



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
IL FUTURO DELLA TERRA, CAMBIAMENTI CLIMATICI E SFIDE
SOCIALI

Ciclo 38

Settore Concorsuale: 04/A4 - GEOFISICA

Settore Scientifico Disciplinare: GEO/10 - GEOFISICA DELLA TERRA SOLIDA

DEEP LEARNING-BASED EARTHQUAKE DETECTION FOR TECTONIC
INTERPRETATION AND HAZARD IMPLICATIONS

Presentata da: Elisa Caredda

Coordinatore Dottorato

Silvana Di Sabatino

Supervisore

Andrea Morelli

Co-supervisore

Maria Elina Belardinelli

Esame finale anno 2026

Abstract

This dissertation explores the detection, location, and seismotectonic interpretation of microseismicity within the Val d’Agri region of Southern Italy. This region is characterized by both active tectonics and anthropogenic activities, such as hydrocarbon extraction and reservoir impoundment. Two deep-learning-based approaches for earthquake detection and phase identification are evaluated and compared, substantially enhancing the sensitivity and comprehensiveness of seismic catalogs compared to conventional STA/LTA methodologies. Deep-learning-based approaches enable the identification of nearly twice as many events while successfully recovering the majority of previously identified earthquakes. Additionally, they provide fully automated workflows that are well-suited for near-real-time monitoring. Building on these enhanced catalogs, I then apply waveform-based clustering and high-precision relocation to reveal the geometry of active fault structures. The clustering approach groups similar events, likely originating from the same fault segments, while the relocation highlights planar features consistent with fault planes. Moment tensor inversion of the largest events ($M > 2$) further constrains the faulting mechanisms, indicating predominantly SW-dipping normal faults aligned with the regional extensional regime. A thorough examination of a representative seismic sequence is undertaken, analyzing its spatial-temporal progression and identifying potential hazards.

This comprehensive work presents a systematic integration of modern machine learning and advanced seismological techniques. By employing these methodologies, I have achieved significant improvements in earthquake detection capabilities. Furthermore, this research contributes to a more nuanced comprehension of active deformation and seismic hazards within an intricate tectonic environment.

Contents

Abstract	i
1 Introduction	1
2 Geological and Seismotectonic context of the Val d’Agri basin	4
2.1 Geodynamic context of the Southern Apennines	4
2.2 Local setting: the Val d’Agri basin	6
2.2.1 Geological setting	6
2.2.2 The Val d’Agri Oilfield	9
2.2.3 Natural and induced seismicity	10
2.2.4 The local seismic networks	11
3 Principles and applications of Deep Learning for seismic event detection	13
3.1 Introduction	13
3.2 Deep Learning and Artificial Neural Networks	14
3.2.1 Convolutional Neural Networks	17
3.2.2 Overfitting and regularization strategies	19

3.2.3	Performance evaluation metrics	20
3.3	Deep Learning in seismic detection methods	21
3.3.1	From traditional methods to deep learning	21
3.3.2	Examples of Deep learning in seismic event detection	23
3.3.3	CNN tested: PhaseNet	24
4	Implementation of DL-based workflows for the detection of micro-seismicity	27
4.1	Introduction	27
4.2	Data	28
4.3	Methods	28
4.3.1	PRN workflow	28
4.3.2	QS workflow	38
4.4	Results	50
4.5	Discussion	57
4.5.1	Comparison of the workflows' performance	57
4.5.2	Noise and performance of the detection	58
4.5.3	Pertusillo water level and seismicity rate	62
4.6	Conclusions	65
5	Magnitude of completeness and b-value	68
5.1	Introduction	68
5.2	Analysis	71
5.3	Discussion	73

6	From moment tensors and waveform similarity to hazard implications	78
6.1	Introduction	78
6.2	Data and Methods	81
6.2.1	Waveform similarity-based clustering	81
6.2.2	Double-difference relocation	82
6.2.3	Moment tensor inversion	85
6.3	Results	86
6.4	Discussion	91
6.4.1	Waveform similarity and focal mechanisms interpretation .	91
6.4.2	July 2023 seismic sequence	94
6.4.3	Hazard Scenarios	94
6.5	Conclusions	97
7	Extending the DL-based detection workflow to other Italian Regions	100
7.1	Introduction	100
7.2	Sardinian microplate	101
7.2.1	Introduction	101
7.2.2	Data and Methods	102
7.2.3	Results	104
7.2.4	Discussion and Conclusions	106
7.3	The Ivrea Verbano Zone	107
7.3.1	Introduction	107

7.3.2	Data and Methods	108
7.3.3	Results	108
7.3.4	Discussion and Conclusions	109
8	Conclusions	111
A	Supplementary Information	115
	Acknowledgments	140

Chapter 1

Introduction

Monitoring microseismicity is essential for understanding fault mechanics and mitigating seismic hazard, particularly in regions affected by industrial activities such as hydrocarbon extraction, petroleum wastewater re-injection, and artificial reservoir operations. High-resolution seismic observations allow for the detection of low-magnitude earthquakes that, while generally not damaging, offer critical insights into the evolution of subsurface processes and the potential for larger events (Keranen & Weingarten, 2018 [1]). Despite their scientific value, the identification and location of microseismic events are often challenged by the low signal-to-noise ratio of the recordings and the large data volumes generated by dense seismic networks.

The Val d’Agri area, located in the southern Apennines of Italy, provides an excellent context for developing and testing advanced methodologies to achieve a more accurate characterization of tectonic and potentially induced microseismicity. This region is characterized by a seismically active extensional regime and hosts both the largest onshore oil field in Europe and the Pertusillo artificial reservoir, both of which have been associated with cases of induced or triggered seismicity. The complexity of this tectonic and industrial setting, combined with the region’s persistent low-magnitude seismicity, calls for advanced monitoring and analysis tools capable of capturing the spatiotemporal evolution of seismicity with high resolution.

In recent years, advances in machine learning have led to significant improvements in automatic seismic data analysis. Deep neural networks, such as PhaseNet (Zhu & Beroza, 2019 [2]), have demonstrated high reliability and scalability (e.g. Dai et al., 2023 [3]) in the detection and picking of seismic arrivals in continuous waveforms, even under low signal conditions. These methods have opened new possibilities for real-time and homogeneous microseismic monitoring, significantly enhancing the detection threshold and reducing the need for manual intervention.

The primary goal of this thesis is to investigate the potential of deep learning techniques for improving the detection and characterization of seismicity in the Val d’Agri region, an area of high tectonic complexity and anthropogenic influence in southern Italy. Two different machine learning-based workflows are implemented and tested: both centered on the use of PhaseNet, a convolutional neural network designed for seismic phase picking, but using different association algorithms. These workflows are applied to continuous waveform data to generate new seismic catalogs, which are then compared in detail with each other and with the official catalog maintained by the *Centro di Monitoraggio delle attività di Sottosuolo* (CMS) (Morelli et al., 2025 [4]) of the *Istituto Nazionale di Geofisica e Vulcanologia* (INGV). The comparison aims to assess the reliability and completeness of each catalog, with particular attention to magnitude thresholds, event rates, and the spatial distribution of seismicity.

After establishing the quality of the newly generated catalogs, the study focuses on the internal structure of the seismicity. One of the machine learning-based catalogs is selected for further analysis using waveform similarity clustering. This approach makes it possible to identify groups of events with similar rupture characteristics, offering insights into fault structures and potential repeating earthquakes.

To deepen the seismotectonic interpretation, the thesis also includes the inversion for moment tensors for the largest events included in the catalog. These solutions are used to infer faulting styles and stress orientations, contributing to a more complete understanding of the active deformation processes in the region.

Finally, the implications of the observed seismicity for ground-motion hazard are explored. A simplified simulation of peak ground velocity (PGV) is performed using synthetic waveforms and source parameters derived from a recent seismic sequence,

providing a first-order estimate of the expected ground shaking and demonstrating the practical utility of enhanced seismic catalogs for hazard assessment.

The thesis is organized as follows.

Chapter 2 describes the geological and seismotectonic framework of the Val d’Agri basin.

Chapter 3 introduces the main concepts of Deep Learning, Neural Networks, and their application to seismic detection.

Chapter 4 presents the implementation and application of two Neural Network-based detection workflows in the Val d’Agri area and compares their performance.

Chapters 5 and 6 build upon the enhanced catalogs obtained in Chapter 4. Chapter 5 focuses on catalog completeness and statistical properties of seismicity, while Chapter 6 provides a detailed physical interpretation of the seismicity through waveform similarity-based clustering, moment tensor inversion, and exploration of potential hazard scenarios.

Chapter 7 extends the implementation of neural-network-based earthquake detection workflows to other low-seismicity regions, demonstrating their robustness and adaptability. These preliminary results highlight the potential of the proposed approach to be generalized beyond the Val d’Agri case study, illustrating the way for its application to diverse geological contexts.

By integrating machine learning, high-precision relocation, clustering, and source mechanism analysis, this thesis contributes to a better understanding of the active deformation in the Val d’Agri region. The results are relevant for both scientific and operational purposes, enhancing seismic hazard assessment in a region of significant societal and industrial relevance.

Chapter 2

Geological and Seismotectonic context of the Val d’Agri basin

2.1 Geodynamic context of the Southern Apennines

The Southern Apennines are part of a Neogene-to-Quaternary orogenic system that developed as a result of the complex interaction between the African and Eurasian plates. This sector of the Apennine chain is currently affected by a regime of active extension, which contrasts with the compressional tectonics that dominated the earlier stages of its evolution (Patacca & Scandone, 2007 [5]). The extensional deformation is interpreted as a consequence of slab retreat and rollback of the subducting Ionian lithosphere toward the southeast, which has driven the progressive eastward migration of the orogenic front and the opening of the Tyrrhenian back-arc basin (Faccenna et al., 2001 [6]).

This dynamic setting results in a crustal-scale segmentation of the belt, characterized by NW–SE striking fault systems that accommodate the extensional deformation. Seismic and geodetic data confirm that the Apenninic belt extensional rate ranges between 2.5 and 5 mm/yr in a NE–SW direction (Hundstad et al., 2003 [7]), associated with the activation of high-angle normal faults that are often responsible for damaging earthquakes (e.g., the 1980 Irpinia M6.9 event;

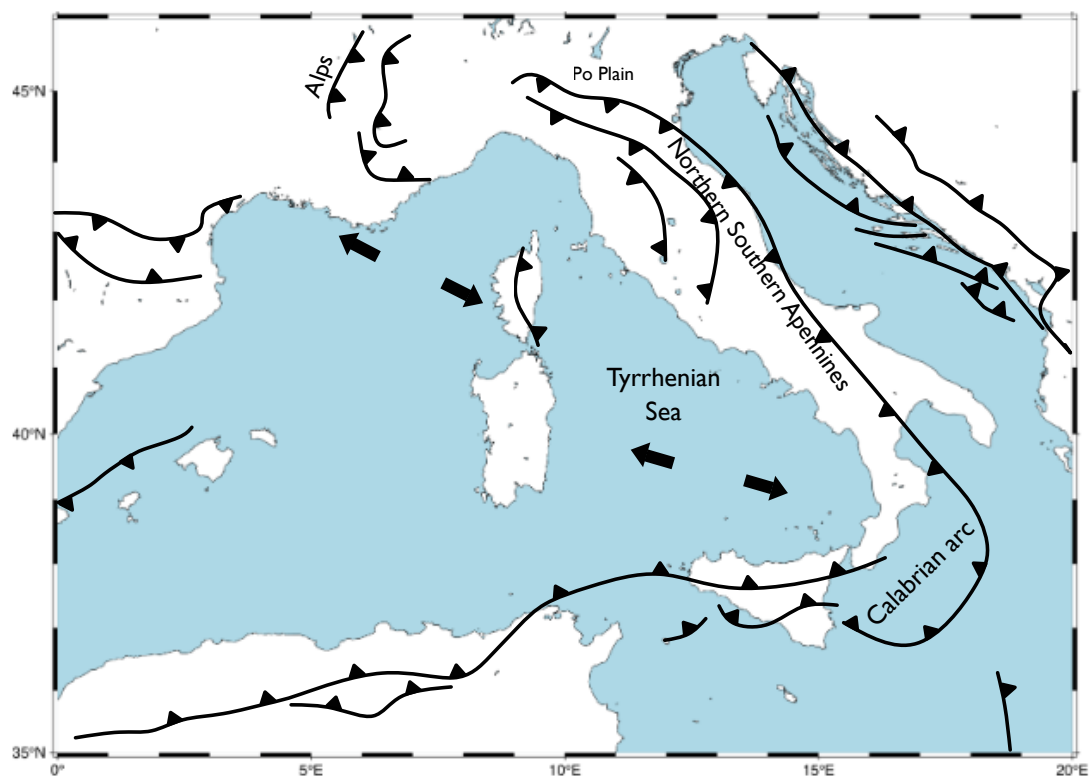


Figure 2.1: Schematic map of the central Mediterranean tectonics. Redrawn after Faccenna et al., 2001 [6].

