у якому переважає виливання лави на землю, часто називають ефузивним виверженням (на відміну від насильницького дроблення магми вибуховими виверженнями). Потоки лави, що утворюються в результаті ефузійних вивержень, різняться за формою, товщиною, довжиною та шириною залежно від типу виверженої лави, викиду, схилу землі, по якій рухається лава, та тривалості виверження.

Наприклад, базальтова лава може перетворитися на «а‘а» або «пахоехо» і текти в глибоких вузьких каналах або тонкими широкими листами. Андезитова лава зазвичай утворює потужні короткі потоки, а дацитова лава часто утворює круті горби, які називаються лавовими куполами.

На рисунку 4.16 можна побачити, як базальтова лава вивергається з конуса на вулкані Кілауеа, Гаваї. Лава, що виливається з конуса, утворила серію ефузивних лавових каналів і потоків.

Найбільші відомі ефузивні виверження на Землі встилили базальтовою лавою сотні і тисячі квадратних кілометрів її поверхні. Вивергаючи сотні потоків лави протягом кількох мільйонів років, вчені називають утворені відклади базальтом повені, а території, що покриваються, — базальтом плато. Однією з таких територій, покритих паводковим базальтом, є регіон плато Колумбія на сході Вашингтона та Орегону. Приблизно від 17 до 14 мільйонів років тому серія виверженъ загальним об’ємом понад 175 мільйонів км3 охопила площу близько 165 000 км2. Ще більш об'ємні базальти плато розташовані в Південній Америці, Південній Африці, Індії.

Рисунок 4.16 — Виверження базальтової лави на вулкані Кілауеа, Гаваї. Гріггс, Дж.Д., 1984-01-31

Осколкове балістичне сміття. Багато бомб набувають округлу аеродинамічну форму під час подорожі в повітрі. Вулканічні бомби включають бомби з хлібною кіркою, стрічкові бомби, бомби веретена (із закрученими кінцями), сфероїдні бомби та бомби з «коров'яним гною». Вулканічні блоки зазвичай складаються із застиглих шматків старих потоків лави, які були частиною конуса вулкана. На рисунку можна побачити відслонення туфу з потоку попелу, північно-західні фланги вулкана Ньюбері, штат Орегон, було виявлено під час виверження, що утворило кальдеру, приблизно 75 000 років тому. Темні лавові бомби переносилися пірокластичними потоками.

Більші темні лавові бомби переносилися в пірокластичному потоці, який підтримувався всередині попелястої матриці.

Рисунок 4.17 — Відслонення туфу з потоку попелу, північно-західні фланги вулкана Ньюбері, штат Орегон, Доннеллі-Нолан, Джулі М. 17.06.2011.

Виверження часто змушують людей, які живуть поблизу вулканів, покинути свою землю та будинки, іноді назавжди. Подалі міста, посіви, промислові підприємства, транспортні системи, літаки та електричні мережі все ще можуть бути пошкоджені тефрою, попелом, лахарами та повенями.

Лахар — також називається вулканічним селем або потоком сміття. Суміш води та вулканічного сміття, яка швидко рухається за течією. Консистенція може варіюватися від каламутної води для посуду до вологого цементу, залежно від співвідношення води до сміття. Вони утворюються різними способами, головним чином у результаті швидкого танення снігу та льоду пірокластичними потоками, інтенсивних опадів на пухких відкладеннях вулканічних порід, прориву озера, перекритого вулканічними відкладеннями, і як наслідок сміттєвих лавин.

Рисунок 4.18 — Лахари з гори Сент-Хеленс перенесли цей великий валун за течією, коли він зривав дерева й залишив товстий відклад грязі після виверження 18 травня 1980 року. Мудді-Рівер, Вашингтон, Топинка, Лин 16.09.1980

Уламкова лавина — рухомі маси каменів, ґрунту та снігу, які виникають, коли схили гори або вулкана обрушуються та сповзають вниз. Коли рухоме сміття мчить вниз по вулкану і в долини річок, воно включає воду, сніг, дерева, мости, будівлі та все інше на шляху. Уламкові лавини можуть пройти кілька кілометрів, перш ніж зупинитися, або вони можуть перетворитися на більш багаті водою лахари, які проходять багато десятків кілометрів за течією.

