تخمین محل و بزرگای زمین‌لرزه‌ها با استفاده از روش‌های نوین یادگیری ماشین

سعید سلطانی مقدم

همکاران پروژه:

انوشیروان انصاری، لیلا اعتمادسعید، محمد تاتار، میثم محمودآبادی

فهرست مطالب

- مروری بر روش‌های مکان‌یابی مبتنی بر یادگیری ماشین،
- معرفی روش مورد استفاده،
- ارزیابی مدل،
- کاربرد در یک نمونه واقعی،

کاتالوگ زمین‌لرزه‌ها چه کاربردهایی دارد؟

Accurate Catalogs Drive Geosciences:

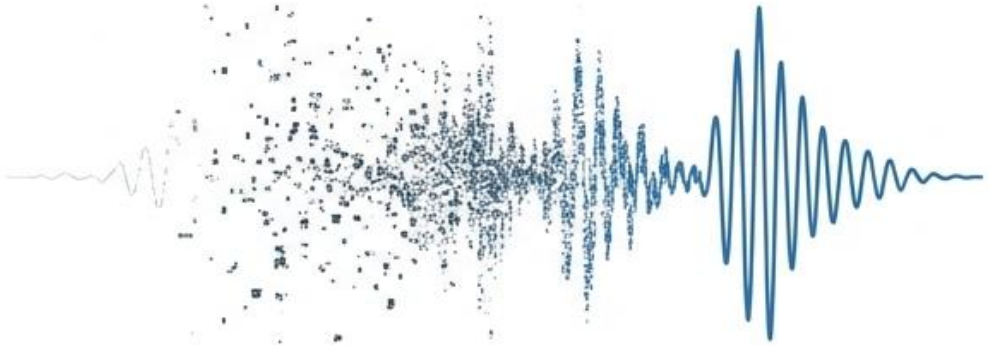
1. Fault geometry investigations

2. Seismic risk & insurance assessments

3. Emergency response planning

4. Earthquake Early Warning Systems (EEWS)

The Fragility of Routine Locators:

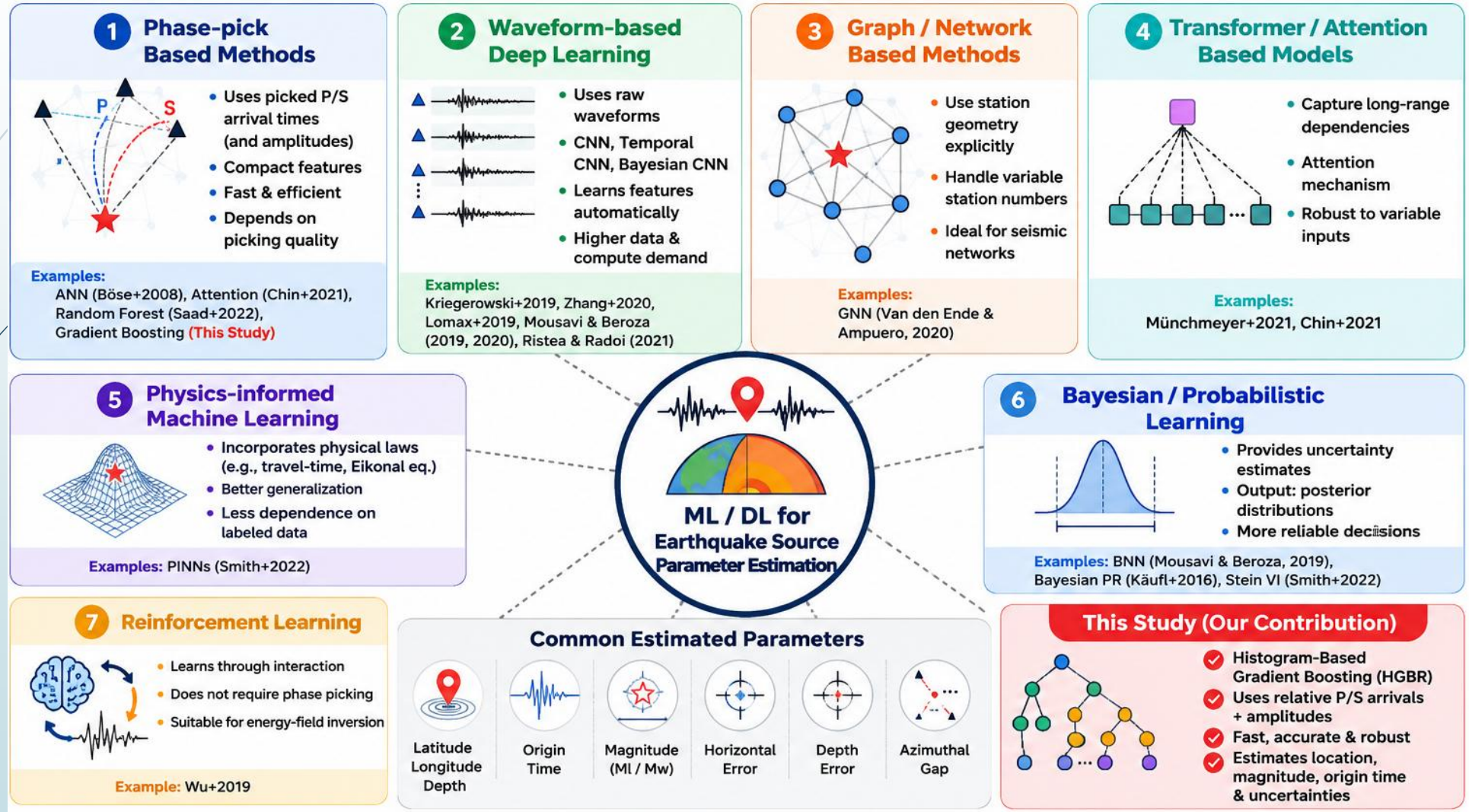


Routine locators (Hypo71, Hypocenter, HypoDD) rely heavily on pristine data. They degrade rapidly when faced with:

- Sparse station coverage or network outages.
- Inaccurate phase picking (noise/GPS drift).
- Velocity model uncertainties.