Westaway & Jackson, 1984 [8]). Figure 2.1 shows a sketch of the tectonics of the central Mediterranean area.

The Southern Apennines are characterized by a high rate of low-magnitude seismicity, as well as occasional destructive earthquakes. Seismic activity in the Southern Apennines is mainly confined to a NW–SE striking belt, 30–50 km in width, broadly aligned with the axis of the orogenic chain (Improta et al., 2000 [9]). The region hosts several intramountain basins, such as the Val d’Agri, which accommodate Quaternary extension and serve as focal points for both natural and anthropogenic seismicity.

In this geodynamic context, the interplay between lithospheric-scale processes and surface faulting generates a highly heterogeneous stress field, making the Southern Apennines an ideal natural laboratory for studying the mechanics of active extension, fault reactivation, and induced seismicity.

2.2 Local setting: the Val d’Agri basin

2.2.1 Geological setting

The Val d’Agri (VA) basin is situated within the southern Apennines fold-and-thrust belt. Its complex structural architecture reflects the Neogene–Quaternary convergence between the Eurasian and African plates, which led to the accretion of Mesozoic–Cenozoic sedimentary sequences detached from the Adria passive margin (Patacca & Scandone, 1989 [10]; Butler et al., 2000 [11])

During the Late Pliocene–Early Pleistocene, the region experienced a thick-skinned compressional phase when high-angle, north- to northwest-trending thrust faults rooted deep within the Apulian Platform generated large-wavelength carbonate anticlines (Shiner et al., 2004 [12]) (Figure 2.2). These structures created the structural traps that today host the Val d’Agri oil field (Shiner et al., 2004 [12]). A thick (~ 1 km) tectonic *mélange*, dominated by overpressured and intensely deformed Miocene–Pliocene mudstones and siltstones, forms a mechanical decoupling horizon between the allochthonous wedge and the Apulian Platform

(Shiner et al., 2004 [12]; Schirripa Spagnolo et al., 2024 [?]) (Figure 2.2). Since the Middle Pleistocene, this compressional framework has been overprinted by a NE-SW extensional regime with a regional deformation rate of 3-5 mm/year (Ferranti et al., 2014 [13]). This younger extensional tectonics reactivated parts of the pre-existing thrust system and currently controls the seismicity of the basin (D'Adda et al., 2017 [14]; Valoroso et al., 2023 [15]).

The Val d'Agri basin is bounded by two main fault systems: the Eastern Agri Fault System (EAFS) to the northeast and the Monti della Maddalena Fault System (MMFS) to the southwest (Cello et al., 2003 [16]; Maschio et al., 2005 [17]) (Figure 2.2). These fault arrays are characterized by segmented geometries and immature extensional features. Seismic profile interpretation and analogue modeling suggest that shallow normal faults are decoupled from deeper Apulian Platform structures due to the intervening ductile *mélange* layer (Candela et al., 2015 [18], D'Adda et al., 2017 [14]).

Recent geophysical studies, including 3D local earthquake tomography, have revealed strong lateral velocity variations in the VA region, highlighting the structural complexity and abrupt depth changes in the Apulian Platform (Valoroso et al., 2011 [19]). Lavecchia et al. (2024) [20] developed a thermo-mechanical model to replicate the primary tectonic setting of the region. Their study revealed the emergence of two distinct fault systems during the contractional phase, originating from pre-existing weak zones within the Apulian Carbonate Platform. These fault systems intersect both the carbonate formation and the overlying units, delineating a region less than 10 kilometers wide that experiences uplift. Subsequently, during the extensional phase, the previously uplifted region evolves into a subsiding basin due to the development of the southwestern and northeastern border fault systems. Furthermore, their model predicts the formation of a series of synthetic and antithetic faults within the basin at this stage. Consequently, the resulting topography exhibits a basin approximately 10 kilometers wide, surrounded by reliefs with elevations ranging from 1,000 to 1,200 meters—a topography that closely resembles the current landscape of the Val d'Agri.

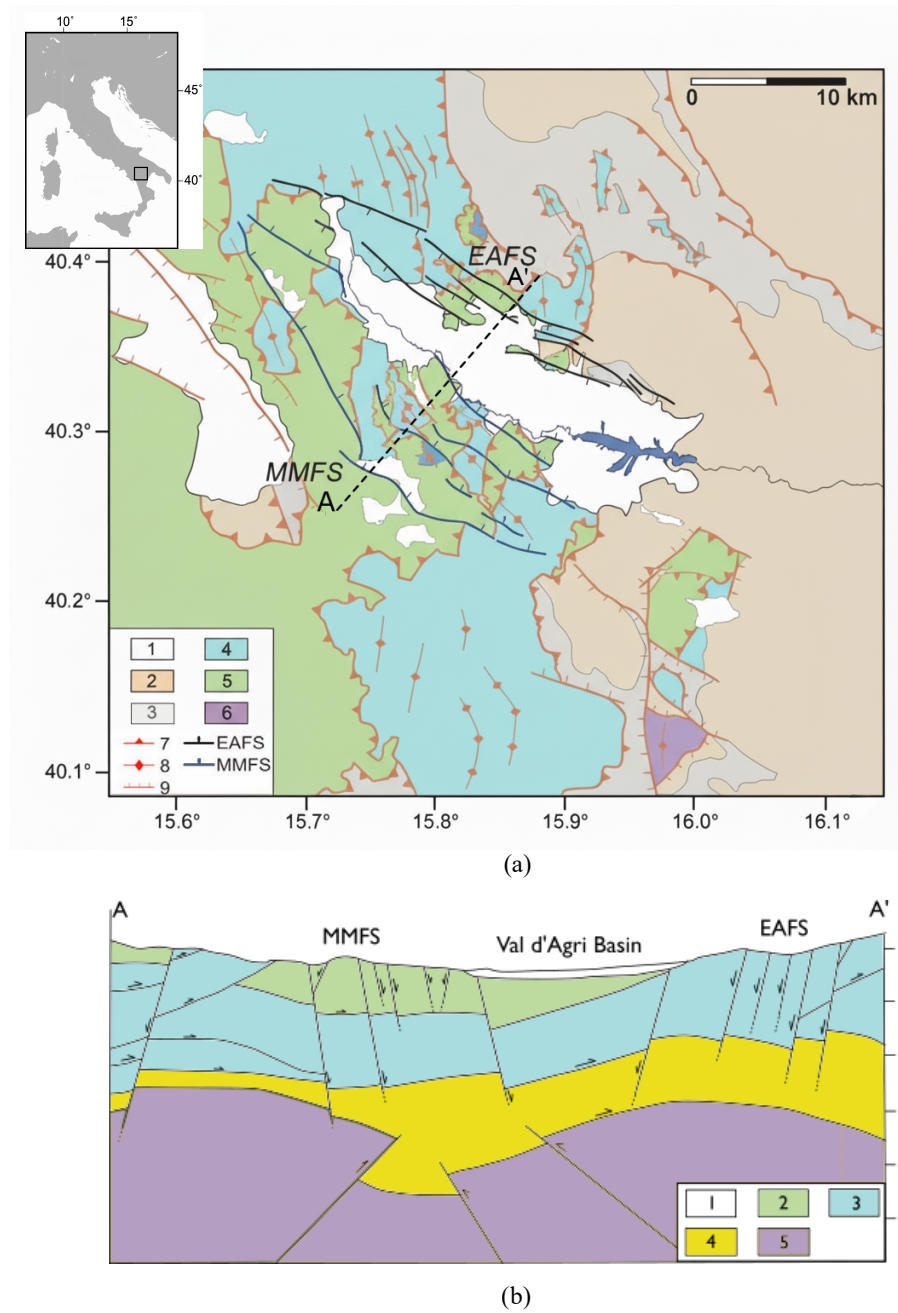


Figure 2.2: a) Geological map of Val d'Agri modified after Valoroso et al. (2011) [19]. 1- Quaternary continental deposits; 2- flysch and terrigenous sediments filling satellite basins (Middle Miocene–Pliocene); 3- pelagic and slope plastic successions (mainly clays and slope carbonates; Cretaceous–Lower Miocene); 4- Mesozoic rocks of the Lagonegro Basin (mainly cherty limestones, cherts and claystones); 5- Western Carbonate Platform (Mesozoic); 6- Apulian Carbonate Platform exposed in the Mt. Alpi tectonic window (Mesozoic–Miocene); 7- main thrusts and overthrusts; 8- antiform axis; 9- normal faults; EAFS- Eastern Agri Fault System; MMFS- Mts. Maddalena Fault System. b) Summary cross section along profile AA' (modified after Schirripa Spagnolo et al., (2024) [?]). 1- Quaternary continental deposits; 2- Western Carbonate Platform; 3- rocks of the Lagonegro basin; 4- Irpinia melange; 5- Apulian Carbonate Platform.

2.2.2 The Val d’Agri Oilfield

The Val d’Agri oilfield represents the largest onshore hydrocarbon reservoir in Europe. Its formation is the result of a complex interplay of multiple tectonic phases that have shaped the Apennine orogenic system since the Late Cretaceous (Cello and Mazzoli, 1999 [21]). Below the Quaternary sediment of the Val d’Agri basin, the stratigraphy consists of upper Messinian to Pliocene terrigenous marine deposits (Doglioni et al., 1999 [22]; Scrocca et al., 2005 [23]), which unconformably overlie a thick succession ($\sim 6\text{--}7$ km) of allochthonous Mesozoic to early Tertiary carbonate units. These units comprise the Lagonegro Basin successions and the Apulian Platform, which were involved in eastward-verging thrusting during the convergence of the African and Eurasian plates (Mazzoli et al., 2001 [24]).

The Apulian Platform carbonates were internally deformed by thrusts and folds, providing the structural traps for hydrocarbon accumulation (Shiner et al., 2004) [12]. The main reservoir consists of Lower Cretaceous to Miocene limestones and dolostones of the Apulian Platform, characterized by low primary porosity but highly productive due to extensive natural fracturing (Trice, 1999 [25]; Wavrek and Mosca, 2004 [26]). These reservoirs are sealed by Pliocene siliciclastic foredeep deposits and overthrustured *mélanges*, which locally exhibit high formation pressures (Wavrek and Mosca, 2004 [26]).

The main trap is a large structure bounded by high-angle reverse faults trending NW–SE, inherited from the compressional Apennine tectonics and subsequently reactivated in an extensional regime (Tavarnelli and Prosser, 2003 [27]). The field extends beneath the central and southern sectors of the basin and along the northeastern ridge, with three principal traps — Cerro Falcone, Monte Enoc, and Monte Alpi — hosting the most productive zones (Wavrek and Mosca, 2004 [26]). Oil production began in the mid-1990s, and by 2010 the field reached full production capacity with 24 wells operating at depths between 1.8 and 3.5 km below sea level. Current production from the Val d’Agri field amounts to about 32,000 bbl/day of crude oil and 2.4×10^6 Sm³/day of natural gas, highlighting its role as the largest onshore hydrocarbon reservoir in Western Europe (ENI website, <https://www.eni.com/eni-basilicata/it-IT/home.html>, last access 16/10/2025).

To manage co-produced saline formation water, ENI began long-term wastewater reinjection in June 2006 using the CM2 (Costa Molina) well, originally drilled in the early 1980s. The open-hole interval of CM2 spans from 2890 to 3096 m b.s.l., targeting a marginal sector of the reservoir in fractured Upper Cretaceous to Miocene carbonates. Injection rates have varied, with monthly volumes between 60,000 and 80,000 m³ and wellhead pressures reaching 130–140 bar (Improta et al., 2015 [28]).