Уламкова лавина мчить по узбіччю вулкана на дно долини. Багато таких сміттєвих лавин перетворюються на лахари і йдуть за десятки кілометрів від

вулкана. Як правило, шрам, утворений лавиною, залишає кратер у формі підкови на стороні вулкана.

На рисунку можна побачити вид на уламкове лавинне відкладення з північного заходу гори Сент-Хеленс після виверження 18 травня 1980 року. Зсувний уступ, кратер у формі підкови вгорі праворуч, відкриває вентиляційний отвір і відкривається на північ.

Після виверження 18 травня 1980 року висота гори Сент-Хеленс становила всього 8364 фути (2550 метрів), а вулкан мав підковоподібний кратер шириною в одну милю (1,5 кілометра) і приблизно 600 м (2000 футів). Вид тут з північного заходу.

Рисунок 4.19 — Вид на уламкове лавинне відкладення з північного заходу гори Сент-Хеленс після виверження 18 травня 1980 року, Казадеваль, Том 16.09.1980

На щастя, вулкани демонструють попередні коливання, які, якщо їх виявити та проаналізувати вчасно, дозволяють передбачити виверження та попередити громади, які загрожують ризику. Час попередження, що передує вулканічним подіям, зазвичай дає достатньо часу для постраждалих громад для реалізації планів реагування та заходів з пом’якшення наслідків.

4.3 Технологія застосування розробленої моделі.

Розроблена модель може бути імплементована як у приватний програмний продукт, так і для подальших наукових досліджень. Перевага цієї моделі, в першу чергу, полягає в часі, затраченому на прогнозування.

Практична імплементація може залежати від багатьох факторів та має широку варіативність застосування. В даній роботі наведено декілька прикладів можливої імплементації, яку будь-хто може використовувати для використання у своєму продукті.

4.3.1 Імплементація моделі за допомогою pickle

Одним з можливих способів серіалізації python об’єктів є бібліотека pickle. За її допомогою можна серіалізувати алгоритми машинного навчання та зберегти серіалізований формат у файл.

Пізніше можна завантажити цей файл, щоб десеріалізувати модель і використовувати його для створення нових передбачень у будь-якому продукті та для будь-яких цілей.

Перш за все, імпортуємо бібліотеку pickle, як показано на рисунку 4.20.

```
# збереження моделі за допомогою pickle

import pickle
```

Рисунок 4.20 — Імпорт пакету pickle

Після імпорту пакету, необхідно зберегти навчену модель у файл. Скористаємося методом pickle.dump та назвемо майбутній файл, в якому буде зберігатися модель, 'pickle_model.save', що можна побачити з рисунку 4.21.

```
# збереження моделі у файл

filename = 'pickle_model.save'
pickle.dump(clf, open(filename, 'wb'))
```

Рисунок 4.21 — Збереження моделі у файл за допомогою pickle

На цьому етапі модель, яка була навчена в розділі 3, серіалізована у файл ‘pickle_model.save`, за допомогою якого вона займає дуже мало місця та може бути легко мобілізована, а потім, при нагоді, десеріалізована, як показано на рисунку 4.22.

```
# вигруження моделі з файлу

loaded_model = pickle.load(open(filename, 'rb'))
result = loaded_model.score(X_test, y_test)
print(result)
```

```
0.9876368600200882
```

Рисунок 4.22 — Десеріалізація моделі за допомогою pickle

На цьому етапі також можна побачити результат роботи моделі за метрикою accuracy на нових даних.

4.3.2 Імплементація моделі за допомогою joblib

Аналогічним способом можна зберігати модель за допомогою ще одного пакета — joblib. Joblib є частиною екосистеми SciPy і надає утиліти для пайплайнів Python задач.

Він надає утиліти для збереження та завантаження об’єктів Python, які ефективно використовують структури даних NumPy. Це може бути корисно для алгоритмів машинного навчання, які вимагають багато параметрів або зберігають весь набір даних.