تهیه شده توسط: https://notebooklm.google/

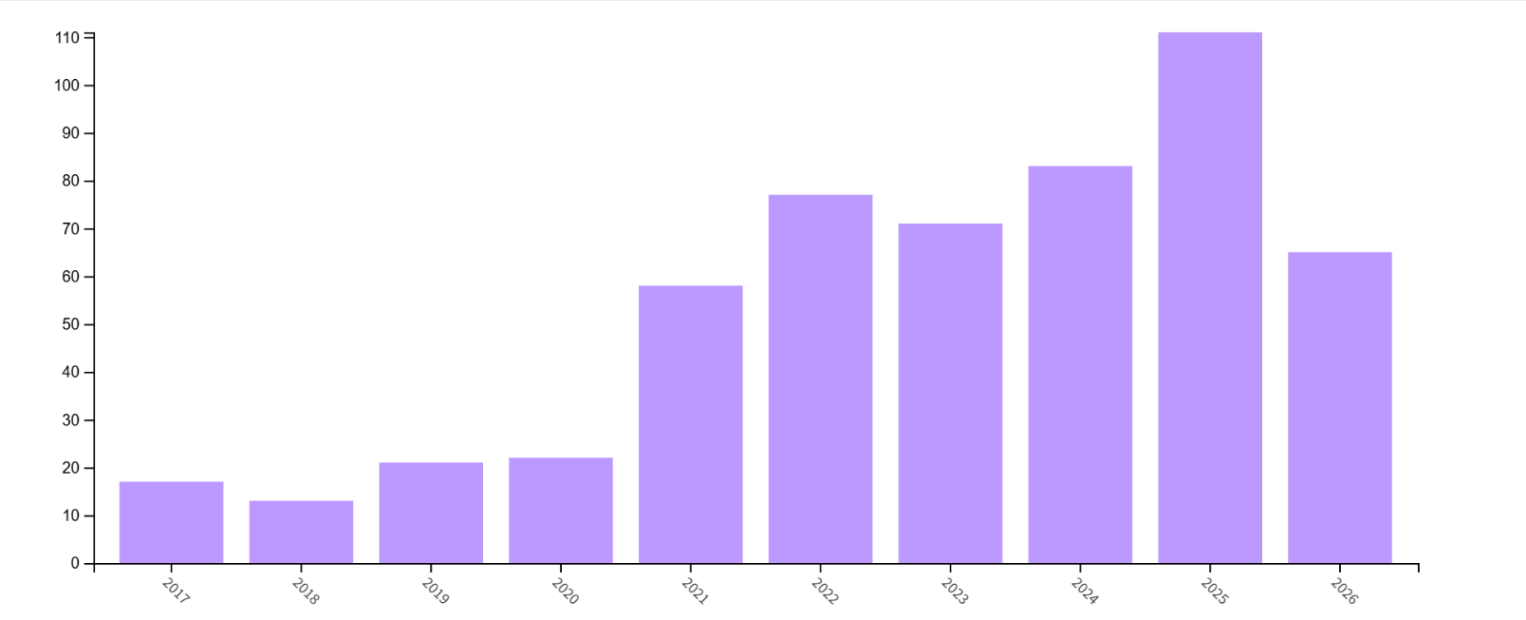
مروری بر مهمترین روش‌های موجود



تهیه شده توسط: https://chatgpt.com/

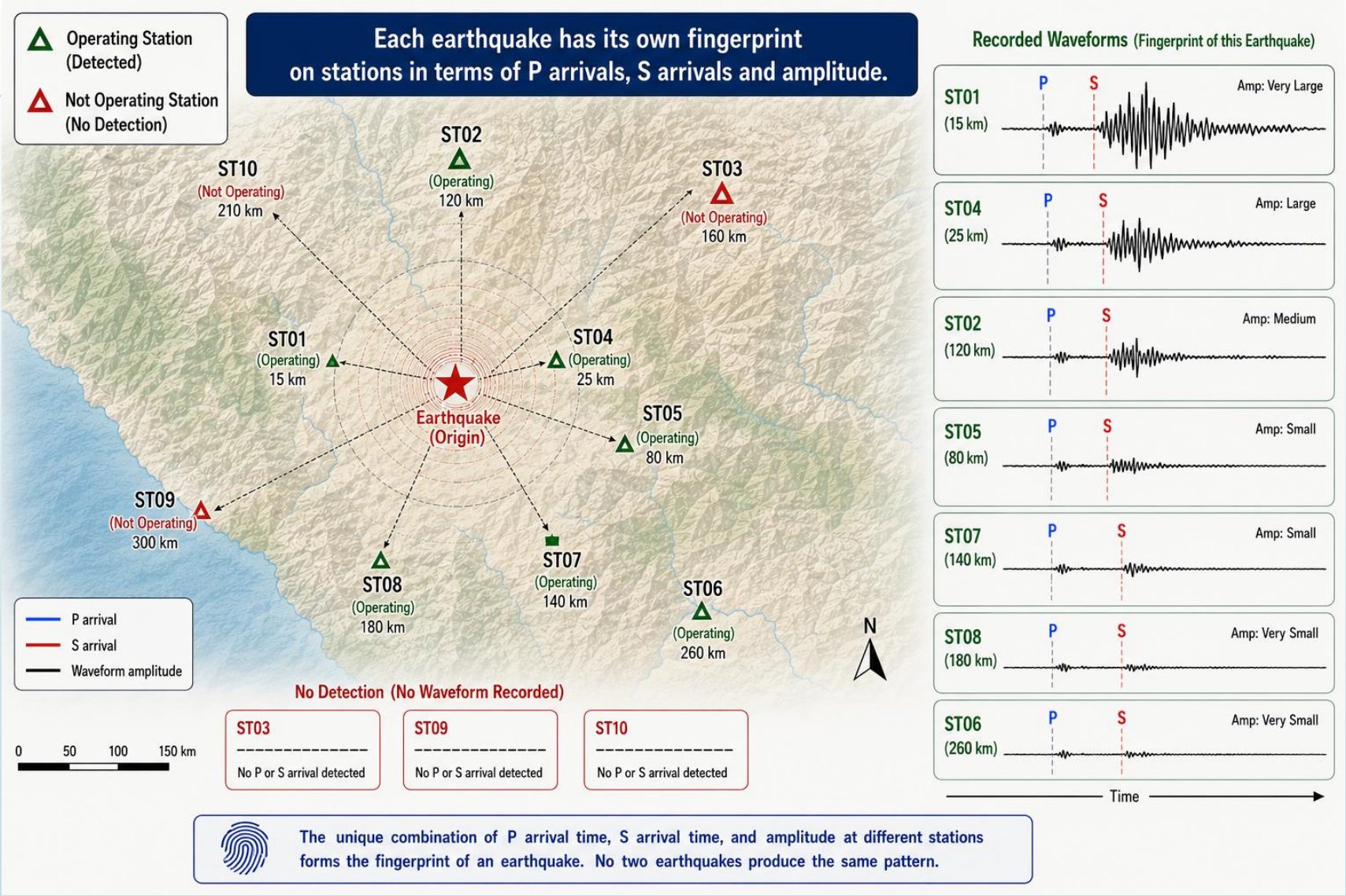
جستجو در پایگاه Webofscience

TS=(("earthquake location" OR "earthquake localization" OR hypocenter* OR "seismic event location" OR "earthquake magnitude" OR" local magnitude" OR "moment magnitude") AND ("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network*" OR CNN OR Transformer* OR LSTM OR "graph neural network*" OR GNN))



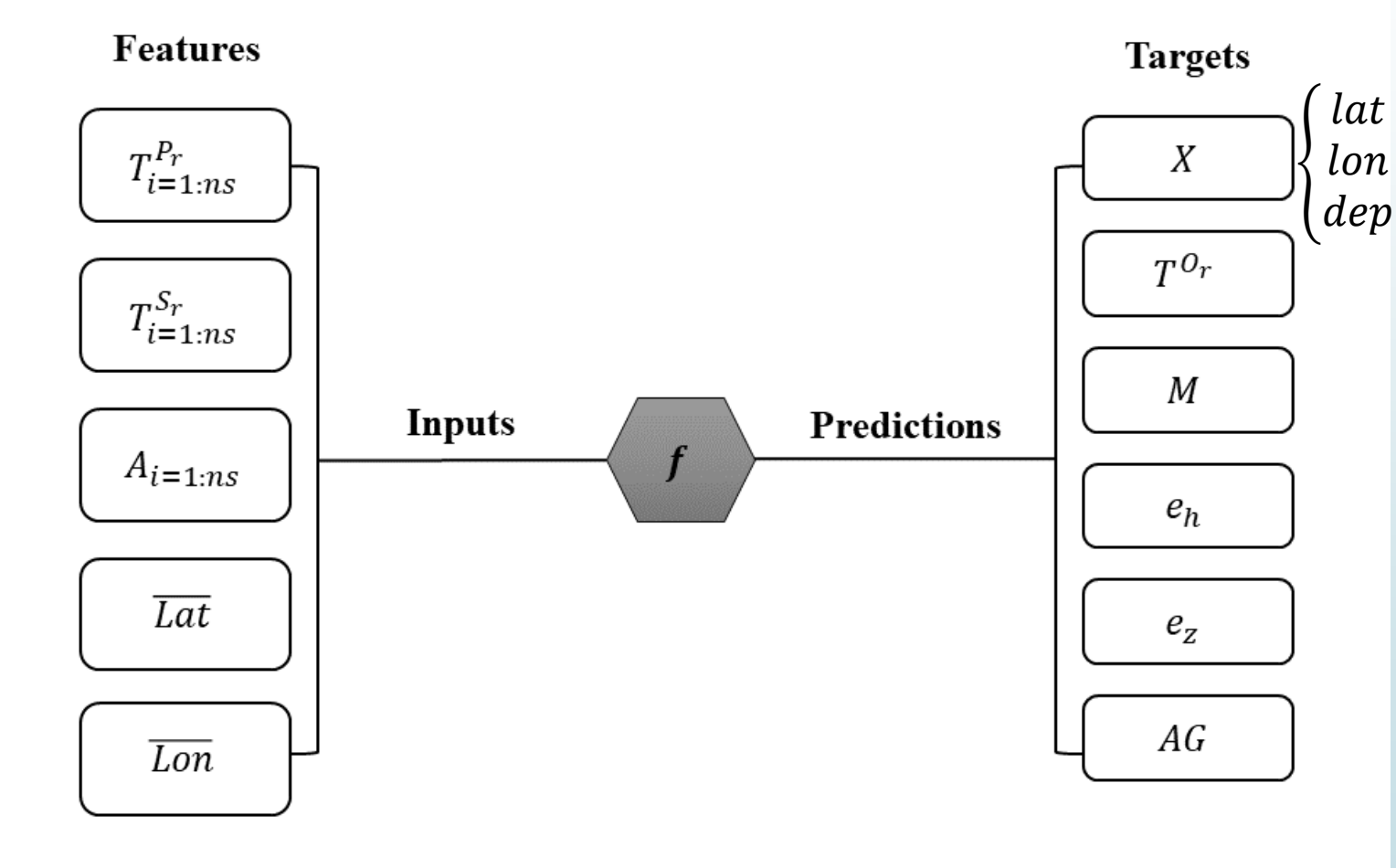
تهیه شده توسط: https://www.webofscience.com/

ایده اصلی: هر زمین‌لرزه = یک رد منحصربفرد در شبکه



تهیه شده توسط: https://chatgpt.com/

پنج بردار ویژگی (ورودی)، هشت بردار هدف (خروجی)

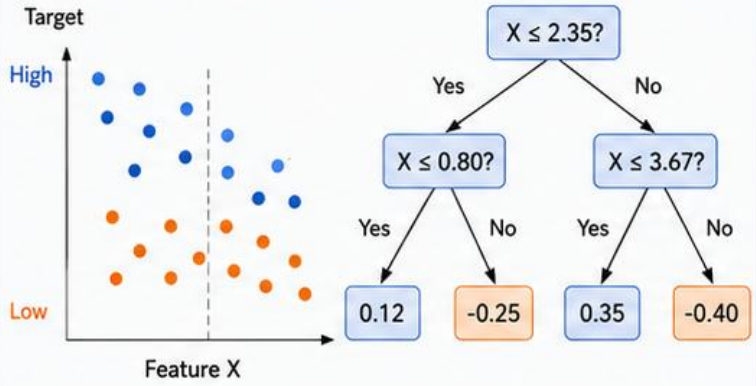


$$[\mathbf{x}, \mathrm{M}, \mathrm{T}^{O_r}, e_h, e_z, AG] = f(\mathrm{T}^{P_r}_{i:1,ns}, \mathrm{T}^{S_r}_{i:1,ns}, A_{i:1,ns}, \overline{Lat}, \overline{Lon})$$

روش مورد استفاده:
Histogram-Based Gradient Boosted Trees (HGBT)

1. Simple Decision Tree

A single tree built on the whole data using exact feature values.