To monitor local seismicity and its possible correlation with industrial activities (e.g., production and reinjection), ENI installed a dedicated seismic network in 2001, later expanded to 15 stations by 2012 (Eni SpA, 2001 [29]; Braun et al., 2020 [30]). The sensitivity of this network, combined with INGV stations and temporary deployments (e.g., Valoroso et al., 2009 [31]; Stabile et al., 2014 [32]), has enabled the detection of microseismicity down to magnitudes of completeness as low as $M_c \approx 1.0$ in the southern sector of the field (Improta et al., 2015 [28]).

2.2.3 Natural and induced seismicity

Microseismicity in the VA region has been extensively documented. Seismic data revealed a persistent swarm-like seismicity south of the Pertusillo artificial lake, concentrated at 1–5 km depth and composed of low-magnitude events (Valoroso et al., 2009 [31]). This seismicity shows seasonal modulation that suggests a temporal correlation with the loading/unloading phases of the water reservoir (Valoroso et al., 2023 [15]). The observed delay of about one month between lake level rise and increased seismic activity suggests a pore pressure diffusion mechanism (Stabile et al., 2014 [32]), consistent with models of reservoir-induced seismicity (Talwani, 1997 [33]).

Further evidence for anthropogenic influence on seismicity comes from induced seismic events associated with re-injection of wastewater produced by hydrocarbon extraction (Improta et al., 2015 [28]).

Historically, the region has been significantly impacted by the catastrophic 1857 Basilicata earthquake (M7.1), whose causative fault continues to be a subject of scientific debate (e.g., Burrato and Valensise, 2008 [34]; Bello et al., 2022

[35]). While such large events are rare, the region has experienced persistent low-magnitude seismicity over recent decades, typically not exceeding magnitude 3 (Improta et al., 2015 [28]). Seismicity in the basin is mainly clustered around the Pertusillo Lake. Focal mechanisms from both natural and induced earthquakes consistently indicate normal faulting along NW-SE structures, in line with the SW-NE regional extensional regime (Cucci et al., 2004 [36]; Frepoli et al. 2004 [37], Improta et al., 2017 [38], Valoroso et al., 2023 [15])

2.2.4 The local seismic networks

The study area is covered by 56 stations belonging to various networks (GE, IV, IX, TP, VA, VD) over an area of approximately 6500 km² (see Figure 2.3). Some stations are part of the INGV National Network (IV, <https://doi.org/10.13127/sd/x0fxnh7qfy>), GEOFON Seismic Network (GE, <https://doi.org/10.14470/TR560404>), CNR-IMAA High Agri Valley geophysical Observatory [VD, <https://doi.org/10.7914/SN/VD>, 39], ISNet - Irpinia Seismic Network (IX, <https://www.fdsn.org/networks/detail/IX/>). While other networks have been installed by private companies: VA (ENI SpA, 2001, <https://doi.org/10.7914/SN/VA>), and TP (TOTAL E& P Italia Spa, 2018 <https://www.fdsn.org/networks/detail/TP/>).

The network configuration is described in detail in the Supplementary Information, where Figure A.1 reports a table with the main characteristics of the 56 stations (e.g. coordinates, sensor type, sampling rate).

Station noise conditions were assessed using Probabilistic Power Spectral Density (PPSD) (McNamara & Buland, 2004) [40] analysis computed over approximately one year for each station. The corresponding plots are provided in Figure A.2 of the Supplementary Information.

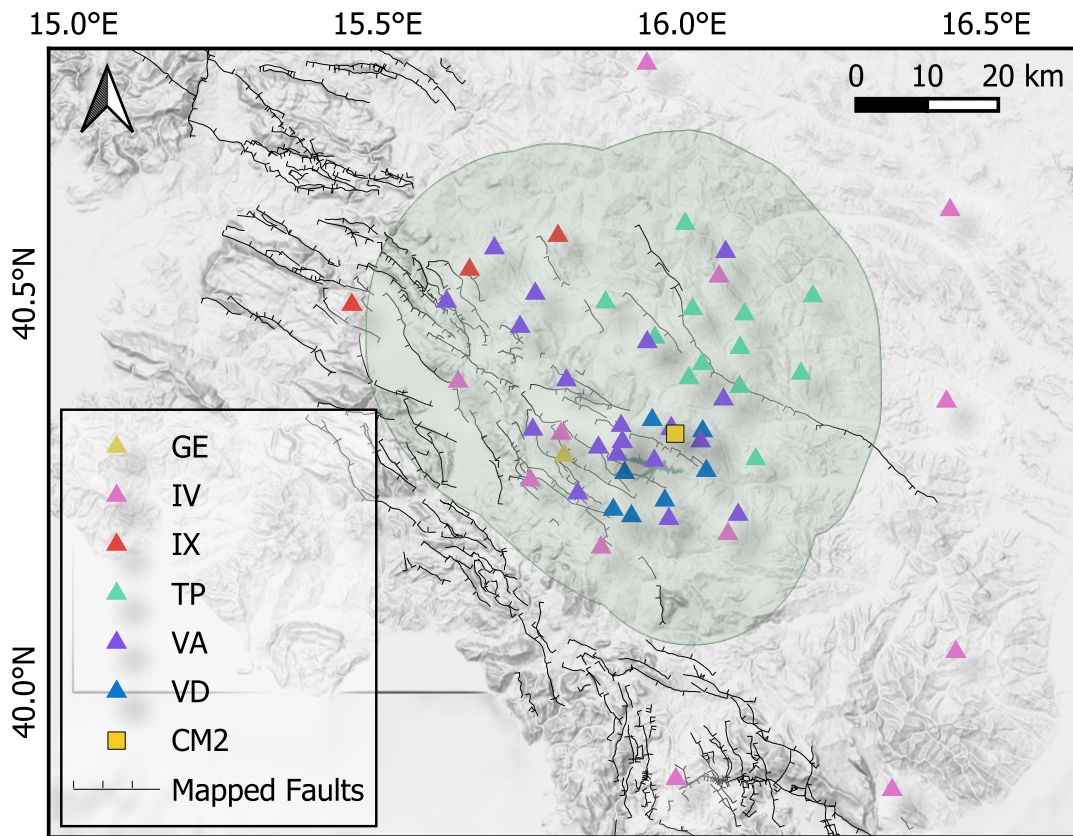


Figure 2.3: Map of the Val d'Agri integrated seismograph network. The triangles show the seismic stations colored by network, whose codes are shown in the legend. The yellow square shows the reinjection well. The shaded green area is the Domain of Attention (DA) for microseismic monitoring, and the focus of this study.

Chapter 3

Principles and applications of Deep Learning for seismic event detection

3.1 Introduction

The accurate detection of seismic events is a fundamental step in seismological analysis, serving as the basis for subsequent processes such as phase picking, event location, and source characterization. Traditional approaches, such as the Short-Term Average/Long-Term Average (STA/LTA) algorithm (Allen, 1978 [41]), have been widely adopted due to their simplicity, computational efficiency, and robustness in a variety of noise conditions. However, these methods rely on manually tuned parameters and may struggle in low signal-to-noise environments, potentially leading to missed detections or false triggers.

In recent years, the increasing availability of large seismic datasets and advances in computational resources have promoted the development of data-driven approaches. Machine Learning (ML) techniques, and in particular Deep Learning (DL), have emerged as powerful tools for pattern recognition tasks, enabling the automatic extraction of complex features directly from raw waveform data. Among DL architectures, Convolutional Neural Networks (CNNs) have proven especially effective in seismic event detection (Ross et al., 2018 [42]), thanks to their ability to capture local and hierarchical structures in time series data.

This chapter first introduces the theoretical foundations of machine learning and deep learning, with an emphasis on architectures relevant to seismic applications. It then reviews the progression from conventional detection algorithms to modern deep learning based approaches, highlighting the advantages, challenges, and potential future directions of these methods in seismology.

By outlining both the theoretical and practical implications of these methods, this chapter lays the foundation for understanding and evaluating the deep learning based tools applied in this study.

3.2 Deep Learning and Artificial Neural Networks

Deep learning is a subfield of machine learning that focuses on algorithms inspired by the structure and function of the brain’s neural networks, with the key distinction of employing architectures composed of multiple layers of non-linear transformations (LeCun et al., 2015 [43]). The “deep” in deep learning refers to the depth of these models, i.e., the number of successive representation levels through which the input data is processed. By stacking many layers, deep learning systems can automatically learn hierarchical representations, where lower layers capture local or simple features and higher layers progressively encode more abstract and complex concepts (Bengio, 2009 [44]).

We can represent a deep learning model as a parametric function (Marquardt, 2021 [45]):

$$y = f_{\theta}(x) \quad (3.1)$$

where x is the input, y is the output, θ are the parameters. In an artificial neural network, the basic computational element is the *neuron*, which stores a scalar value. Its output y is computed from the outputs y_k of connected neurons through two steps: first, a linear combination of the inputs is calculated as

$$z = \sum_k \omega_k y_k + b \quad (3.2)$$

with ω_k being the weights and b the bias (Figure 3.1). Then, a nonlinear activation

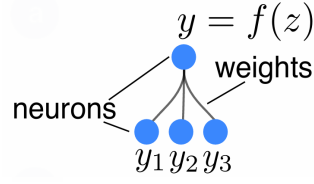


Figure 3.1: Operation of a single neuron in a neural network. Image from Marquardt, 2021 [45].

function f is applied to produce

$$y = f(x) \quad (3.3)$$

To approximate arbitrary nonlinear functions $F(x)$, multiple layers of neurons are required, where each neuron in layer $n + 1$ receives weighted inputs from all neurons in the previous layer n (see Figure 3.2). In this case, the neuron output in layer n ($y_k^{(n)}$) influences the neuron j in the next layer via the weight $\omega_{jk}^{(n+1)}$, according to the general equations:

$$z_j^{(n+1)} = \sum_k \omega_{jk}^{(n+1)} y_k^{(n)} + b_j^{(n+1)} \quad (3.4)$$

$$y_j^{(n+1)} = f(z_j^{(n+1)}) \quad (3.5)$$

This multi-layer architecture allows neural networks to model complex, hierarchical patterns in data. The output is provided by solving these equations for each layer. Weights and biases together are the network parameters θ .

Common activation functions f can be the *sigmoid*: $f(x) = 1/(1 + e^{-x})$ or the *ReLU*: $f(x) = 0$ if $x < 0$ and $f(x) = x$ if $x > 0$.

Training—The training of deep networks is typically carried out by minimizing a loss function $L(\theta)$, which measures the discrepancy between the predicted output and the ground truth labels, i.e. the function that the activation function f is trying to approximate. The loss function can be defined as:

$$L_x(\theta) = |F_\theta(x) - F(x)|^2 \quad (3.6)$$

depending on the specific input x . By averaging over all the possible inputs we

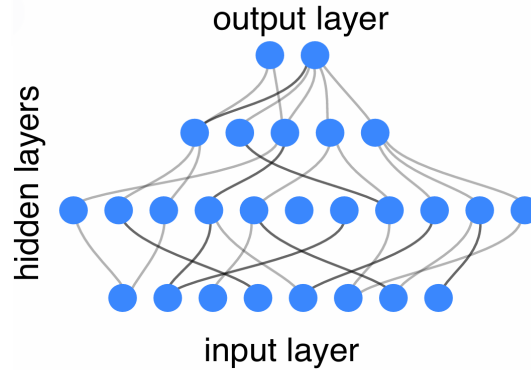


Figure 3.2: Neural network architecture. Each circle represents a neuron. Image from Marquardt, 2021 [45].

obtain:

$$L(\theta) = \langle L_x(\theta) \rangle_x \quad (3.7)$$

Parameters are optimized to minimize the loss function via stochastic gradient descent (SGD) or its variants (Bottou, 2010 [46]), by moving along the negative gradient of the loss function. Gradients are computed efficiently using the back-propagation algorithm (Rumelhart et al., 1986 [47]). This optimization process uses the chain rule of calculus to efficiently propagate the error from the output layer back through the network to all previous layers, updating each weight according to its contribution to the final prediction error.