Імпортуємо бібліотеку joblib, як показано на рисунку 4.23.

```
# збереження моделі за допомогою joblib

import joblib
```

Рисунок 4.23 — Імпорт пакету joblib

Після імпорту пакету, необхідно зберегти навчену модель у файл. Скористаємося методом pickle.dump та назвемо майбутній файл, в якому буде зберігатися модель, 'joblib_model.save', що можна побачити з рисунку 4.24.

```
# збереження моделі у файл

filename = 'joblib_model.save'

joblib.dump(clf, filename)
```

```
['joblib_model.save']
```

Рисунок 4.24 — Збереження моделі у файл за допомогою joblib

На цьому етапі модель, яка була навчена в розділі 3, серіалізована у файл ‘joblib_model.save`, за допомогою якого вона займає дуже мало місця та може бути легко мобілізована, а потім, при нагоді, десеріалізована, як показано на рисунку 4.25.

```
# вигруження моделі з файлу

loaded_model = joblib.load(filename)

result = loaded_model.score(X_test, y_test)

print(result)
```

```
0.9876368600200882
```

Рисунок 4.25 — Десеріалізація моделі за допомогою joblib

Зазвичай прогнозування вулканів роблять висококваліфіковані аналітики та науковці, які витрачають багато часу та ресурсів, які могли би бути застосовані в більш важливих задачах, які не можна виконати автоматично, так як тепер прогнозування вулканів можна зробити за допомогою запропонованої моделі.

Результати цієї роботи дозволять по-новому поглянути на проблему стихійних лих. Отримані результати неймовірно важливі не лише для наукової спільноти в подальших дослідженнях, а і в практичному застосуванні для порятунку життів людей.

ВИСНОВКИ

Було досліджено поведінку вулканів на основі їх сейсмічних сигналів, створено математичний апарат для тлумачення предикторів незвичайних подій в контексті вулканів для систематизації залежності від них часу до виверження, розглянуто локальні та зарубіжні теоретичні джерела для глибокого аналізу даних сенсорів.

Виконано ґрунтовний аналіз проблеми виверження вулканів, окреслена географічна розповсюдженість та наслідки подій. Проаналізовані існуючі методи дослідження та вирішення поставленої задачі, враховано переваги та недоліки кожного з них. Окреслені методи та джерела збору даних.

Детально описано обраний метод збору даних для дослідження, доступні методи вирішення задачі регресії, особлива увага приділяється CatBoost, який був обраний для імплементації.

Реалізовано модель, використовуючи методи збору інформації, обробки та очищення даних, побудови архітектури моделі машинного навчання з використанням алгоритму CatBoost для навчання, оптимізації моделі заа допомогою KFold та оцінки моделі за допомогою метрики MAPE. Ця модель не може бути порівняна з будь-якими іншими моделями, а, отже, не потребує доказів своїх переваг, оскільки вона унікальна в контексті вибору алгоритму при підході до вирішення задачі.

Результати цієї роботи дозволять по-новому поглянути на проблему стихійних лих. Отримані результати важливі не лише для наукової спільноти в подальших дослідженнях, а і в практичному застосуванні для порятунку життів людей.

Результати роботи можуть стати дуже впливовим інструментом для завчасного інформування про природну катастрофу, що наближається, попередження негативних наслідків та значного зменшення кінцевих проблем.