Key idea

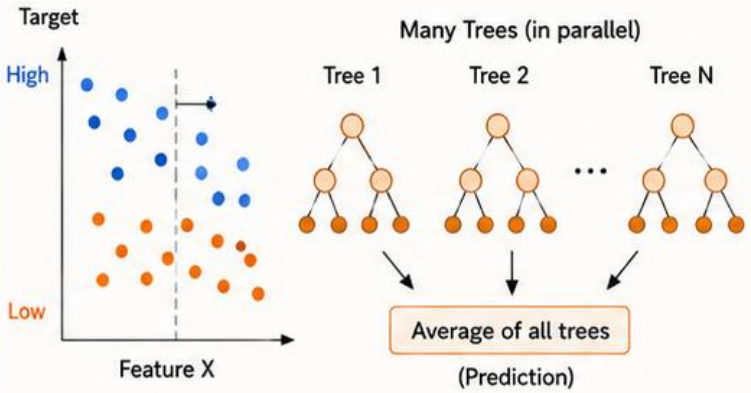
Finds the best splits using exact threshold values.

Pros

- Easy to interpret
- Can capture complex patterns
- Prone to overfitting (high variance)

2. Random Forest

Many trees built on bootstrapped samples with random feature selection.



Key idea

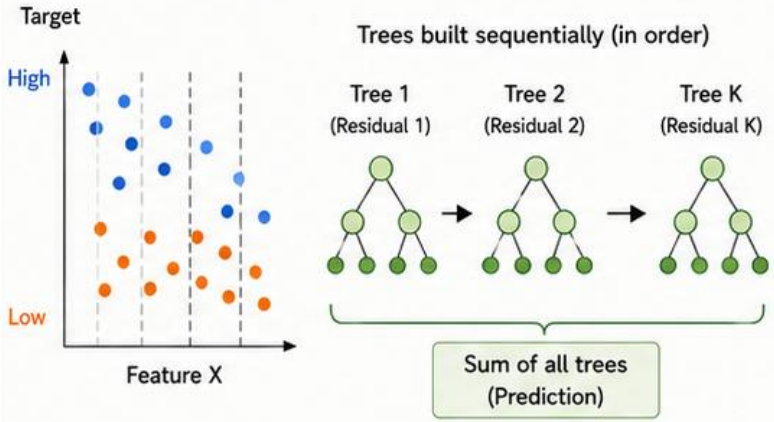
Reduces overfitting by averaging many diverse trees.

Pros

- Lower variance, more robust
- Works well on many datasets
- Less interpretable than a single tree
- Does not model sequential improvement

3. Histogram-Based Gradient Boosted Tree

Trees are built sequentially; each new tree learns from the errors of previous trees using histogram bins.



Key idea

Uses histogram bins to find split candidates and builds trees sequentially to minimize errors (Gradient Boosting).

Pros

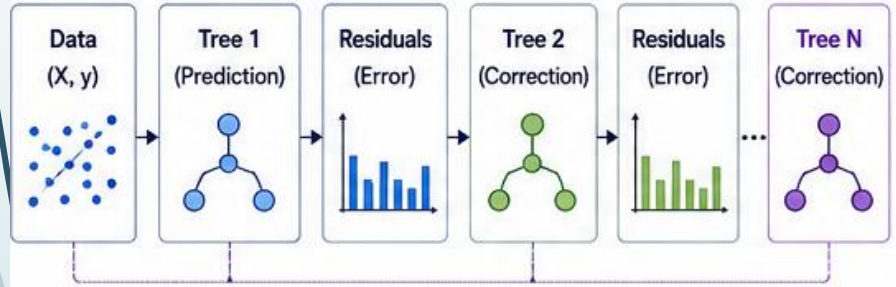
- High predictive performance
- Captures complex patterns
- More robust to overfitting (with regularization)
- Computationally efficient with histogram bins

تهیه شده توسط: https://chatgpt.com/

مزایای روش HGBT:

1. GRADIENT BOOSTING: HOW IT WORKS

Build trees sequentially, where each new tree learns the residual errors of the previous ensemble.



Final Prediction = Tree 1 + Tree 2 + ... + Tree N

Each tree is a weak learner that tries to reduce the remaining error (negative gradient of the loss function). The ensemble becomes stronger as more trees are added.

2. HISTOGRAM BINNING: FASTER SPLIT SEARCH

Traditional Tree Split Search	Histogram-based Split Search
All unique values are considered as potential splits.	Values are grouped into a limited number of bins.
2.11 2.13 2.15 2.18 2.19 ... many more	0–1 1–2 2–3 3–4 ...
Millions of candidate splits	Only ~255 candidate splits
✗ Computationally expensive	✓ Much faster Lower memory usage

Continuous features → fixed bins (e.g., 255 bins per feature)
Greatly reduces training time and memory.

NATIVE MISSING VALUE HANDLING

Is feature missing? → Learn the best direction (Left or Right) during training → Send missing values to the learned direction

✓ No imputation required. Missing values are handled naturally.

3. KEY ADVANTAGES OF HGBR

High Accuracy
Boosting reduces bias and captures complex non-linear relationships.

Fast Training
Histogram binning dramatically speeds up split search.

Memory Efficient
Storing bins instead of raw values lowers memory usage.

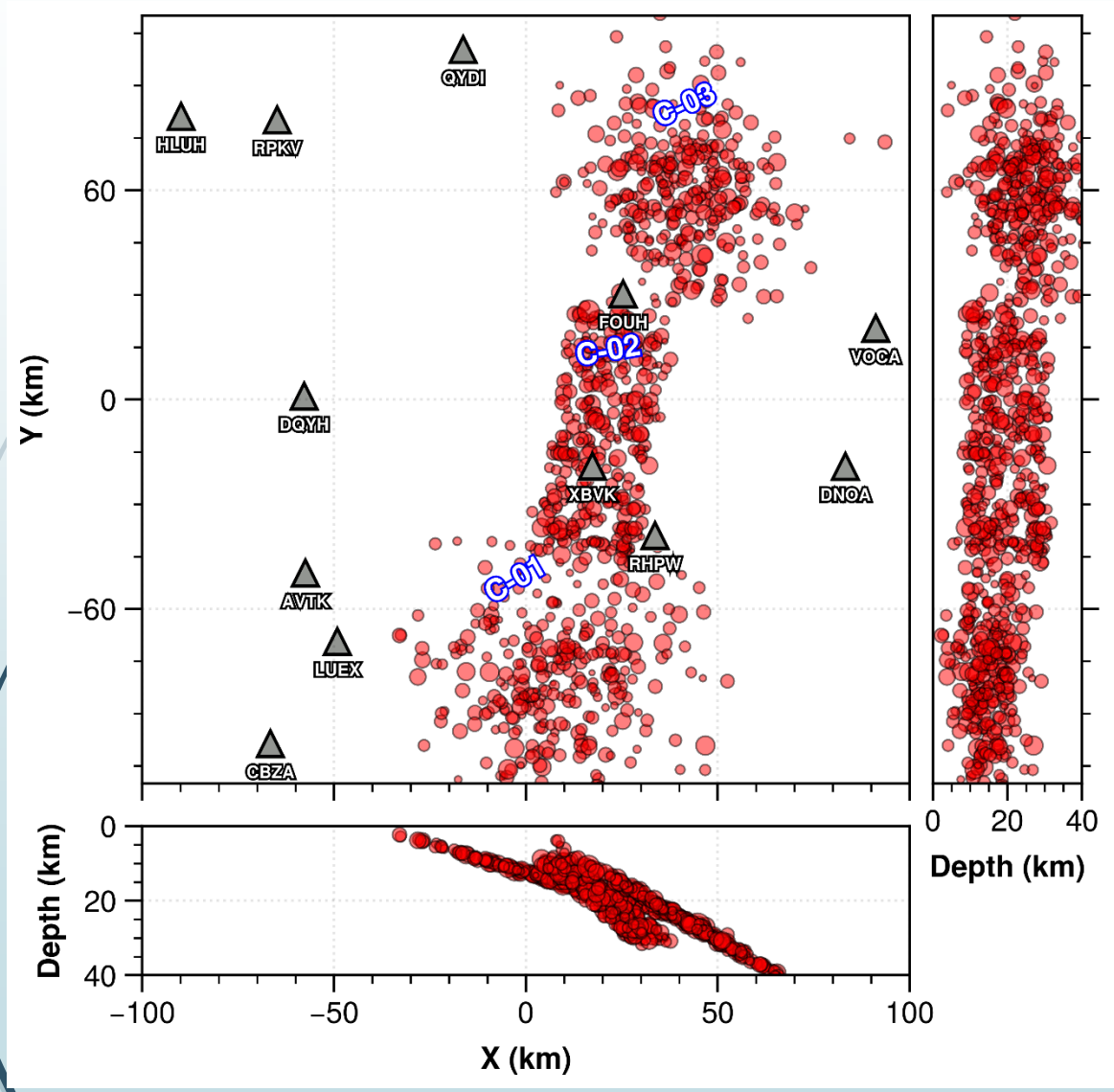
Handles Missing Values
Learns optimal routing for missing values automatically.

Built-in Regularization
Multiple parameters (learning rate, max depth, L2, early stopping, ...) control overfitting.

Scalable
Works efficiently on large datasets with many features.