Without going through the details of the mathematical formulation, the process consists of two main phases:

- Forward pass – The input data propagate through the network, producing an output $\hat{y} = F_{\theta}(x)$ and computing the value of the loss function $L(\theta)$.
- Backward pass – Starting from the output layer, the partial derivatives of L with respect to each parameter are computed layer-by-layer, moving backwards. This is done by reusing intermediate quantities (activations and derivatives) from the forward pass, which avoids redundant computations and significantly reduces the computational cost.

Backpropagation, combined with stochastic gradient descent, allows networks with millions of parameters to be trained efficiently, and remains the core algorithm

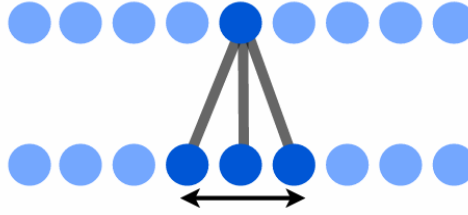


Figure 3.3: A one-dimensional convolutional neural network in which each neuron’s value is influenced only by a limited set of neighboring neurons in the preceding layer, with this relationship remaining consistent across the entire input through translational invariance (from Marquardt, 2021 [45]).

of modern deep learning despite numerous recent innovations in optimization strategies.

3.2.1 Convolutional Neural Networks

CNNs are a class of deep neural networks designed to process data with a grid-like topology, such as time series or images (LeCun et al., 1998 [48]; LeCun et al., 2015 [43]). Their key feature is the use of convolutional layers, which apply a set of learnable filters (kernels) across the input, producing feature maps that highlight local patterns. Unlike fully connected layers, where each neuron is linked to all inputs, convolutional layers exploit local connectivity and weight sharing, drastically reducing the number of parameters and improving generalization (see Figure 3.3).

This kind of architecture is inspired by a physical property called translational invariance (Marquardt, 2021[45]): in certain scenarios, the response $A(x)$ of a system at location x to a perturbation depends solely on the displacement from that point, rather than on the absolute position. Mathematically, such behaviour is described by a convolution, an operation that combines two functions based on their relative displacement:

$$A(x) = \int G(x - x')F(x')dx' \quad (3.8)$$

Convolutional Neural Networks (CNNs) leverage this property by imposing a

strong structural constraint on the network weights: instead of assigning a distinct weight to each connection, the weights depend only on the distance between input positions. In practice, the same kernel (filter) is applied across different regions of the input, enabling efficient parameter sharing and improved generalization. In the one-dimensional case, this corresponds to sliding the kernel along the signal and computing the convolution at each position.

In 1D the weights can be formulated as (Marquardt, 2021[45]):

$$w_{ij}^{(n+1)} = w^{(n+1,n)}(i - j) \quad (3.9)$$

where $w^{(n+1,n)}(i - j)$ is called *kernel*, that is set to 0 over a certain distance, dramatically reducing the number of weights that have to be stored and updated during training. A typical CNN architecture alternates three main components (LeCun, 2010 [49]):

- *Filter bank layer* extracts increasingly abstract features from the input through successive convolutions,
- *Non-linearity layer* introduces non-linearity, allowing the network to learn complex mappings,
- *Pooling layer* performs a downsampling operation (e.g., max pooling) to reduce spatial resolution, control overfitting, and increase translational invariance.

By stacking multiple convolution–activation–pooling stages, CNNs learn hierarchical representations: lower layers capture local, low-level features (edges, wavelet-like patterns), while higher layers encode more abstract and global patterns relevant for classification or regression tasks. The output of the final convolutional stage is usually flattened and passed through one or more fully connected layers before the final prediction.

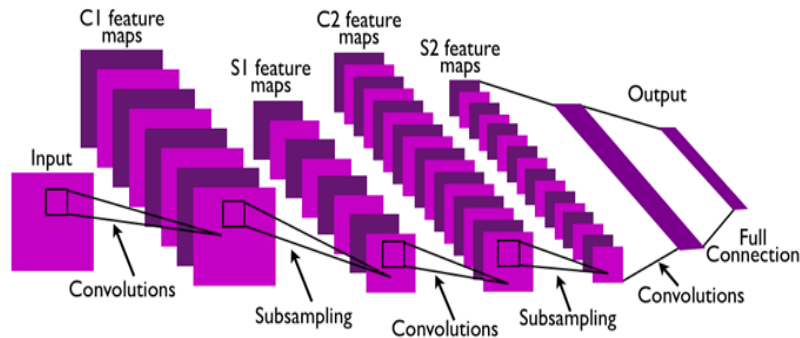


Figure 3.4: Example of a CNN architecture with two feature stages. Image from LeCun, 2010 [49]

3.2.2 Overfitting and regularization strategies

In neural network training, an epoch corresponds to a complete pass of the entire training dataset through the model (Goodfellow et al., 2016 [50]). During each epoch, the network’s internal parameters are iteratively updated to minimize the chosen loss function. However, if training is prolonged beyond the point necessary for optimal performance, the model may begin to exhibit overfitting—a phenomenon in which it learns patterns and noise specific to the training set rather than generalizable features. Overfitting typically results in excellent performance on the training data but poor accuracy when evaluated on unseen test data, indicating a loss of generalization capability (Goodfellow et al., 2016 [50]).

A widely adopted approach to address this issue is to set aside a portion of the training data as validation data. The model’s performance on this independent subset is monitored after each epoch to detect the onset of overfitting. When the validation performance ceases to improve and begins to degrade, training should be stopped. This technique is known as early stopping (Goodfellow et al., 2016 [50]). The parameter patience defines the number of consecutive epochs without improvement tolerated before terminating the training process.

Another effective regularization strategy is dropout (Srivastava et al., 2014 [51]), which introduces controlled noise into the training process. Dropout operates by randomly deactivating a fraction of neurons during each training step, thereby preventing the network from relying excessively on specific activations. This stochastic

removal of connections forces the model to develop more robust, distributed feature representations, significantly improving generalization.

3.2.3 Performance evaluation metrics

The performance of a machine learning model, particularly in classification tasks, is commonly assessed using a set of statistical metrics that quantify different aspects of predictive accuracy. Among the most widely used are precision, recall, F1-score, and accuracy (Powers, 2020 [52]). Precision, defined as

$$Precision = \frac{TP}{TP + FP} \quad (3.10)$$

measures the proportion of true positive predictions (TP) among all positive predictions (TP+FP), thus indicating how reliable positive classifications are. Recall, or sensitivity, given by

$$Recall = \frac{TP}{TP + FN} \quad (3.11)$$

quantifies the proportion of correctly identified positive cases relative to all actual positives (TP+FN), providing insight into the model's ability to detect relevant instances.

The F1-score combines precision and recall into:

$$F1 - score = 2 \frac{Precision \cdot Recall}{Precision + Recall} \quad (3.12)$$

balancing the trade-off between the two and being particularly useful when dealing with imbalanced datasets. Accuracy, defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.13)$$

expresses the proportion of correctly classified elements over the total, but may be misleading in the presence of class imbalance.

Selecting the appropriate metric depends on the specific application: in scenarios where missing a positive instance carries a high cost (e.g., medical diagnosis or

event detection in seismology), recall may be prioritized over precision; conversely, in applications where false positives are more problematic, precision may be favored. The F1-score serves as a balanced indicator when both types of errors are equally critical.

3.3 Deep Learning in seismic detection methods

3.3.1 From traditional methods to deep learning

Traditionally, earthquake detection within regional seismic networks has relied on algorithmic techniques such as the *STA/LTA* ratio and manual verification. This method operates on the continuous waveform by computing the ratio between the average amplitude calculated over a short time window (STA), sensitive to sudden changes in the signal, and a longer time window (LTA), which captures the background noise level. When the STA/LTA ratio exceeds a predefined threshold, the algorithm identifies a potential phase arrival, typically corresponding to the onset of a seismic event (Allen, 1978 [41]; Baer and Kradolfer, 1987 [53]). Due to its simplicity, computational efficiency, and low latency, the STA/LTA method has become a standard choice in operational seismic monitoring. This method is widely implemented in standard monitoring software like SeisComP3 (Weber, 2007 [54]), which automates phase picking, association, and event location in near-real-time operational workflows. However, its performance is known to degrade in noisy environments or in the presence of overlapping events, and it often requires manual tuning of parameters such as window lengths, threshold values, and filter settings—which are highly dependent on network configuration, noise conditions, and local geological features (Gentili and Michelini, 2006 [55]).

In practice, seismic analysts often perform manual inspection and correction of automatically generated picks, especially for low-magnitude or ambiguous events. This process is time-consuming, subjective, and does not scale to dense network datasets or high event volumes (Storchak et al., 2015 [56]). Moreover, pick-based methods can struggle when events overlap in time (e.g., during aftershock sequences), leading to mis-association or missed events altogether (Trnkoczy et al.,

2012 [57]).

Detecting and locating microseismic events presents significant challenges compared to larger earthquakes, primarily due to their low signal-to-noise ratio (SNR) and limited spatial detectability across seismic networks. Many conventional localization techniques rely on the assumption that a seismic event is recorded clearly at multiple stations, allowing for accurate arrival-time picking and triangulation. However, weak microearthquakes often fail to produce discernible waveforms at a sufficient number of stations, resulting in a substantial number of missed or mislocated events (Meier et al., 2019 [58]; Gajewski and Tessmer, 2005 [59]).

To overcome these limitations, various advanced techniques have been developed over the past two decades, offering alternative strategies for microseismic monitoring. One early approach is reverse time migration (RTM), as proposed by Gajewski and Tessmer (2005) [59]. This method uses time-reversed wavefields, where recorded seismograms are numerically propagated backward through a velocity model. If the model is accurate, seismic energy will focus at the true source location at time zero. RTM has shown promising results even under high noise conditions, especially when a dense receiver array is available.

Another key advancement is the Source Scanning Algorithm (SSA), originally introduced by Kao and Shan (2004) [60] and later refined by Grigoli et al. (2013) [61]. In this method, a grid search is performed over potential source locations by stacking a *brightness function* derived from waveform attributes across stations. Grigoli et al. (2013) [61] improved the algorithm by stacking STA/LTA traces using both P- and S-wave arrivals, making the method automatic and picking-free, while also enhancing performance in noisy environments and reducing dependence on precise arrival-time measurements.

In a subsequent development, Grigoli et al. (2016) [62] presented a modified waveform-stacking location method that incorporates source-specific station corrections and a master-event approach. This strategy improves location accuracy even with poorly constrained velocity models by leveraging a well-located reference event. The method's high level of automation and precision makes it particularly suitable for real-time monitoring of induced seismicity in industrial contexts, such as oil and gas operations.

Another widely adopted approach for detecting small and repeating earthquakes is the template matching technique [e.g., Gibbons et al., 2006 [63](#)]. This method involves cross-correlating continuous seismic data with pre-defined waveform templates, typically extracted from previously detected events. By exploiting the high waveform similarity of earthquakes occurring on the same fault patch, template matching enables the identification of events that are often missed by standard detection algorithms, especially those with low signal-to-noise ratios. Over the past decade, template matching has proven effective in reconstructing seismic catalogs with enhanced completeness, temporal resolution, and clustering of events, particularly in regions affected by induced seismicity [e.g., Skoumal et al., 2014 [64](#)]. Despite its high sensitivity, the technique relies heavily on the availability and representativeness of templates, making it computationally intensive and less suitable for real-time applications at large scale. These limitations have motivated the development of more scalable and adaptive approaches, such as those based on deep learning, which aim to combine the sensitivity of template matching with the generalization capability of neural networks.

3.3.2 Examples of Deep learning in seismic event detection

In recent years, deep learning has emerged as an excellent tool for seismic signal processing, particularly in low SNR contexts. CNN-based models have been successfully applied to a range of seismological tasks, such as phase picking (Ross et al., 2018 [\[42\]](#); Zhu & Beroza, 2019 [\[2\]](#), Mousavi et al., 2020 [\[65\]](#)), phase association (Ross et al., 2019 [\[66\]](#)), first-motion polarity assignment (Ross et al., 2018 [\[42\]](#)).

These models have demonstrated significant improvements over traditional methods in terms of accuracy, robustness, and scalability, particularly in the detection of small-magnitude events and in the presence of noisy data (Mousavi et al., 2020 [\[65\]](#)). Since deep learning pickers learn features directly from the waveform, they reduce phase misdetections, minimize false alarms, and deliver better performance in noisy and low SNR environments—a clear advantage over STA/LTA and manual-based workflows (Mousavi and Beroza, 2023 [\[67\]](#)).