ПЕРЕЛІК ПОСИЛАНЬ

1. Acocella, V. 2021. Volcano-Tectonic Processes. Berlin: Springer.
2. Sparks, R. S. J., and Aspinall, W. P. 2004. Volcanic activity: frontiers and challenges in forecasting, prediction and risk assessment. State Planet Front. Challenges Geophys. Geophys. Monogr. 19:359–73. doi: 10.1029/150GM28
3. Marzocchi, W., and Bebbington, M. S. 2012. Probabilistic eruption forecasting at short and long time scales. Bull. Volcanol. 74:1777–805. doi: 10.1007/s00445-012-0633-x
4. Winson, A. E. G., Costa, F., Newhall, C. G., and Woo, G. 2014. An analysis of the issuance of volcanic alert levels during volcanic crises. J. Appl. Volcanol. 3:14. doi: 10.1186/s13617-014-0014-6
5. Poland, M. P., and Anderson, K. R. 2020. Partly cloudy with a chance of lava flows: forecasting volcanic eruptions in the twenty-first century. J. Geophys. Res. 125:e2018JB016974. doi: 10.1029/2018JB016974
6. Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V., INSTANCE – the Italian seismic dataset for machine learning, Earth Syst. Sci. Data, 13 (12), 5509 – 5544, doi:10.5194/essd-13-5509-2021.
7. Веб-сайт. URL: https://eida.ingv.it/it/ (дата звернення: 20.12.2021).
8. Ross, Z. E., Meier, M.-A., and Hauksson, E.: P Wave Arrival Picking and First-Motion Polarity Determination With Deep Learning, J. Geophys. Res.-Sol. Ea., 123, 5120–5129, https://doi.org/10.1029/2017JB015251, 2018b
9. Meier, M.-A., Ross, Z. E., Ramachandran, A., Balakrishna, A., Nair, S., Kundzicz, P., Li, Z., Andrews, J., Hauksson, E., and Yue, Y.: Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning, J. Geophys. Res.-Sol. Ea., 124, 788–800,https://doi.org/10.1029/2018JB016661, 2019.

10. SCEDC (2013): Southern California Earthquake Center. Caltech.Dataset. doi:10.7909/C3WD3xH1. The SCEDC and SCSN are funded through U.S. Geological Survey Grant G20AP00037, and the Southern California Earthquake Center, which is funded by NSF Cooperative Agreement EAR-0529922 and USGS Cooperative Agreement 07HQAG0008.
11. Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C.: STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI, IEEE Access, 7, 179464–179476, https://doi.org/10.1109/ACCESS.2019.2947848, 2019.
12. "Enti di ricerca" [Research Institutions]. Ministry of Education, Universities and Research (in Italian). Archived from the original on 19 March 2012. Retrieved 5 July 2012.
13. "Earthquakes: Top 20 Institutions". Sciencewatch. Archived from the original on 26 June 2015.
14. EMSO, European Multidisciplinary Seafloor and water column Observatory
15. EPOS, European Plate Observing System
16. "Volcano Observatory Best Practices Workshop: Eruption Forecasting". U.S. Geological Survey. September 2011. Archived from the original on 22 February 2014.
17. Веб-сайт. URL: http://terremoti.ingv.it/bsi (дата звернення: 10.01.2022).
18. INSTANCE The Italian Seismic Dataset For Machine Learning, Alberto Michelini, Spina Cianetti, Sonja Gaviano, Carlo Giunchi, Dario Jozinović & Valentino Lauciani, Seismic Waveforms And Associated Metadata published 2021 in Istituto Nazionale di Geofisica e Vulcanologia (INGV) https://doi.org/10.13127/instance
19. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M.,

Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, arXiv [preprint], arXiv:1603.04467, 14 March 2016.

20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems 32, edited by: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R., Curran Associates, Inc., 8024–8035, available at:http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (last access: 19 November 2021), 2019.
21. Chollet, F. and others: Keras [code], available at: https://keras.io (last access: 25 November 2021), 2015.
22. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding, arXiv [preprint], arXiv:1408.5093, 20 June 2014
23. Набір даних: веб-сайт. URL: https://github.com/smousavi05/STEAD
24. Magrini, Fabrizio, Jozinović, Dario, Cammarano, Fabio, Michelini, Alberto, & Boschi, Lapo. (2020). LEN-DB - Local earthquakes detection: a benchmark dataset of 3-component seismograms built on a global scale [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3648232

25. Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., and Gerstoft, P.: Machine Learning in Seismology: Turning Data into Insights, Seismol. Res. Lett., 90, 3–14, https://doi.org/10.1785/0220180259, 2018.

26. Bergen, K. J., Johnson, P. A., de Hoop, M. V., and Beroza, G. C.: Machine learning for data-driven discovery in solid Earth geoscience, Science, 363, eaau0323, https://doi.org/10.1126/science.aau0323, 2019.

27. Dramsch, J. S.: Chapter One – 70 years of machine learning in geoscience in review, Adv. Geophys., 61, 1–55, https://doi.org/10.1016/bs.agph.2020.08.002, 2020.