تهیه شده توسط: https://chatgpt.com/

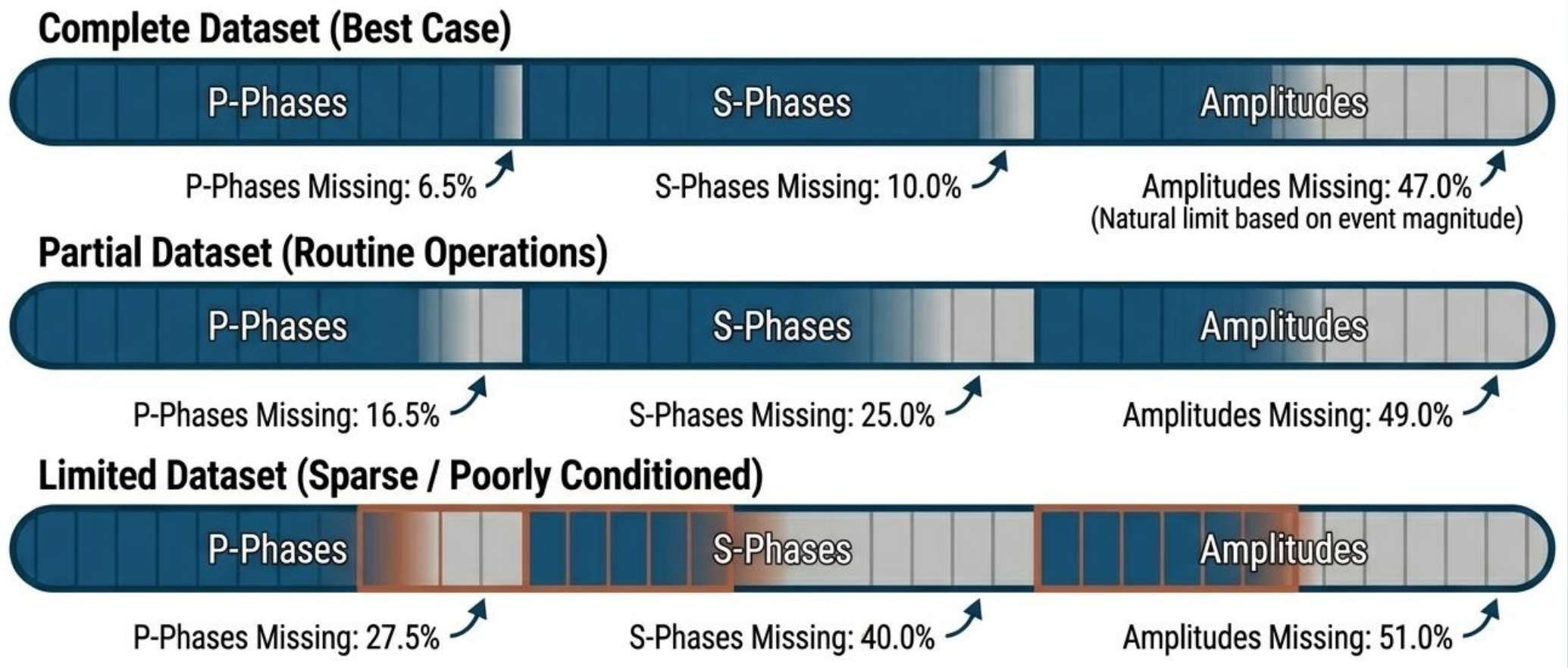
ارزیابی اولیه با کاتالوگ مصنوعی، تولید بانک داده



شبیه‌سازی کاتالوگ:

- منطقه‌ای به شعاع ۱۵۰ کیلومتر، شامل ۱۵ ایستگاه با توزیع تصادفی و فاصله ۱۰ کیلومتر،
- سه گسل با امتداد، شیب و عمق مختلف،
- ۹۰۰ زمین‌لرزه مصنوعی (۳۰۰ رویداد برای هر گسل)، با توزیع نرمال،
- تولید فازها و دامنه بصورت تصادفی (بر اساس فاصله و بزرگا)،

ارزیابی اولیه با کاتالوگ مصنوعی، شبیه‌سازی با کاتالوگ واقعی



تهیه شده توسط: https://notebooklm.google/

ارزیابی اولیه با کاتالوگ مصنوعی، یافتن پارامترهای مدل

۷۵درصد داده آموزشی (آموزش + اعتبارسنجی)، ۲۵ درصد داده آزمون

تنظیم فراپارامترهای مهم:

استفاده از RandomizedSearchCV این امکان را می‌دهد که بجای آزمایش تمام ترکیب‌های ممکن (GridSearchCV) تعدادی ترکیب تصادفی از فضای پارامترها بررسی شود.

Dataset	learning_rate	max_depth	max_iter	L2_reg	R^2
Complete	0.09	62	107	25.4	0.86
Partial	0.09	62	107	25.4	0.77
Limited	0.06	93	84	0.05	0.64
Distribution	$U(0.05, 0.15)$	$U(2, 100)$	$U(50, 150)$	$U(0.01, 100)$	-