Overall, the shift from threshold-based detection to waveform-based and learning-based methodologies reflects the growing demand for high-resolution, real-time

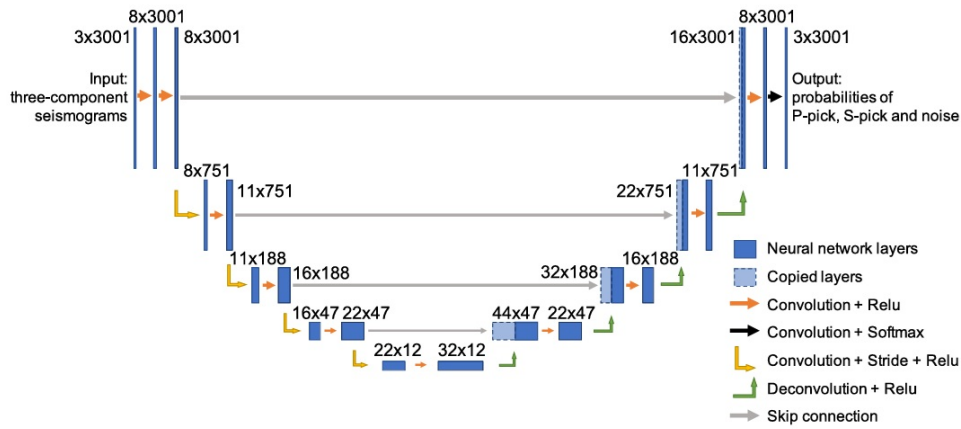


Figure 3.5: PhaseNet architecture, from Zhu & Beroza (2019) [2]

microseismic monitoring in complex environments. By moving to data-driven models, seismic monitoring workflows have become increasingly automated and efficient, enabling the generation of richer and higher-resolution catalogs.

3.3.3 CNN tested: PhaseNet

A significant example of application of these methods is PhaseNet —a convolutional neural network designed for picking seismic phase arrivals directly from raw waveforms (Zhu & Beroza, 2019 [2]).

PhaseNet is trained on a large dataset of manually picked phase arrivals from the Northern California Earthquake Data Center. During training, the model minimizes a categorical cross-entropy loss, encouraging accurate classification of each sample. The architecture of PhaseNet (Figure 3.5) is adapted from the U-Net model (Ronneberger et al., 2015 [68]) to process 1D time-series data. Originally developed for biomedical image segmentation, U-Net is designed to localize specific features within an image. In our context, this approach is applied to seismic time-series data to identify and classify three categories: P-wave arrivals, S-wave arrivals, and noise. The input consists of three-component seismograms from known earthquake events, each containing 3001 data points per component (corresponding to 30 seconds sampled at 100 Hz).

The network processes the input through four downsampling layers followed by four

upsampling layers. Each layer applies 1-D convolutions paired with rectified linear unit (ReLU) activations. The downsampling path compresses the raw seismic signals, concentrating relevant features into a smaller representation where each neuron corresponds to a broader receptive field. Subsequently, the upsampling path reconstructs this compressed information into probability distributions for P-wave, S-wave, and noise at each time step. Skip connections link the corresponding layers in the downsampling and upsampling paths, allowing feature maps to bypass deeper layers (Zhu & Beroza, 2019 [2]). This structure improves the stability and speed of training convergence (Ronneberger et al., 2015 [68]).

In the final layer, a softmax activation function converts the raw output values $z_i(x)$ for each class i belonging to $(noise, P, S)$ into normalized probabilities:

$$q_i(x) = \frac{e^{z_i(x)}}{\sum_{k=1}^3 e^{z_k(x)}} \quad (3.14)$$

where $q_i(x)$ represents the predicted probability of class i at time point x . The network is trained by minimizing the cross-entropy loss between the true probability distribution $p(x)$ and the predicted distribution $q(x)$, defined as:

$$H(p, q) = - \sum_{i=1}^3 p_i(x) \log(q_i(x)) \quad (3.15)$$

Finally, the P- and S-wave arrival times are identified by detecting the peaks in the corresponding output probability distributions. Figure 3.6 visually shows an example of the input seismograms and the corresponding output of PhaseNet in terms of P and S picks probability.

Thanks to its speed and generalization capabilities, PhaseNet is well-suited for real-time seismic monitoring and large-scale catalog generation. Its architecture is flexible enough to be retrained or fine-tuned on region-specific datasets, making it applicable across different tectonic settings and instrumentation schemes.

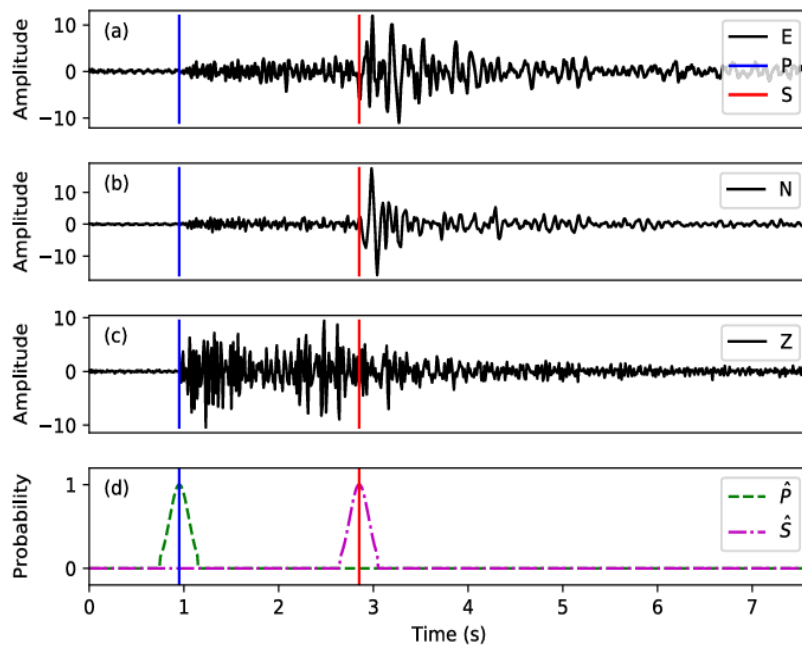


Figure 3.6: Example of PhaseNet input and output: (a)– (c) Input seismograms of the 3 components: East, North, Vertical. The blue and red vertical lines are the manually picked P and S arrival times. (d) The output is a converted probability for P and S picks. The shape is a truncated Gaussian distribution where the mean (μ) is the arrival time and has a standard deviation (σ) of 0.1 s. Image from Zhu & Beroza (2019) [2].

Chapter 4

Implementation of DL-based workflows for the detection of micro-seismicity

4.1 Introduction

Building on the theoretical framework outlined in the previous chapter, this section¹ focuses on the practical implementation of deep learning techniques for seismic monitoring in the Val d’Agri region, a tectonically active and anthropogenically influenced area in Southern Italy. The primary goal is to evaluate the performance of the PhaseNet algorithm in detecting and characterizing local seismicity, and to compare its output with that of the existing institutional catalog produced by the Centro di Monitoraggio delle attività di Sottosuolo (CMS-INGV).

I begin by describing the dataset. The subsequent sections detail the practical implementation of two automatic detection workflows tailored to the Val d’Agri dataset. While both rely on PhaseNet for phase picking, they diverge in the subsequent processing steps, especially in the event association and location procedures. Then, I show the results obtained, outlining differences in terms

¹This section is taken from: E. Caredda, M.P. Isken, S. Cesca, M. Errico, G. Zerbinato, and A. Morelli. Improving microearthquake detection in the Val d’Agri region (southern Italy) with deep learning, 2025, *Seismica* 4 (2)

of catalog completeness, magnitude distribution, spatial patterns, and overall detection sensitivity. Particular attention is given to the identification of previously undetected microseismic events, which are often missed by conventional workflows.

By comparing deep-learning-generated catalogs with STA/LTA-based reference datasets, this chapter aims to assess the added value of modern algorithms in enhancing seismic surveillance in complex settings. The chapter is taken from an article published in *Seismica* [69]: *Improving microearthquake detection in the Val d’Agri region (Southern Italy) with deep learning* (Caredda et al., 2025). This work provides a foundation for the subsequent analyses of spatio-temporal waveform-similarity-based clustering, source mechanisms, and hazard implications developed in the following chapters.

4.2 Data

For each of the 56 stations shown in section 2.2.4 and for each component (one vertical and two horizontal), I select one of the channels HH, EH or CH (sampling rates 100 to 250 Hz), in this prioritizing order, depending on availability. I analyse data from January 2021 to October 2023 (33 months).

I employ the earthquake bulletin from CMS monitoring operations as a reference catalog. It has been obtained using the STA/LTA algorithm, with detections revised daily by operators. It contains 5,030 earthquakes in the selected period, with a local magnitude range from -1.1 to 3.8 (Figure 4.1).

4.3 Methods

4.3.1 PRN workflow

I refer to the first of the two workflows as PRN, after the names of the software employed at each stage of the procedure: PhaseNet for phase picking (Zhu & Beroza, 2019 [2]), REAL for phase association (Zhang et al., 2019 [71]), and NonLinLoc for hypocenter location (Lomax et al., 2000, 2014 [72, 73]). The workflow is structured

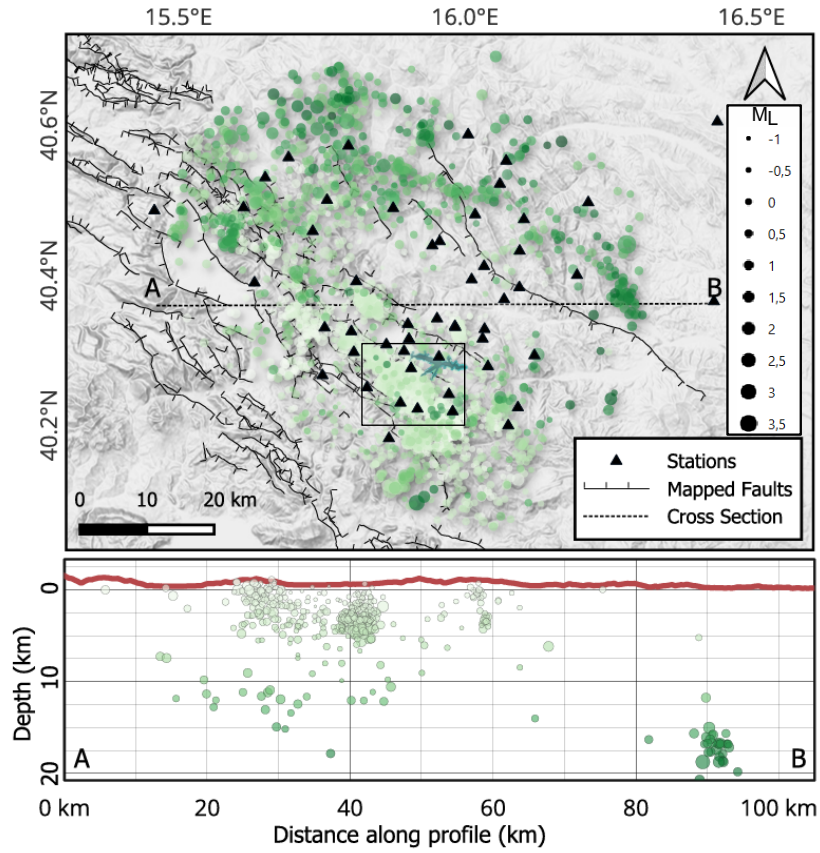


Figure 4.1: Distribution of epicenters and relative vertical cross sections for the reference CMS catalog. Vertical cross section represent a sliver with a thickness of 5 km on each side. The dot size is proportional to Magnitude, while the color scales with the depth. Fault traces from the dataset by Lavecchia (2023) [70]. The black rectangle around the Pertusillo lake area indicates the region selected for analysis in the Discussion section.

into sequential steps—picking, association, and location—reflecting a modular design that ensures flexibility in the processing of seismic data. An additional step of magnitude estimation is applied subsequently, based on the amplitude measurements associated with the located events. Figure 4.2 visually represents the PRN workflow steps. To improve computational efficiency, the workflow is parallelized at two levels. First, job-array parallelization on HPC is used to process each day independently. Second, within each day, waveform files from individual stations are processed concurrently. This hierarchical parallelization allows the pipeline to scale efficiently with both data volume and network size.