28. Strollo, A., Cambaz, D., Clinton, J., Danecek, P., Evangelidis, C. P., Marmureanu, A., Ottemöller, L., Pedersen, H., Sleeman, R., Stammler, K., Armbruster, D., Bienkowski, J., Boukouras, K., Evans, P. L., Fares, M., Neagoe, C., Heimers, S., Heinloo, A., Hoffmann, M., Kaestli, P., Lauciani, V., Michalek, J., Odon Muhire, E., Ozer, M., Palangeanu, L., Pardo, C., Quinteros, J., Quintiliani, M., Antonio Jara-Salvador, J., Schaeffer, J., Schloemer, A., and Triantafyllis, N.: EIDA: The European Integrated Data Archive and Service Infrastructure within ORFEUS, Seismol. Res. Lett., 92, 1788–1795, https://doi.org/10.1785/0220200413, 2021.

29. Ingate, S.: The IRIS Consortium: Community Based Facilities and Data Management for Seismology, in: Earthquake Monitoring and Seismic Hazard Mitigation in Balkan Countries, NATO Science Series: IV: Earth and Environmental Sciences, edited by: Husebye, E. S., vol. 81, 121–132, Springer, Dordrecht, https://doi.org/10.1007/978-1-4020-6815-7_8, 2008.

30. Quinteros, J., Carter, J. A., Schaeffer, J., Trabant, C., and Pedersen, H. A.: Exploring Approaches for Large Data in Seismology: User and Data Repository Perspectives, Seismol. Res. Lett., 92, 1531–1540, https://doi.org/10.1785/0220200390, 2021.

31. Danecek, P., Pintore, S., Mazza, S., Mandiello, A., Fares, M., Carluccio, I., Della Bina, E., Franceschi, D., Moretti, M., Lauciani, V., Quintiliani, M., and Michelini, A.: The Italian Node of the European Integrated Data Archive, Seismol. Res. Lett., 92, 1726–1737, https://doi.org/10.1785/0220200409, 2021.
32. Квантування: веб-сайт. URL: https://catboost.ai/en/docs/concepts/quantization
33. Michelini, A., Margheriti, L., Cattaneo, M., Cecere, G., D'Anna, G., Delladio, A., Moretti, M., Pintore, S., Amato, A., Basili, A., Bono, A., Casale, P., Danecek, P., Demartin, M., Faenza, L., Lauciani, V., Mandiello, A. G., Marchetti, A., Marcocci, C., Mazza, S., Mele, F. M., Nardi, A., Nostro, C., Pignone, M., Quintiliani, M., Rao, S., Scognamiglio, L., and Selvaggi, G.: The Italian National Seismic Network and the earthquake and tsunami monitoring and surveillance systems, Adv. Geosci., 43, 31–38, https://doi.org/10.5194/adgeo-43-31-2016, 2016.
34. Margheriti, L., Nostro, C., Cocina, O., Castellano, M., Moretti, M., Lauciani, V., Quintiliani, M., Bono, A., Mele, F. M., Pintore, S., Montalto, P., Peluso, R., Scarpato, G., Rao, S., Alparone, S., Di Prima, S., Orazi, M., Piersanti, A., Cecere, G., Cattaneo, M., Vicari, A., Sepe, V., Bignami, C., Valoroso, L., Aliotta, M., Azzarone, A., Baccheschi, P., Benincasa, A., Bernardi, F., Carluccio, I., Casarotti, E., Cassisi, C., Castello, B., Cirilli, F., D'Agostino, M., D'Ambrosio, C., Danecek, P., Cesare, W. D., Bina, E. D., Di Filippo, A., Di Stefano, R., Faenza, L., Falco, L., Fares, M., Ficeli, P., Latorre, D., Lorenzino, M. C., Mandiello, A., Marchetti, A., Mazza, S., Michelini, A., Nardi, A., Pastori, M., Pignone, M., Prestifilippo, M., Ricciolino, P., Sensale, G., Scognamiglio, L., Selvaggi, G., Torrisi, O., Zanolin, F., Amato, A., Bianco, F., Branca, S., Privitera, E., and Stramondo, S.: Seismic Surveillance

and Earthquake Monitoring in Italy, Seismol. Res. Lett., 92, 1659–1671, https://doi.org/10.1785/0220200380, 2021.