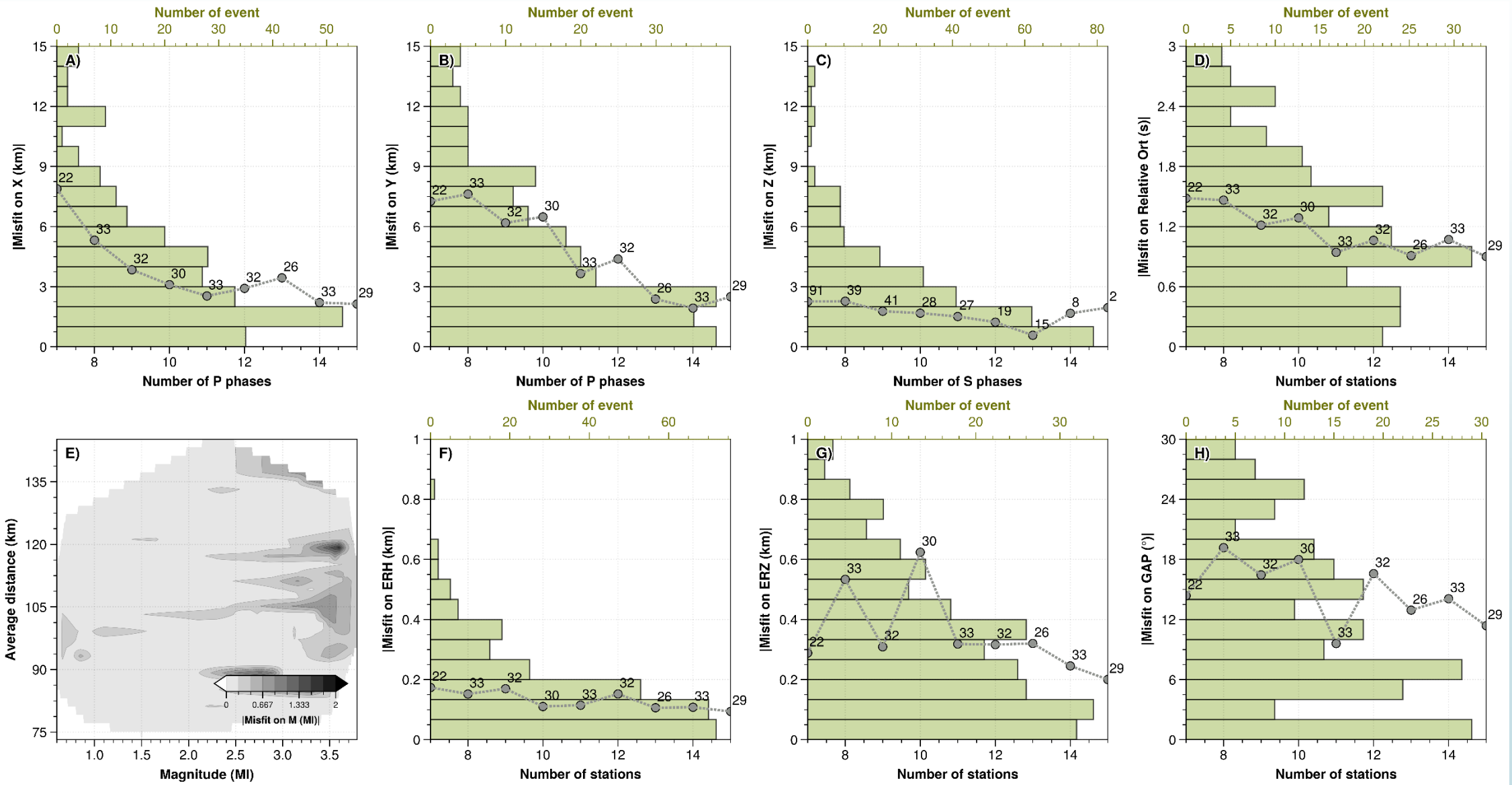
میزان تاثیر هر درخت جدید در اصلاح خطا

تعداد مراحل Boosting

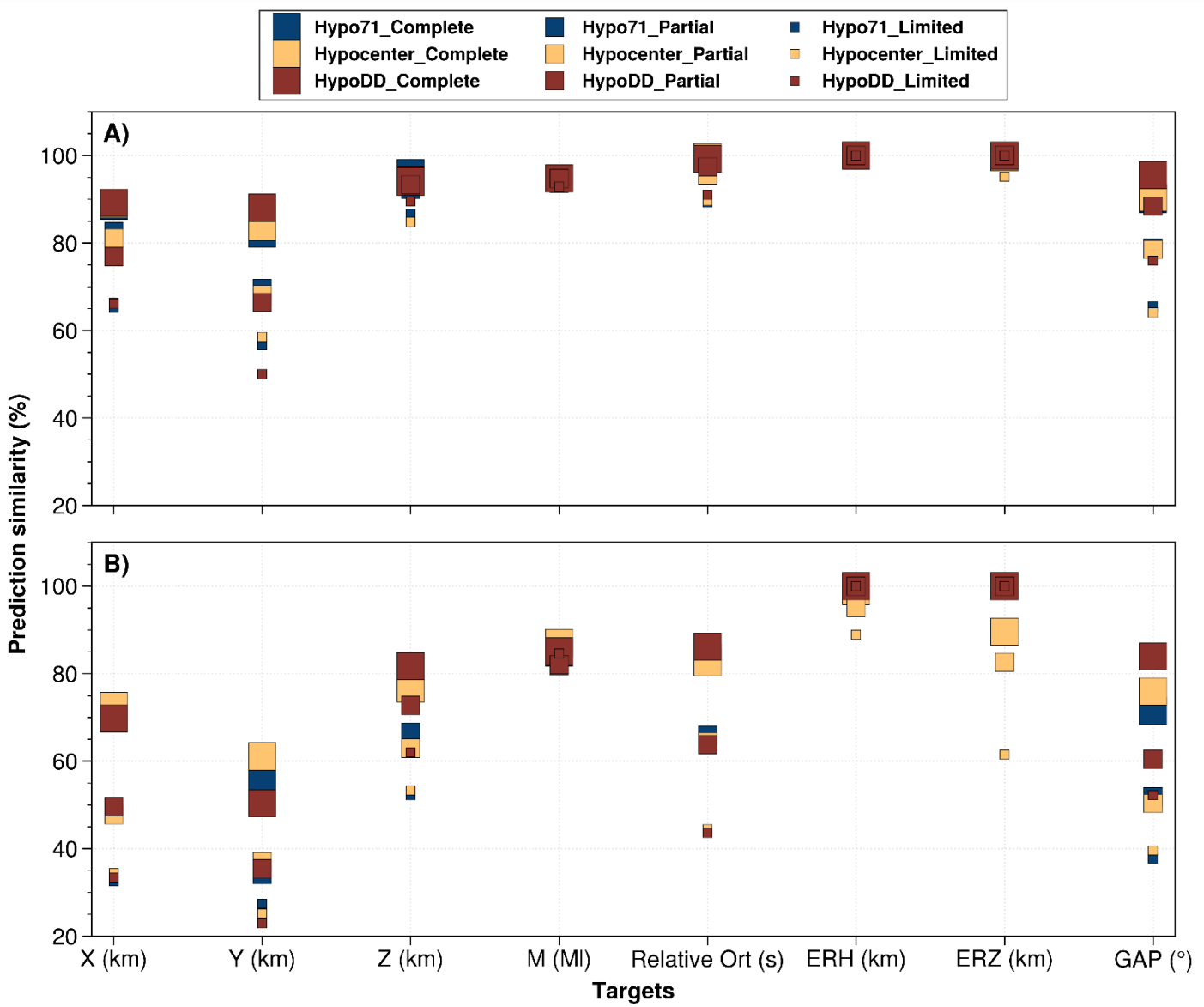
کنترل پیچیدگی مدل

میزان جریمه برای کاهش بیش‌برازشی

ارزیابی اولیه با کاتالوگ مصنوعی، دقت قابل قبول حتی با داده کم



ارزیابی اولیه با کاتالوگ مصنوعی، مقایسه سه مدل مختلف



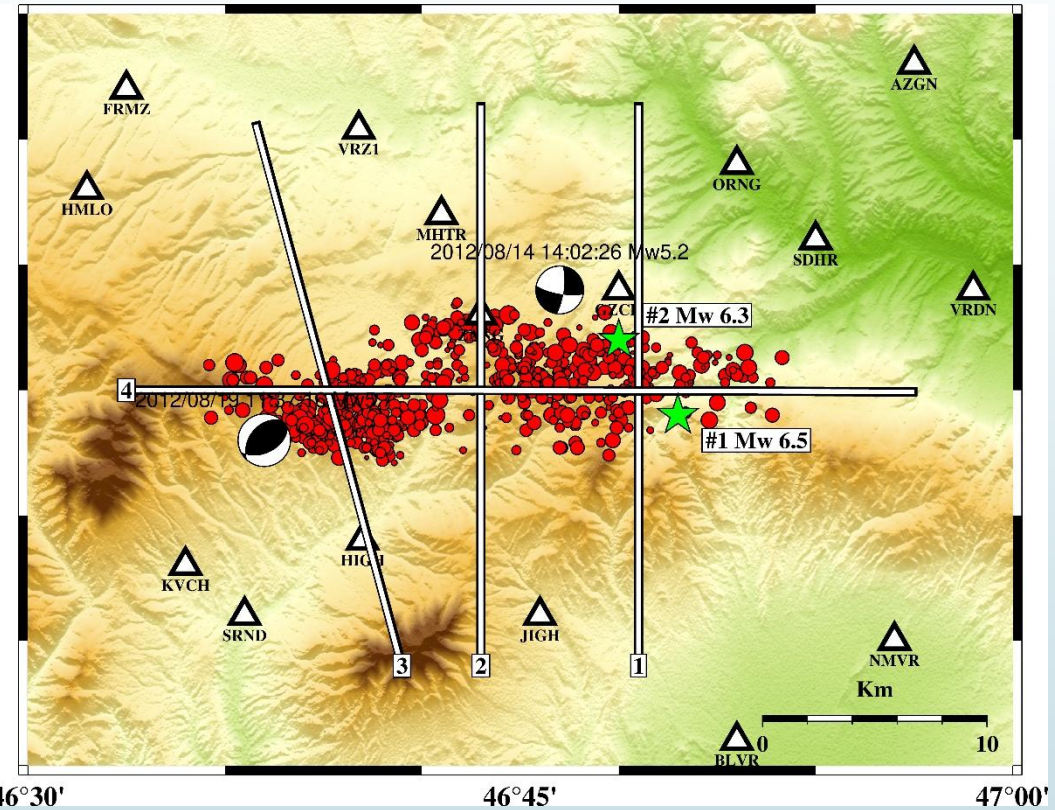
Condition	A	B
X(km)	<2.0	<5.0
Y(km)	<2.0	<5.0
Z(km)	<2.0	<5.0
OT(s)	<1.0	<3.0
M(km)	<0.2	<0.5
Gap(km)	<10	<20
ERH(km)	<1.0	<3.0
ERZ(km)	<1.0	<3.0

ارزیابی اولیه با کاتالوگ مصنوعی، مقایسه سرعت مکان‌یابی با سایر برنامه‌ها

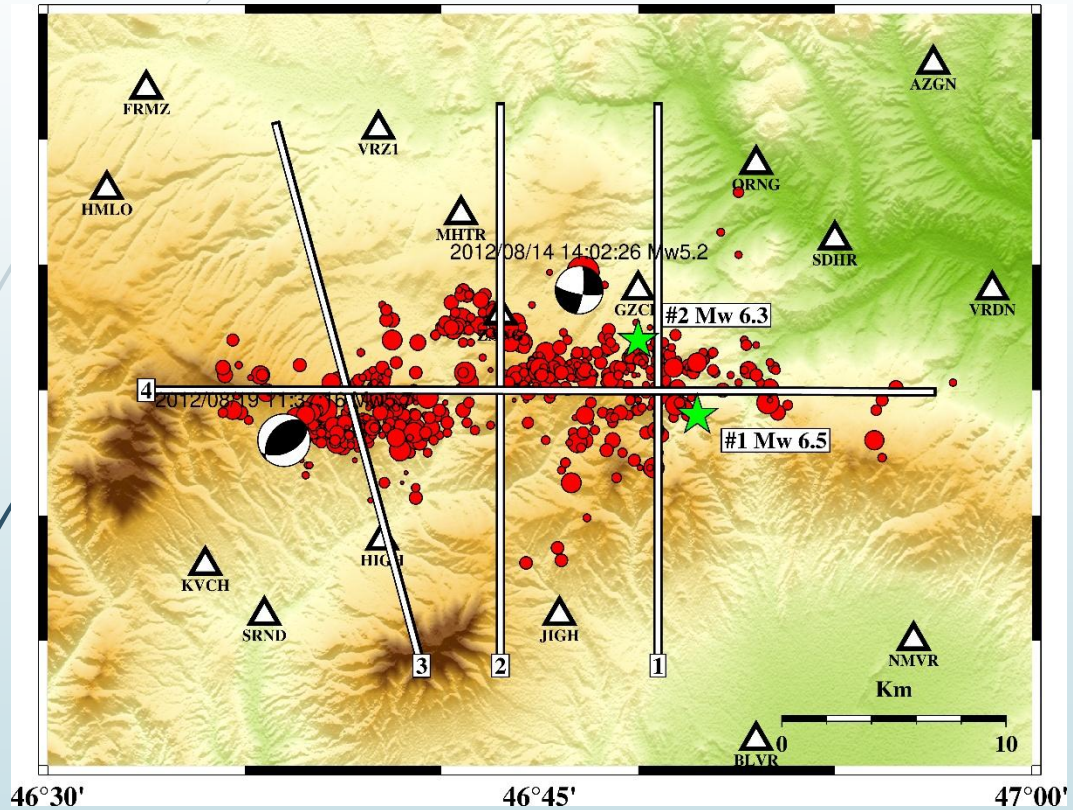
تهیه شده توسط: https://notebooklm.google/