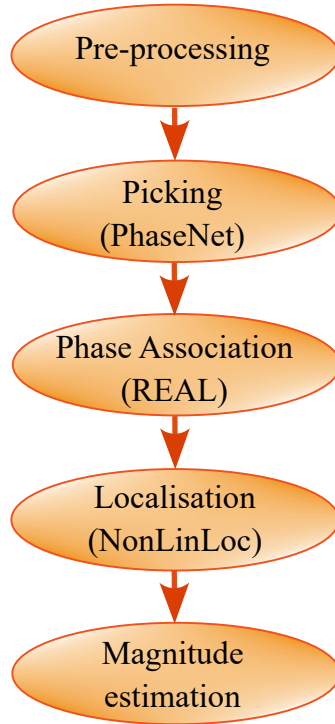


Figure 4.2: PRN workflow steps.

Pre-processing

Before the picking step, I preprocess the seismograms by applying a bandpass filter between 1 and 45 Hz. This frequency range is chosen to maximize signal retention while enhancing the recall rate of seismic phase detection, particularly for micro-earthquakes, which radiate significant energy at high frequencies. The upper bound is constrained by the Nyquist frequency (given the 100 Hz sampling rate required by PhaseNet), while the lower bound helps suppress low-frequency noise. This choice also reflects common practice in similar studies [e.g., Mousavi et al., 2020; Wickham-Piotrowski et al., 2023 [65](#), [74](#)], supporting its general applicability and reproducibility.

Picking

I use the PhaseNet original model trained on the Northern California Earthquake Data Center Catalog. The PhaseNet CNN description is in section [3.3.3](#).

I set the PhaseNet pick score (a pseudo-probability) threshold to 0.3 for the P and S phase arrivals to filter out low-confidence picks. This value was chosen based on empirical tests on this dataset and is consistent with thresholds adopted in previous studies using PhaseNet (e.g., Zhu & Beroza, 2019 [2]). This threshold offers a good balance between minimizing false picks and retaining valid arrivals, especially in low-SNR conditions. Although a PhaseNet variant trained on the Italian INSTANCE dataset (Michellini et al., 2021 [75]) exists, a recent comparative study in the Central Apennines (Cianetti et al., 2025 [76]) showed only marginal improvements over the original model. Given the similar tectonic context and waveform characteristics, I expect comparable performance in this study area and therefore opted for the original model.

Association

I apply the phase association algorithm REAL (Rapid Earthquake Association and Location; Zhang et al, 2019 [71]). Unlike traditional waveform-based approaches, REAL applies the delay-and-sum strategy directly to seismic picks rather than continuous waveform data. The delay-and-sum strategy is a classic beamforming technique in which the observed arrival times are shifted (*delayed*) according to predicted travel times from a trial source location and then summed. If the trial location is correct, the shifted arrivals align in time and the summed signal (or, in REAL's case, the number of coherent picks) reaches a maximum.

The algorithm simultaneously associates P- and S-phase picks to seismic events and estimates their locations by evaluating the number of associated arrivals and minimizing travel-time residuals. Instead of scanning the entire waveform or computing a brightness or coherence function sample-by-sample, REAL focuses on discrete pick times, which makes the method computationally efficient and well-suited for processing large datasets [71].

REAL operates in three main steps (Zhang et al., 2019 [71]):

- Phase picking and amplitude estimation: seismic waveforms are processed to compute vertical and horizontal energy traces. In my case, phase picks (P and S) are read from PhaseNet output files. If multiple picks appear

within a specified time window, only the one with the highest PhaseNet pseudo-probability is retained;

- Grid search and event association: a 3D grid search is initiated around the station with the earliest P pick. Theoretical P and S travel times are precomputed using constant velocities. Observed picks are associated with potential hypocenters if they fall within time windows around predicted arrivals. The grid point with the highest number of associated picks and the smallest residual is selected as the event location;
- Event validation and magnitude estimation: an event is confirmed if the number of associated picks exceeds a defined threshold. To avoid duplicate detections, only the most reliable event within a short time window is retained. The local magnitude is then estimated using corrected phase amplitudes and an empirical relationship.

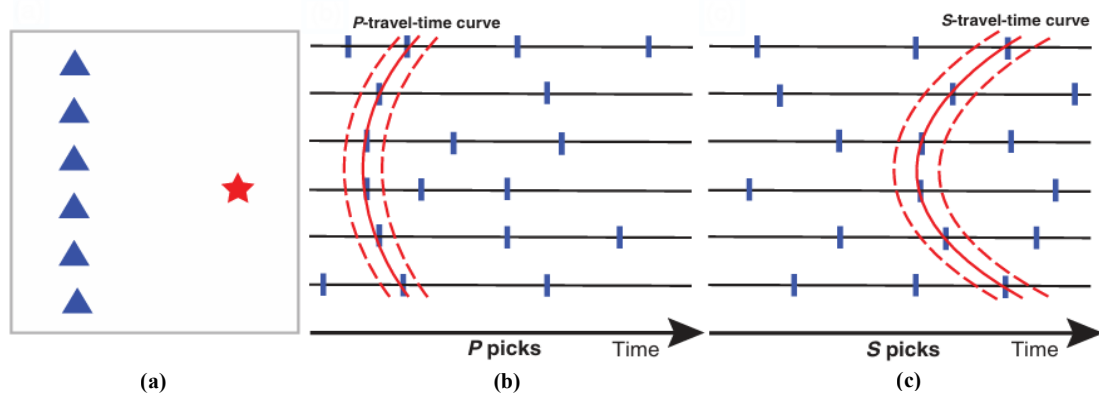


Figure 4.3: Schematic representation of the REAL method for rapid earthquake association and location: (a) The red star indicates the true earthquake location, while blue triangles represent the distribution of seismic stations. (b) The theoretical P-wave arrival time is shown as a red curve, with dashed red lines representing the uncertainty range caused by limitations in the velocity model and the finite grid search. Both true and false P picks are shown relative to this theoretical curve. (c) Same as (b), but for S-wave arrivals, illustrating the expected S arrival-time curve, associated picks, and uncertainties. Modified after Zhang et al. (2019)

In this workflow, I require a minimum of 7 phase arrivals per earthquake (at least 4 P phases, 3 S phases, and three stations with both P and S phases). I chose this threshold because it is a good compromise between reducing false positives

and detecting small earthquakes. Furthermore, it is a slightly more restrictive threshold compared to the CMS catalog, which uses a minimum number of 5. Since PhaseNet is more sensitive than STA/LTA method and generally provides a higher number of picks for the same event, I opted for a higher threshold to ensure robustness and reduce the inclusion of low-quality detections.

I set the maximum allowed residual for the preliminary locations computed by REAL to 2 seconds in order to include a larger number of candidate events during the initial phase association step. This approach is particularly useful when working with low-magnitude events, where stricter residual thresholds might exclude valid detections. Unlike NonLinLoc, REAL does not employ a detailed velocity model but rather uses constant P- and S-wave velocities, which justifies adopting more permissive residual limits at this stage.

S arrivals without a corresponding P arrival at the same station for the same event are discarded because they are generally less reliable.

Localisation

After phase association, I use NonLinLoc software (Lomax et al., 2000, 2014 [72, 73]) to solve for the absolute hypocentral location of the seismic events. Unlike traditional linearized inversion methods, NonLinLoc explores the full posterior probability density function (PDF) of the hypocenter by sampling a predefined 3D spatial grid and computing travel times using precomputed travel-time grids (Lomax et al., 2000, 2014 [72, 73]). For each grid point, the algorithm evaluates a misfit function based on the difference between observed and theoretical arrival times, incorporating data uncertainties and possible station corrections. The final location is represented as the maximum likelihood point or as the expectation value of the PDF, accompanied by full uncertainty estimates in space and time. This approach is particularly advantageous in regions with heterogeneous structures or sparse station coverage, as it avoids the bias introduced by linear assumptions and provides robust confidence regions for the estimated hypocenters.

Here, I briefly explore the theory behind the method by Lomax et al., 2000 [72]. The NLLoc algorithm for earthquake hypocenter determination (Lomax et al., 2000 [72]) is grounded in a probabilistic approach to inverse problems, originally

formulated by Tarantola and Valette (1982) [77] and later expanded by Tarantola (1987) [78].

An inverse problem is a mathematical approach used to infer unknown parameters of a system from observed data, typically by working backward from observations to model parameters (Tarantola and Valette, 1982 [77]). This methodology interprets both prior knowledge and observational data in terms of probability distributions, enabling a statistical representation of uncertainties in model parameters. Specifically, a normalized probability density function $f(x)$ defines the likelihood that a given parameter x lies within a specified interval $(X, X + \Delta X)$ (from Lomax et al., 2000 [72]):

$$P(X \leq x \leq X + \Delta X) = \int_x^{x+\Delta X} f(x)dx \quad (4.1)$$

In the context of geophysical inversion, the problem consists of estimating a vector of unknown parameters $\mathbf{p}$, such as hypocentral coordinates and origin time, based on a set of observed data $\mathbf{d}$, and a forward model $\theta(\mathbf{d}|\mathbf{p})$ that predicts observations from given parameters. When the prior information on the model $\rho_p(\mathbf{p})$ and the data $\rho_d(\mathbf{d})$ are considered statistically independent, the forward model can be recast as a conditional probability $\theta(\mathbf{d}|\mathbf{p})$. The combination of this likelihood with the prior model information yields a posterior probability density function $\sigma_p(\mathbf{p})$, which characterizes the most probable values of the unknowns given the observed data.

$$\sigma_p(\mathbf{p}) = \rho_p(\mathbf{p}) \int \frac{\rho_d(\mathbf{d})\theta(\mathbf{d}|\mathbf{p})}{\mu_d(\mathbf{d})} d\mathbf{d} \quad (4.2)$$

where $\mu_d(\mathbf{d})$ represents the absence of information about parameters and data.

In the context of earthquake location, the parameters to be estimated are the hypocenter coordinates $\mathbf{x}=(x, y, z)$ along with the origin time T . At the same time, the observational dataset consists of seismic arrival times $\mathbf{t}$. The forward model describes the theoretical travel times $\mathbf{h}$, linking the observed and model parameters.

According to the probabilistic framework introduced by Tarantola and Valette (1982)[77], if both the theoretical model predictions and the arrival time measurements are characterized by Gaussian uncertainty distributions, associated with covariance matrices $\mathbf{C}_T$ (for the model) and $\mathbf{C}_t$ (for the observed arrival times), and

assuming a uniform prior distribution for the origin time T , one can analytically integrate over the data space as described in their formulation. This includes integrating over T to derive the marginal probability density function (PDF) for the spatial coordinates of the hypocenter, denoted as $\sigma(\mathbf{x})$. Under these assumptions, this marginal PDF simplifies to the form presented in Tarantola and Valette (1982) and later refined by Moser et al. (1992) [79]. This probabilistic formulation allows for a comprehensive quantification of uncertainty and naturally incorporates observational and modeling errors into the earthquake location process.

$$\sigma(\mathbf{x}) = A \rho(\mathbf{x}) \exp\left[-\frac{1}{2}g(\mathbf{x})\right] \quad (4.3)$$

where A is a normalization factor, $\rho(\mathbf{x})$ is the density function of the prior information about the model parameters, and $g(\mathbf{x})$ is an L2 misfit function, which quantifies the difference between observed data and model predictions as the sum of the squared residuals, in this case formally written as:

$$g(\mathbf{x}) = [\hat{\mathbf{t}}_0 - \hat{\mathbf{h}}(\mathbf{x})]^T (\mathbf{C}_t + \mathbf{C}_T)^{-1} [\hat{\mathbf{t}}_0 - \hat{\mathbf{h}}(\mathbf{x})] \quad (4.4)$$

In Equation 4.4 $\mathbf{t}_0$ represents the vector of the difference between observed arrival times $\hat{\mathbf{t}}_0$ and their weighted mean, while $\hat{\mathbf{h}}$ the difference between theoretical travel times and their weighted mean. The weights are given by:

$$w_i = \sum_j w_{ij}; \quad w_{ij} = [(\mathbf{C}_t + \mathbf{C}_T)^{-1}]_{ij} \quad (4.5)$$

According to Moser et al., (1992) [79] the maximum likelihood of the origin time corresponding to a hypocenter located at (x, y, z) can be written as:

$$T_{ML}(\mathbf{x}) = \frac{\sum_i \sum_j w_{ij} [t_i - h_i(\mathbf{x})]}{\sum_i \sum_j w_{ij}} \quad (4.6)$$

The grid search for the maximum of the posterior probability density function adopts an octree-based algorithm. This hierarchical spatial subdivision allows for a coarse-to-fine sampling of the search volume, increasing resolution in areas with high probability and reducing computational cost elsewhere (Meagher, 1982 [80];

Lomax, 2000 [72]).