35. Gutenberg, B. and Richter, C. F.: Frequency of earthquakes in California, B. Seismol. Soc. Am., 34, 185–188, 1944.
36. http://terremoti.ingv.it/en/webservices_and_software, останній доступ: 19 листопада 2021 р.).
37. Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J.: ObsPy: A Python Toolbox for Seismology, Seismol. Res. Lett., 81, 530–533, https://doi.org/10.1785/gssrl.81.3.530, 2010.
38. Megies, T., Beyreuther, M., Barsch, R., Krischer, L., and Wassermann, J.: ObsPy – What can it do for data centers and observatories?, Ann. Geophys.-Italy, 54, 47–58, https://doi.org/10.4401/ag-4838, 2011.
39. Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., and Wassermann, J.: ObsPy: a bridge for seismology into the scientific Python ecosystem, Computational Science & Discovery, 8, 014003, https://doi.org/10.1088/1749-4699/8/1/014003, 2015.
40. cf. http://www.fdsn.org/seed_manual/SEEDManual_V2.4.pdf, last access: 19 November 2021;
41. Boore, D. M.: Estimating Vs(30) (or NEHRP Site Classes) from Shallow Velocity Models (Depths < 30 m), B. Seismol. Soc. Am., 94, 591–597, https://doi.org/10.1785/0120030105, 2004.
42. Michelini, A., Faenza, L., Lanzano, G., Lauciani, V., Jozinović, D., Puglia, R., and Luzi, L.: The New ShakeMap in Italy: Progress and Advances in the Last 10 Yr, Seismol. Res. Lett., 91, 317–333,https://doi.org/10.1785/0220190130, 2019.
43. Ross, Z. E., Meier, M., Hauksson, E., and Heaton, T. H.: Generalized Seismic Phase Detection with Deep Learning, B. Seismol. Soc. Am., 108, 2894–2901, https://doi.org/10.1785/0120180080, 2018

44. Jozinović, D., Lomax, A., Štajduhar, I., and Michelini, A.: Rapid prediction of earthquake ground shaking intensity using raw waveform data and a convolutional neural network, Geophys. J. Int., 222, 1379–1389, https://doi.org/10.1093/gji/ggaa233, 2020.
45. Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C.: Earthquake transformer-an attentive deep-learning model for simultaneous earthquake detection and phase picking, Nat. Commun., 11, 3952, https://doi.org/10.1038/s41467-020-17591-w, 2020.
46. Веб-сайт. URL: https://esm-db.eu/, останній доступ: 19 листопада 2021
47. Lanzano, G., Luzi, L., Russo, E., Felicetta, C., D'Amico, M., Sgobba, S., and Pacor, F.: Engineering Strong Motion Database (ESM) flatfile [data set], Tech. Rep., Istituto Nazionale di Geofisica e Vulcanologia (INGV),https://doi.org/10.13127/esm/flatfile.1.0, 2018.

ДОДАТКИ

ДОДАТОК А

Код імпорту пакетів

```
import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as plt
import os
from tqdm import tqdm
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from catboost import CatBoostRegressor, Pool
from sklearn.metrics import mean_absolute_error, mean_squared_error
```