ارزیابی بروی کاتالوگ واقعی، پس‌لرزه‌های اهر–ورزقان ۲۰۱۲

کاتالوگ بدست آمده با HGBT

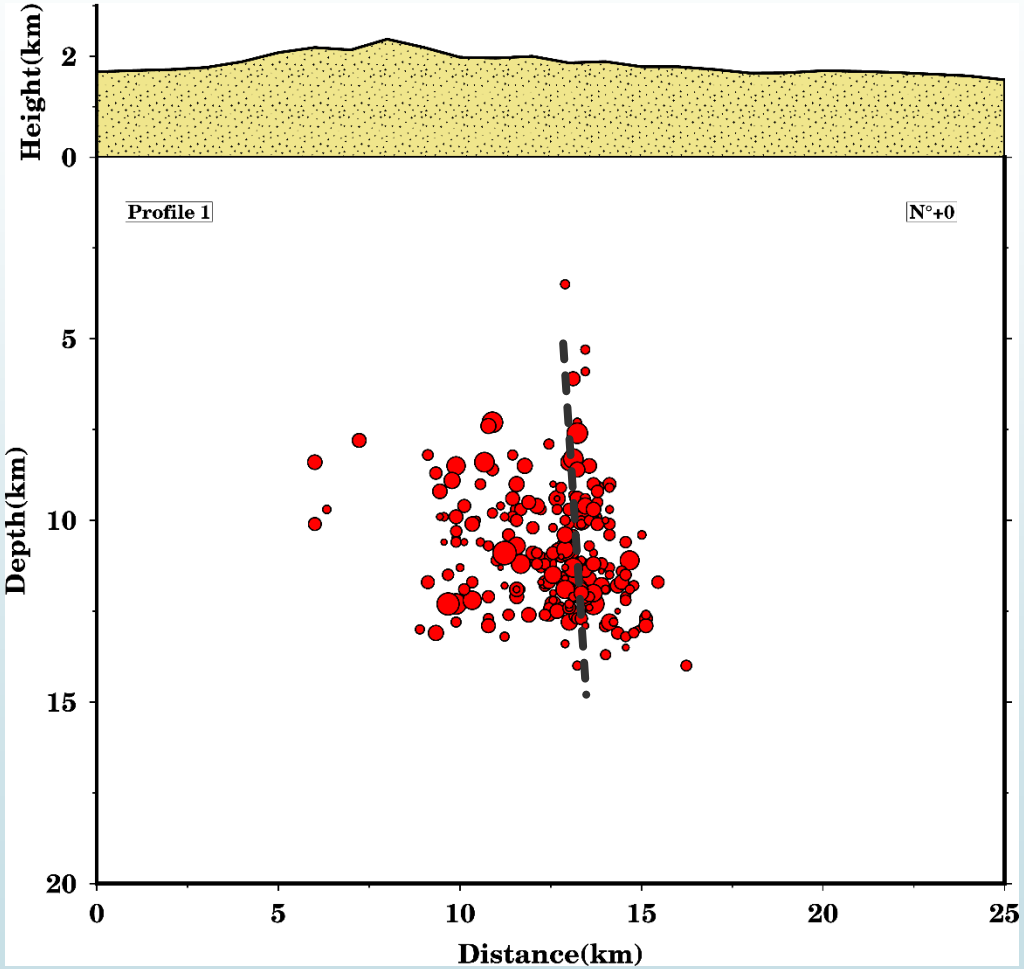


کاتالوگ مومنی و همکاران (۲۰۱۸)

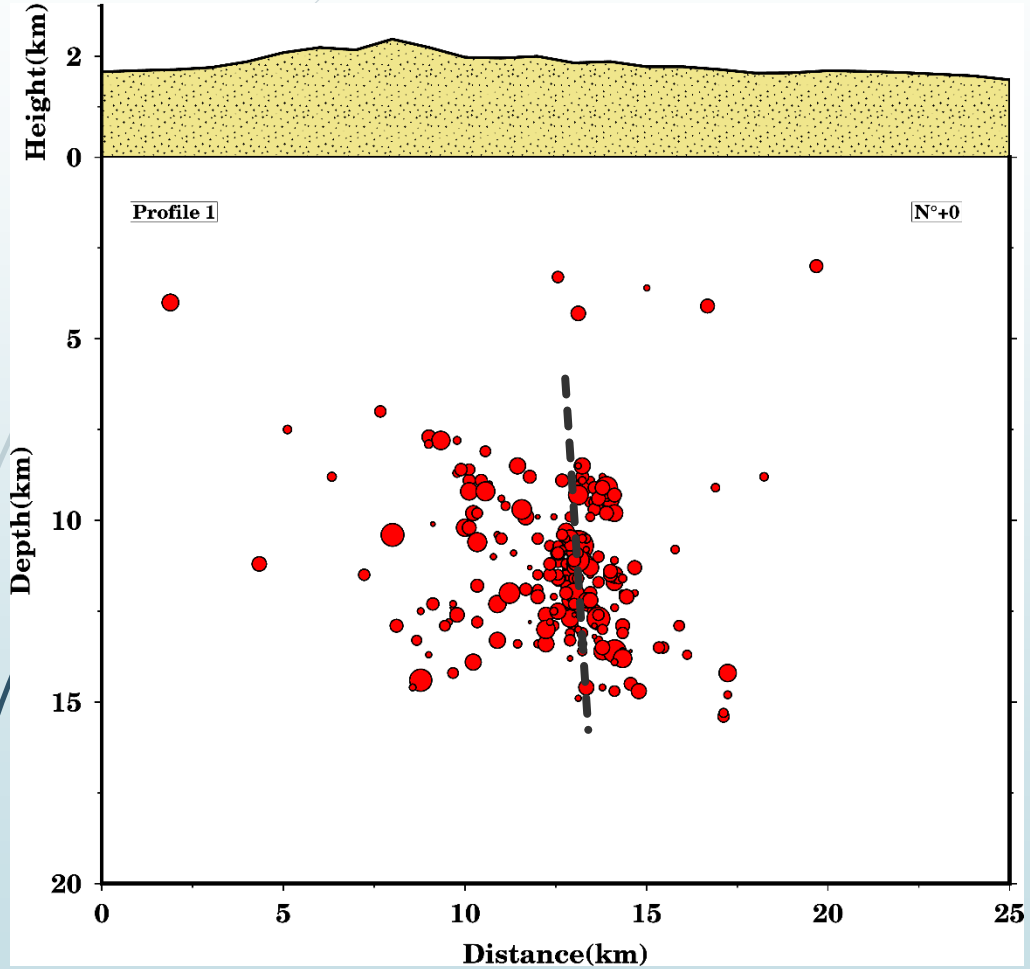


از مجموع ۲۶۶۲ زمین‌لرزه به ثبت رسیده در شبکه موقت، ۷۵ درصد (۱۹۹۶) به عنوان داده آموزشی و ۲۵درصد (۶۶۶) به عنوان داده آزمون انتخاب گردید و مدل HGBT محاسبه شد.