The posterior probability density function $\sigma(\mathbf{x})$, as defined in Equation 4.3, provides a comprehensive probabilistic solution for the earthquake location. It includes both the uncertainty associated with the solution and the spatial resolution of the inferred hypocenter.

Building upon this theoretical framework, my implementation adopts the 1-D layered velocity model from Improta et al., (2017)[38], the same used to generate the reference catalog, to ease comparison. This choice ensures consistency between the two location procedures and facilitates a more direct comparison of the resulting hypocentral solutions.

After a first run, I run NonLinLoc a second time, considering P and S wave station-specific delays, derived from the average residuals after the first run (Lomax et al., 2000, 2014 [72, 73]). This improves location accuracy by reducing travel-time uncertainties.

At this stage, the workflow yields 21,617 earthquake locations.

Filtering out false positives

To assess the reliability of the detections, I manually inspected a representative subset of 100 randomly selected events. Of these, 52 were confirmed as true earthquakes with reliable hypocentral solutions, yielding a precision of $\sim 52\%$ at this stage. To filter out poor locations, events with horizontal and vertical uncertainties larger than 5 km, and those with RMS larger than 1 s are removed. Such tuning offers a good compromise: it reduces the number of false positives and bad associations while recovering the majority of earthquakes previously identified from the reference catalog. However, the new catalog does not include a small number of earthquakes from the original catalog due to the conservative thresholds used to filter out bad associations and locations. Among the rejected events, some correspond to earthquakes that actually occurred outside the study area but were mislocated within it, while others exhibit very low signal-to-noise ratios, making a clear visual confirmation difficult.

A common feature among these mislocated events was a strong imbalance between the number of P and S phases, with S picks often accounting for less than one-third of the total associated phases. Although a minimum of three S phases was required for event validation, such a low P/S ratio suggests limited spatial constraint, potentially leading to unstable or biased hypocenter solutions.

To mitigate this issue, a quality check based on P/S phase number ratio and a visual inspection were introduced.

To reassess the reliability after the refinements, I manually inspected a new subset of 100 randomly selected events. This analysis indicates that 88% of the detections correspond to true earthquakes with reliable hypocentral solutions, demonstrating a substantial improvement in precision compared to the first evaluation. This value reflects the intended balance between completeness and precision.

The number of PhaseNet picks associated per station was analysed to highlight possible differences in detection efficiency and noise conditions (Figure A.3 in Supplementary Information). Differences in the number of picks per station can arise from several factors. Stations located near clusters of earthquakes record more picks due to higher local seismicity. In fact, microearthquakes may not be detected at more distant stations even if they are well-functioning. Environmental noise, site conditions, and sensor characteristics (sensitivity, frequency response) can also affect pick detection. Furthermore, temporary data gaps may further reduce the number of picks at some stations.

Magnitude estimation

Finally, I calculate the local magnitude M_L using the attenuation formula by Bakun and Joyner (1984) [81]:

$$M_L = \log_{10} A + \log_{10} \frac{r}{100} + 0.00301(r - 100) + 3 \quad (4.7)$$

where A is the amplitude in mm of the simulated record of a Wood-Anderson seismometer, and r is the epicentral distance in km. I use this equation because it is the same local magnitude relationship adopted for the reference CMS catalog for the Val d'Agri region. Although originally developed for Californian data, this

relationship was chosen by the CMS team as a practical proxy for M_L in this specific region (Morelli et al., 2025 [82]). Therefore, I apply the same formula to ensure consistency and comparability with the reference catalog.

To calculate the magnitude, I first deconvolve the instrument response, simulate the corresponding Wood-Anderson response, and calculate amplitude A as the mean of the two horizontal half peak-to-peak amplitudes in millimeters (mm) of the recordings for each triggered station. The peak-to-peak amplitude is calculated within a window that starts 0.5 seconds before the P-wave arrival and ends 3 seconds after the expected S-wave arrival time. The final local magnitude is calculated as the median of all station magnitudes. The results are compared to the reference CMS (INGV) catalog.

4.3.2 QS workflow

Qseek

This workflow employs the *Qseek* (QS) framework (Isken et al., 2025 [83]), an automatic and data-driven earthquake detector and locator. It uses neural network phase arrival annotations and waveform stacking for detection and localisation of seismicity using an adaptive octree grid search, an algorithm that uses a hierarchical octree (Meagher, 1982 [84]) data structure to represent and efficiently search a 3D space by recursively subdividing it into eight smaller cubes. The location accuracy is further improved by applying source-specific station corrections (SSSCs), defined as station-specific 3D volumes of delay times that account for the natural 3D variability of seismic velocities (Lomax and Savvaidis, 2022 [85]). For each station, a regular octree grid is used to estimate travel-time corrections by averaging the residuals of nearby events, weighted by both their distance from each node and their pick confidence. The resulting 3D delay-time grids allow interpolation of phase- and station-dependent correction terms for arbitrary event locations, thus compensating for local velocity anomalies not captured by the 1D model (Isken et al., 2025 [83]; Richards-Dinger and Shearer (2000)[86]).

I use PhaseNet (with the original model, trained on the Northern California Earthquake Catalog [2]) for P and S first arrival picking. *Qseek* uses the neural

network output (annotation function) as image function $I(t)$ (from Isken et al., 2025)[83]:

$$I^\theta(t) = PN^\theta(\phi(t)) \quad (4.8)$$

where θ can be the P or S phase, PN is the annotation function of the employed neural network (in this case PhaseNet), representing a probability, and $\phi(t)$ is the seismogram. As I show in paragraph 3.3.3, the neural network provides Gaussian-shaped annotations for each pick. The peak value (between 0 and 1) reflects the certainty of the phase identification (Zhu & Beroza, 2019 [2]). In *Qseek* it corresponds to $\max(I(t))$. Theoretical travel times from each potential hypocentral location to the seismic stations are computed based on an a 1D velocity model (Improta et al., 2017 [38]). These travel times are then used to temporally align the phase picks during the stacking process, which is essential for accurate event detection and location. The method consists in back-projecting the recorded seismic energy toward possible hypocentral locations by aligning and stacking image functions from multiple stations based on predicted travel times. This stacking is performed over all grid nodes within the search volume, producing a coherence metric referred to as the semblance S (from Grigoli et al., 2016 [87]), which indicates the likelihood of a true source at each location:

$$S_\theta(\mathbf{x}, t) = \sum_s^N I_s^\theta(t + \tau_s^\theta(\mathbf{x})) \quad (4.9)$$

where τ_s^θ indicates the theoretical travel time of a seismic wave (P or S) at the seismic station s in a network of stations N . Figure 4.4b visually represents the process for an hypothetical detection in 2D.

The semblance for P and S phases are combined into a single function $\hat{S}(\mathbf{x}, t)$ using the power mean, defined as:

$$\hat{S}(\mathbf{x}, t) \equiv \hat{S}(S_\theta(\mathbf{x}, t), \eta) = \left(\frac{\sum_\theta \omega_\theta S_\theta^\eta}{\sum_\theta \omega_\theta} \right)^{\frac{1}{\eta}} \quad (4.10)$$

where ω_θ is the phase weight defined by the user, and η is a tunable parameter useful to shape the stacked function, with typical values between 1 and 1.5. The hypocenter location is identified as the one that maximizes the semblance $\hat{S}(\mathbf{x}, t)$. The semblance function $\hat{S}(\mathbf{x}, t)$ is evaluated at each grid point $\mathbf{x}$ and time t , and

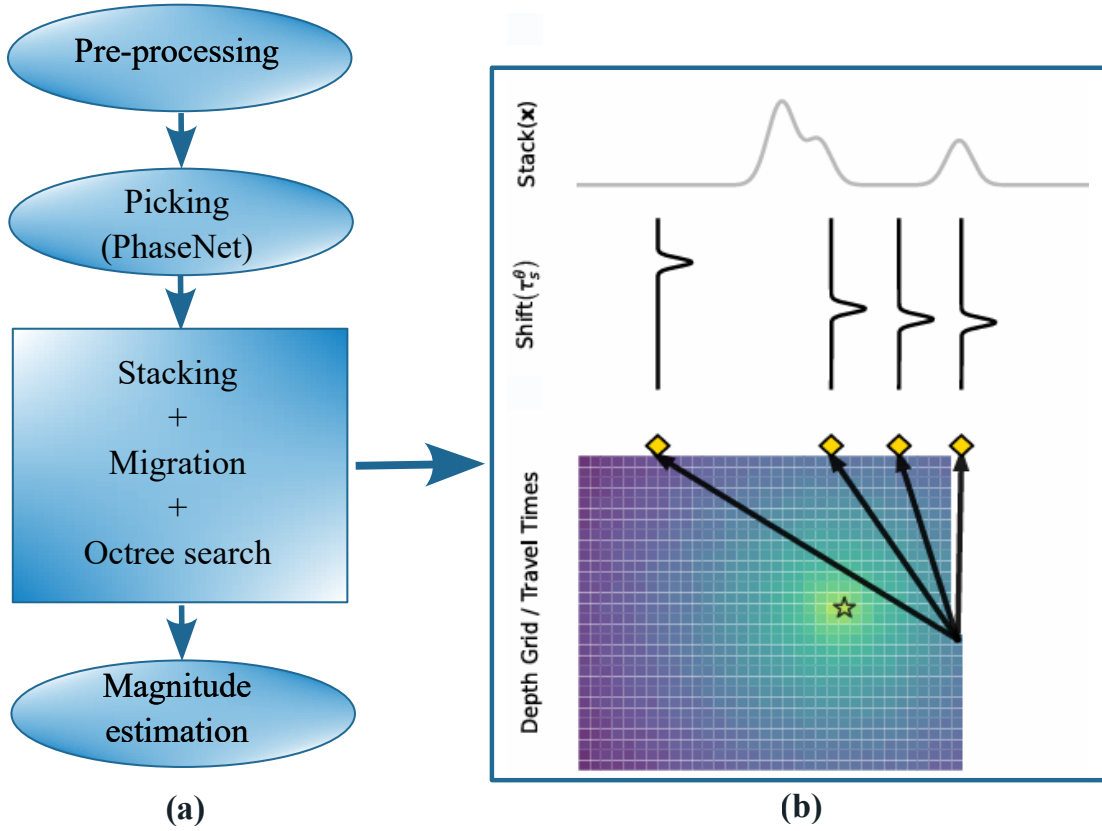


Figure 4.4: a) QS workflow employed. b) Stacking and migration grid search method (2D concept): the method involves shifting the positive portions of the seismic trace annotation functions according to theoretical travel-times τ_s^θ computed for each grid node $\mathbf{x}$. These shifted traces are then stacked for each grid point. As a result, for every node in the grid, a time series of semblance values is obtained. The location and origin time of the seismic event are inferred from the grid point and time at which the semblance stack reaches its maximum, indicating the most likely hypocenter position. This maximum is typically represented by a star in schematic diagrams (modified after Isken et al., 2025 [83]).

seismic events are identified as local maxima in this function. To detect these maxima, a two-step procedure is applied. First, the maximum semblance value across space is computed for each time step, yielding a reduced time series

$$\bar{S}(t) = \max_{\mathbf{x}}(\hat{S}(\mathbf{x}, t)) \quad (4.11)$$

Peaks in this reduced function are then identified using a threshold-based trigger. However, this approach may fail to distinguish closely spaced events occurring

within less than 1 second.