ДОДАТОК Б

Код обробки даних

```
def create_frame(data, data_time=None, type_data='train'):
    data = data.fillna(0)

    # основні статистики
    data_transform = data.describe().iloc[1:, :]

    # додаткові параметри
    # коефіцієнт асиметрії
```

```
        data_transform.loc['skew'] = data.skew().tolist()

        #середнє абсолютне відхилення
        data_transform.loc['mad'] = data.mad().tolist()

        #коефіцієнт ексцесу - міра гостроти піку розподілу випадкової
величини
        data_transform.loc['kurtosis'] = data.kurtosis().tolist()

         # додавання квантилів
         for i in range(0, 100, 5):
            if ((i!=25) & (i!=50)):
                str_col = f"{i}%"
                int_col = float(i)/100
                data_transform.loc[str_col] =
data_transform.quantile(int_col).tolist()
            else:
                continue
        data_transform = pd.DataFrame(data_transform.unstack()).reset_index()
        data_transform = data_transform.rename(columns={0: 'value'})
        data_transform['feature'] = data_transform['level_0'] + '_' +
data_transform['level_1']
        data_transform = data_transform.drop(['level_0', 'level_1'],
axis=1).set_index('feature').T

        if type_data=='train':
            data_transform['time'] = data_time
        return data_transform
```

```
all_train = pd.DataFrame()

for file in tqdm(train_list):
    df = pd.read_csv(PATH + 'train/' + file)
    data_time = train_time[train_time.segment_id ==
int(file.split('.')[0])].time_to_eruption.values[0]
    df = create_frame(df, data_time, type_data='train')
    all_train = all_train.append(df)

all_train = all_train.reset_index(drop=True)

all_test = pd.DataFrame()

for file in tqdm(test_list):
    df = pd.read_csv(PATH + 'test/' + file)
    df = create_frame(df, data_time=None, type_data='test')
    all_test = all_test.append(df)

all_test = all_test.reset_index(drop=True)
```

ДОДАТОК В

Код побудови алгоритму

```
X = all_train.drop('time',axis=1)
y = all_train['time']
test = all_test.copy()
X_train, X_test, y_train, y_test = train_test_split(
```

```
    X, y, test_size=0.25, shuffle=True, random_state=10)

X_train, X_val, y_train, y_val = train_test_split(
    X_train, y_train, test_size=0.25, shuffle=True, random_state=10)
def mape(y_true, y_pred):
    return np.mean(np.abs((y_pred-y_true)/y_true))
clf = CatBoostRegressor(loss_function='MAPE')
train_dataset = Pool(data=X_train,
                     label=y_train,
                     )

eval_dataset = Pool(data=X_val,
                    label=y_val,
                    )

clf.fit(train_dataset,
          use_best_model=True,
          verbose = 1,
          eval_set=eval_dataset)
y_pred = clf.predict(Pool(data=X_test))
```

ДОДАТОК Г

Код оптимізації моделі

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.25, shuffle=True, random_state=10)
n_fold = 5
cv = KFold(n_splits=n_fold, shuffle=True, random_state=10)
```

```
prediction = np.zeros(len(test))
mape_, mae, rmse = [], [], []
params = {  'iterations':1000,
            'learning_rate':0.1,
            'depth':6,
            'eval_metric':'MAPE' }

for fold, (train_index, val_index) in enumerate(cv.split(X)):
    X_train = X.iloc[train_index,:]
    X_val  =  X.iloc[val_index,:]
    y_train  =  y.iloc[train_index]
    y_val = y.iloc[val_index]
    clf = CatBoostRegressor(**params)
     train_dataset = Pool(data=X_train,
                  label=y_train, )
    eval_dataset = Pool(data=X_val,
                  label=y_val, )
    clf.fit(train_dataset,
            use_best_model=True,
            verbose = 0,
            eval_set=eval_dataset)
    y_pred = clf.predict(Pool(data=X_test))

    mape_.append(mape(y_test, y_pred))
    mae.append(mean_absolute_error(y_test, y_pred))
    rmse.append(np.sqrt(mean_squared_error(y_test, y_pred)))
    print(f"fold: {fold}, MAPE: {mape(y_test, y_pred)}")
    print(f"fold: {fold}, MAE: {mean_absolute_error(y_test, y_pred)}")
```

```
            print(f"fold: {fold}, RMSE: {np.sqrt(mean_squared_error(y_test,
y_pred))}")
        prediction += clf.predict(Pool(data=test))
        prediction /= n_fold
```

ДОДАТОК Д

Код імплементації моделі за допомогою pickle

```
import pickle

filename = 'pickle_model.save' pickle.dump(clf,

open(filename, 'wb')) loaded_model =

pickle.load(open(filename, 'rb')) result =

loaded_model.score(X_test, y_test) print(result)
```

ДОДАТОК Е

Код імплементації моделі за допомогою joblib

```
import joblib

filename = 'joblib_model.save'

joblib.dump(clf, filename)
```

```
loaded_model = joblib.load(filename)

result = loaded_model.score(X_test, y_test)

print(result)
```