ارزیابی بروی کاتالوگ واقعی، پس‌لرزه‌های اهر–ورزقان ۲۰۱۲

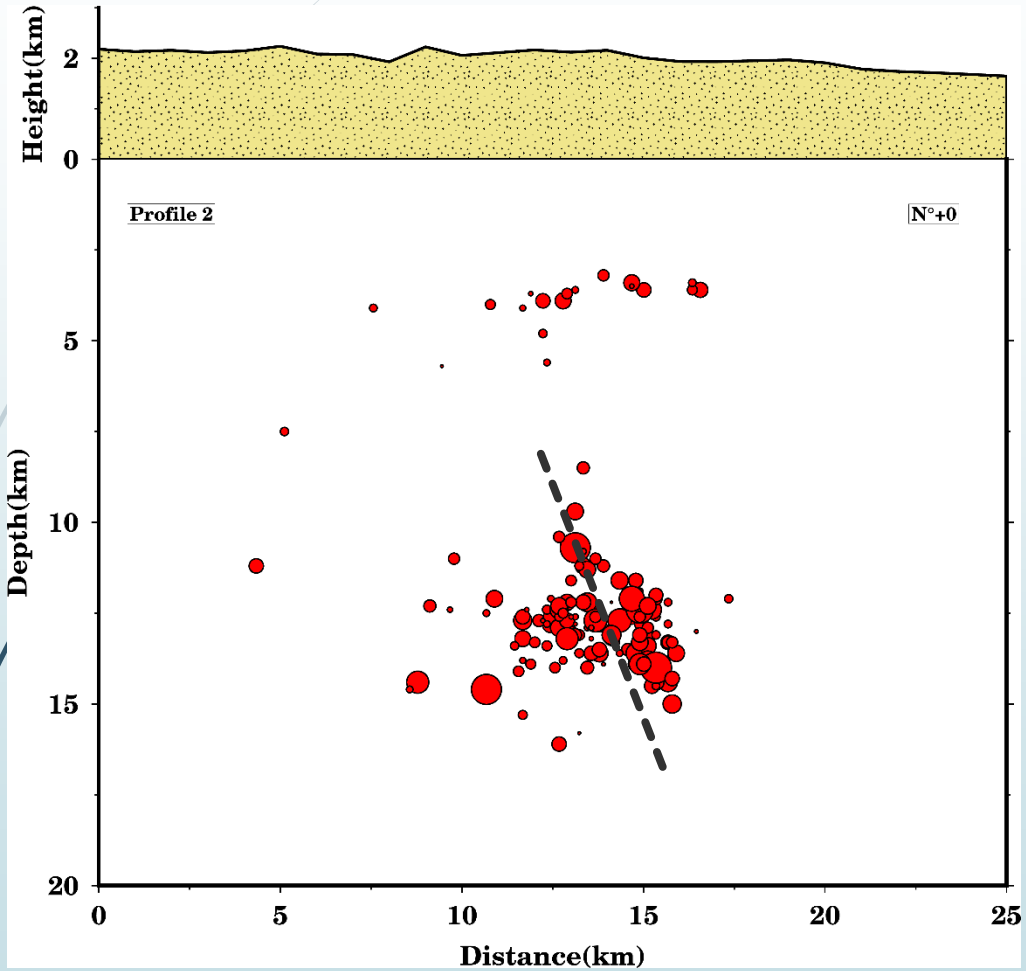


کاتالوگ بدست آمده با HGBT

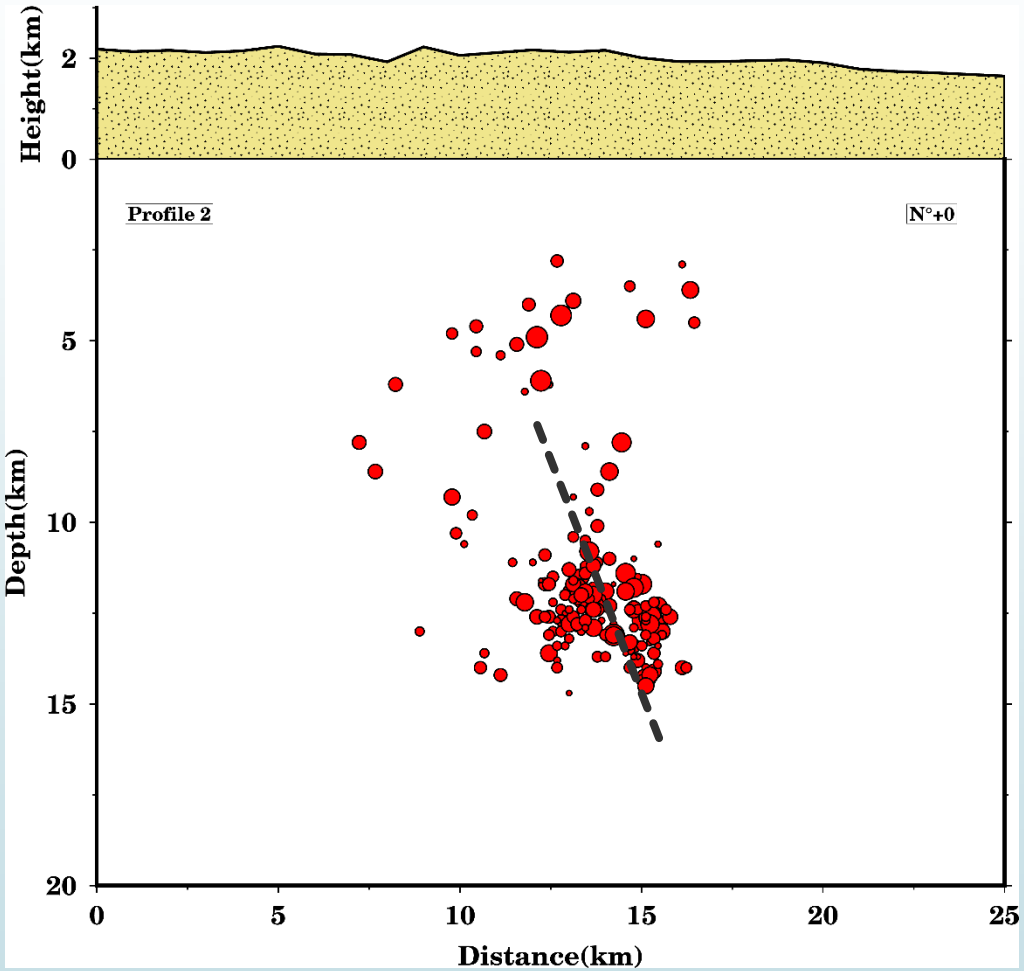


کاتالوگ مومنی و همکاران (۲۰۱۸)

ارزیابی بروی کاتالوگ واقعی، پس‌لرزه‌های اهر–ورزقان ۲۰۱۲

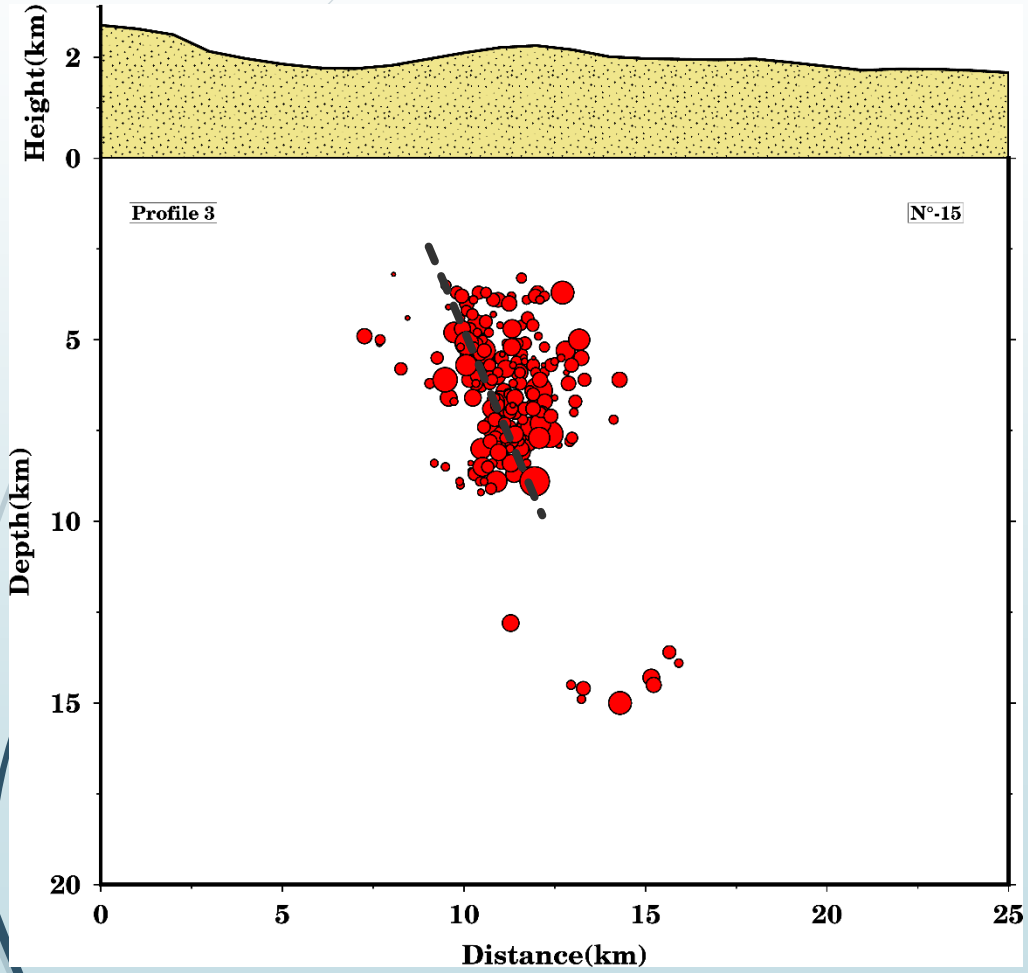


کاتالوگ مومنی و همکاران (۲۰۱۸)

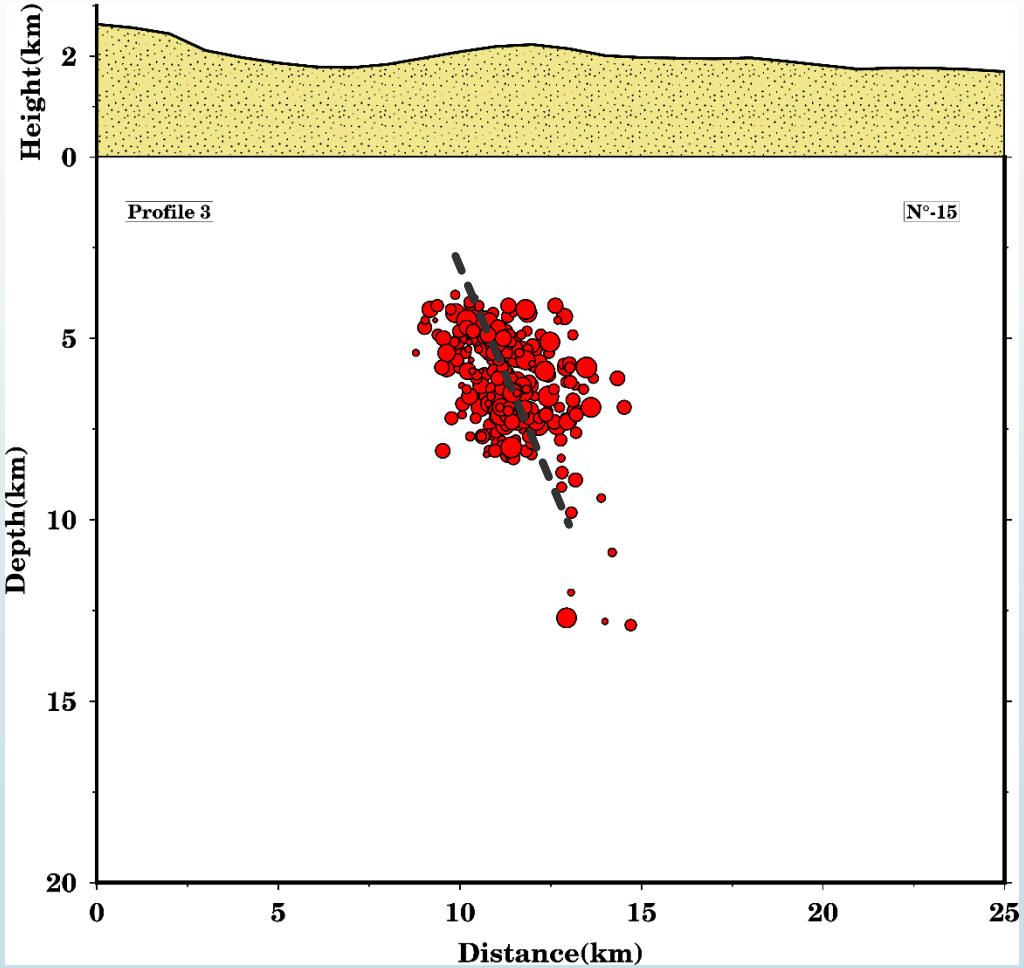


کاتالوگ بدست آمده با HGBT

ارزیابی بروی کاتالوگ واقعی، پس‌لرزه‌های اهر–ورزقان ۲۰۱۲

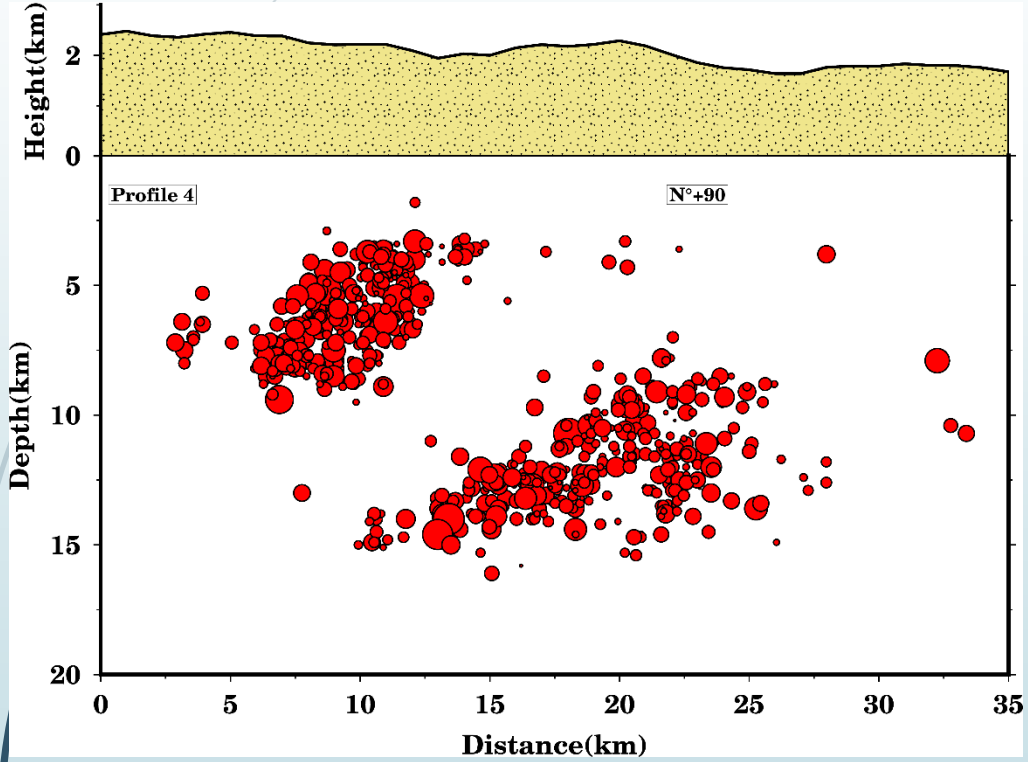


کاتالوگ مومنی و همکاران (۲۰۱۸)

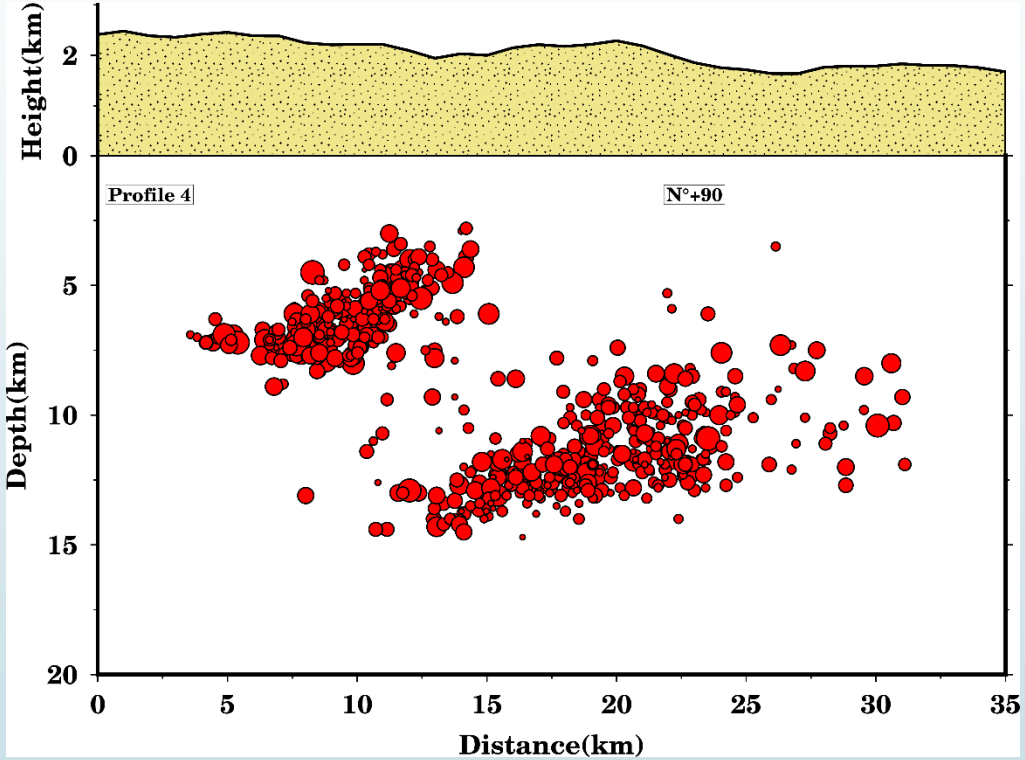


کاتالوگ بدست آمده با HGBT

ارزیابی بروی کاتالوگ واقعی، پس‌لرزه‌های اهر–ورزقان ۲۰۱۲



کاتالوگ مومنی و همکاران (۲۰۱۸)



کاتالوگ بدست آمده با HGBT

نتایج

- روش ارائه شده دارای دقت مشابه اما سرعت عمل بمراتب بالاتری نسبت به برنامه‌های مکان‌یابی کلاسیک است،
- قابلیت مدیریت داده‌های گمشده در الگوریتم HGBT، سبب می‌گردد تا بتوان از این روش در غیاب داده‌های دامنه تا ۵۰ دردصد و فازهای S تا ٤٠ درصد بدون افت قابل توجه در دقت مکان‌یابی، استفاده کرد،
- مدل نهایی پس از آموزش، برای مکان‌یابی نیازی به مدل سرعتی منطقه نخواهد داشت و عدم قطعیت پایین در تخمین عمق و ساختار هندسه گسل را ارائه می‌دهد،
- روش ارائه شده می‌تواند دقتی معادل برنامه HypoDD و سرعت Hypo71 داشته باشد و برای هر نوع برنامه مکان‌یابی دیگر قابل تعمیم است.

محدودیت‌ها

- روش ارائه شده صرفاً برای استفاده در شبکه‌های لرزه‌نگاری دائم، طراحی شده است، اما در صورت بکارگیری الگوریتم‌های مبتنی بر گراف (GNN)، می‌توان این وابستگی را از میان برد،
- عملکرد این روش برای شبکه‌های با تعداد بسیار کم ایستگاه (کمتر از ۵ ایستگاه) روشن نیست و نیاز به بررسی بیشتر دارد،
- با توجه به اینکه بطور طبیعی توزیع آماری بزرگای زمین‌لرزه‌های به توزیع گاما نزدیک است، احتمال وقوع زمین‌لرزه‌ها با بزرگای بالاتر بیشتر است. این موضوع می‌تواند عملکرد مدل را به منظور تخمین بزرگای چنین زمین‌لرزه‌هایی کاهش دهد. راهکار پیشنهادی برای این منظور، استفاده از داده‌های مصنوعی برای Data augmentation است تا تعادل کافی میان بزرگاهای مختلف ایجاد گردد.

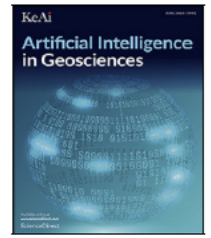
«با تشکر از توجه شما»

Artificial Intelligence in Geosciences 6 (2025) 100149

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

journal homepage: www.keaipublishing.com/en/journals/artificial-intelligence-in-geosciences



Original research articles

Earthquake location and magnitude estimation using seismic arrival times pattern and gradient boosted decision trees

Saeed SoltaniMoghadam *, Anooshiravan Ansari, Leila Etemadsaeed, Mohammad Tatar, Meysam Mahmoodabadi

International Institute of Earthquake Engineering and Seismology (IIEES), No. 21, Arghavan St., North Dibajee, Farmanieh, Tehran, 19537-14453, Iran

ARTICLE INFO

Keywords:
Earthquake location
Machine learning
Gradient-boosted-decision-trees
Synthetic and real data

ABSTRACT

We present a machine learning approach for earthquake location and magnitude estimation based on seismic arrival time patterns, using Histogram-Based Gradient Boosting for its high accuracy and computational efficiency. The model is first evaluated using a synthetic earthquake bulletin that simulates realistic network geometry, station-event distributions, and incorporates a 3D velocity model for accurate travel-time computation. Input features include P and S arrival times and amplitudes, while targets consist of location, origin time, magnitude, and uncertainty measures (horizontal and depth errors, azimuthal gap). Model performance is evaluated using R^2, Mean Absolute Error (MAE), and Median Absolute Error (MEDAE), demonstrating high accuracy across datasets with varying levels of completeness. Finally, we validate the model using real-world data from the Ahar-Varzaghan 2012 aftershock sequence in NW Iran. The model accurately recovers key spatial patterns of seismicity despite significant missing data, and the results align with previous high-resolution studies. These findings confirm that the proposed method generalizes well beyond synthetic settings and offers