To improve spatial resolution near high-semblance areas while reducing computational cost elsewhere, a hierarchical octree structure is used. Starting from a coarse grid, the octree adaptively refines the search volume by subdividing nodes where semblance is high (Meagher, 1982 [84]). This iterative process converges toward the most likely event location by progressively increasing spatial resolution only where necessary.

The advantage of this method is that stacking the arrivals enhances the signal-to-noise ratio (SNR) and allows for assessing the coherence of phase picks across multiple stations. Coherent energy from true seismic arrivals is reinforced, while incoherent noise is suppressed. The semblance quantifies the coherence of signals by analyzing their energy distribution over time and space. According to this definition, a higher semblance value increases the likelihood of detecting a true positive. To identify coherent seismic sources, I compute the maximum semblance over time and apply a detection threshold based on the Median Absolute Deviation (MAD) of this trace. The MAD is a robust measure of variability that allows us to adapt to varying noise levels. Although semblance is primarily a measure of the coherence of pick confidence across stations, it is also influenced by the signal-to-noise ratio (SNR) and data quality. For smaller events, especially those distant from the network, low SNR and fewer picks at some stations tend to reduce semblance values. Conversely, larger events generally produce clearer signals at most stations, resulting in higher semblance values. Therefore, semblance is often indirectly correlated with earthquake magnitude through these effects. To prevent spurious detections due to very few stations, event detection requires a minimum of three stations; windows with fewer stations are discarded.

Magnitude estimation

The framework includes the calculation of local magnitudes using the method as for the PRN catalog (Bakun & Joyner, 1984 [81]). In this case, I consider only stations within 20 km of the epicenter for magnitude estimation. This choice arises because Qseek, by default, includes in the magnitude estimation all stations with SNR above a certain threshold (2 in this case) in the time window of the theoretical

arrival, regardless of whether the station is triggered. For micro-earthquakes, distant stations often record mostly noise, which can be mistaken for seismic signal, leading to systematic overestimation –in comparison to the one of the same event in the reference catalog– of local magnitude. Visual inspection of waveforms from distant stations confirms poor signal quality and random noise spikes causing false coherent signals. Given the network density, sufficient reliable stations are available within 20 km, so excluding more distant stations improves magnitude consistency and reliability for small events. This distance cutoff is not applied to the PRN workflow, which only considers triggered stations, thereby naturally filtering out noisy distant recordings. I verify that this approach results in more reliable local magnitudes for small events, while larger events are unaffected.

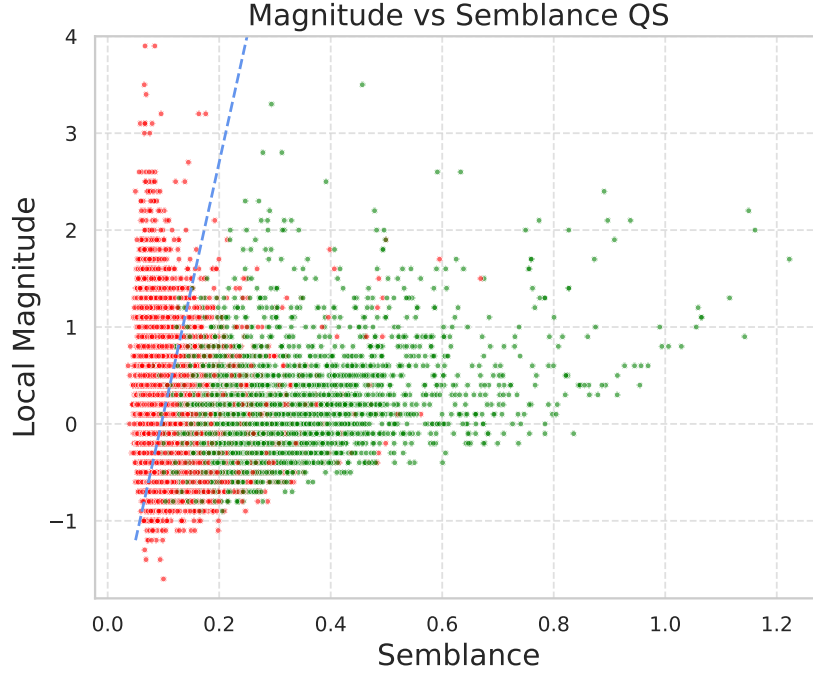
Filtering out False Positives

I follow an approach similar to the PRN workflow to filter false positives and events with low-quality locations. I set the minimum number of picks to 5 to allow a recall aligned with PRN, at the cost of more false positives that will be filtered out later. This approach ensures a coherent comparison between the two methods while maximizing the recall of the reference catalog.

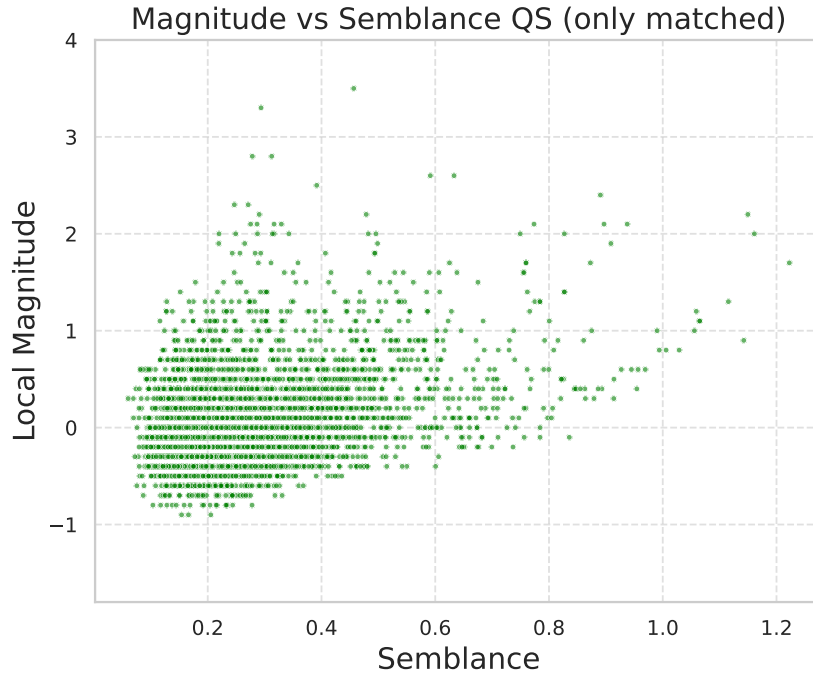
Despite the preliminary filtering applied, the QS catalog still contains a significant number of false positives. These often correspond to events with higher magnitude (e.g., $M > 2$) that are absent from the original catalog but with low semblance. I expect to detect the highest magnitudes with traditional methods too.

Given the size of the catalog, it is impractical to verify each earthquake manually. While a full manual labeling of true and false positives was not performed, a visual inspection of 100 randomly selected events from the initial catalog (consisting of 21,784 detections) was carried out to obtain a rough estimate of precision. Among these, 55 events were visually confirmed as true earthquakes, yielding an estimated precision of approximately 55% for the unfiltered catalog. This relatively low precision is an expected consequence of my deliberate choice to adopt low detection thresholds in the initial stage of the workflow, in order to maximize completeness and ensure that weak but potentially real events are not missed.

I propose an automated approach to filter false positives further. First of all, I compare detections in the QS and reference CMS catalogs and define a match if the origin time difference is less than 1.5 seconds and the epicentral distance is less than 10 km. Events appearing in both catalogs are considered true positives (TP). Those only appearing in QS may either be false positives or new detections. Plotting the two families of earthquakes in a Magnitude-Semblance diagram (Figure 4.5) illustrates how TP (green dots) are clustered within a specific region of the graph, where the semblance correlates with the magnitude. Events with low semblance and high magnitude are believed to be false positives (Figure 4.5a).



(a)



(b)

Figure 4.5: Magnitude vs. QS semblance from the QS catalogs: a) earthquakes not present in the reference catalog are plotted as red dots, together with those that appear in both (green). The blue dashed line marks a possible boundary for low-semblance, high-magnitude dubious events, b) only earthquakes in the QS catalog that appear also in the reference catalog (CMS).

Figure 4.6 shows the Magnitude-Semblance matrix for the QS catalog. The 2D magnitude-semblance space is rendered with pixels. The color scale represents the rate of green dots (ratio between the number of green detections and total detections in each Magnitude-Semblance cell) as a function of magnitude and semblance. The diagram shows only magnitudes equal to or higher than the magnitude of completeness estimated for the CMS catalog (-0.2 with Maximum Curvature method), as we cannot compare those below this threshold.

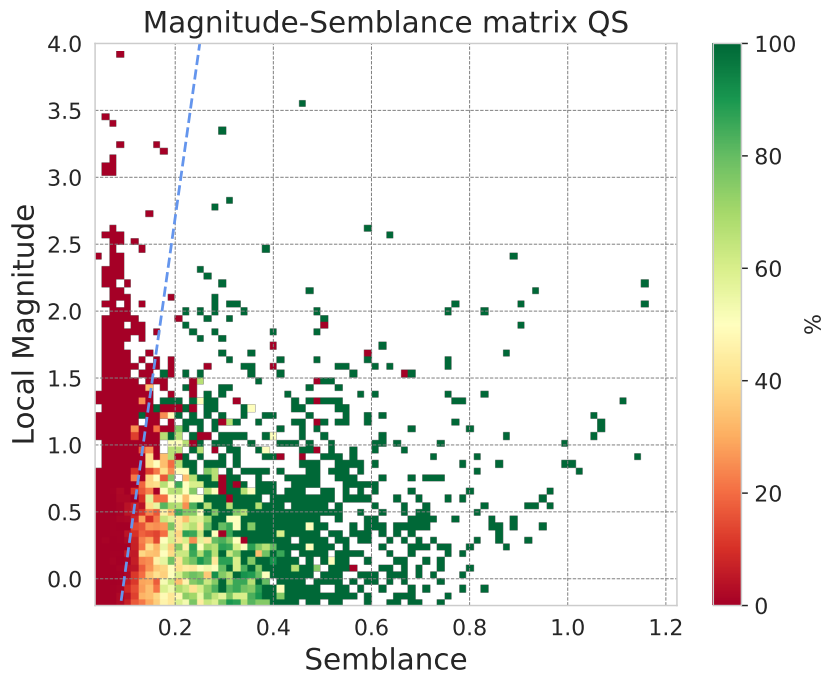
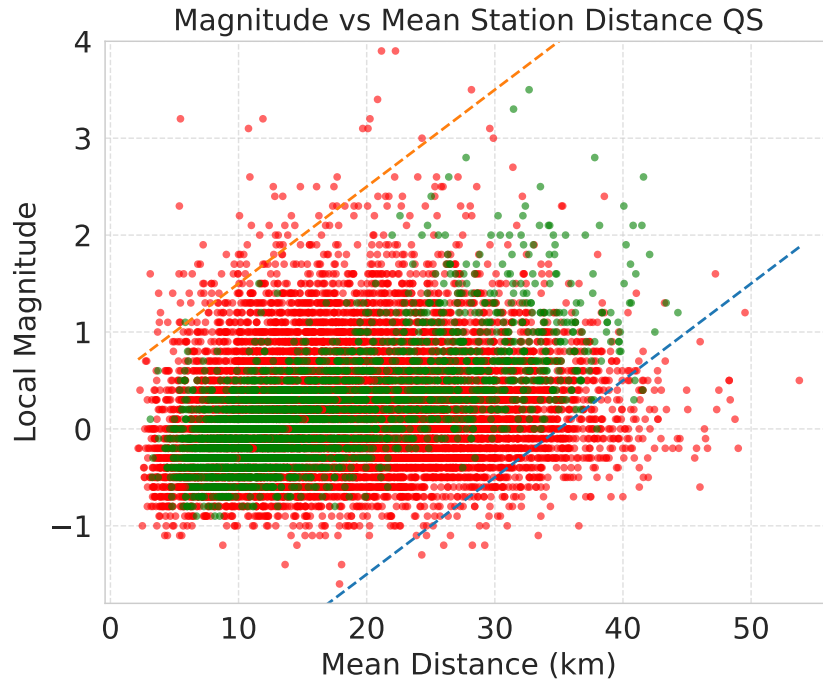


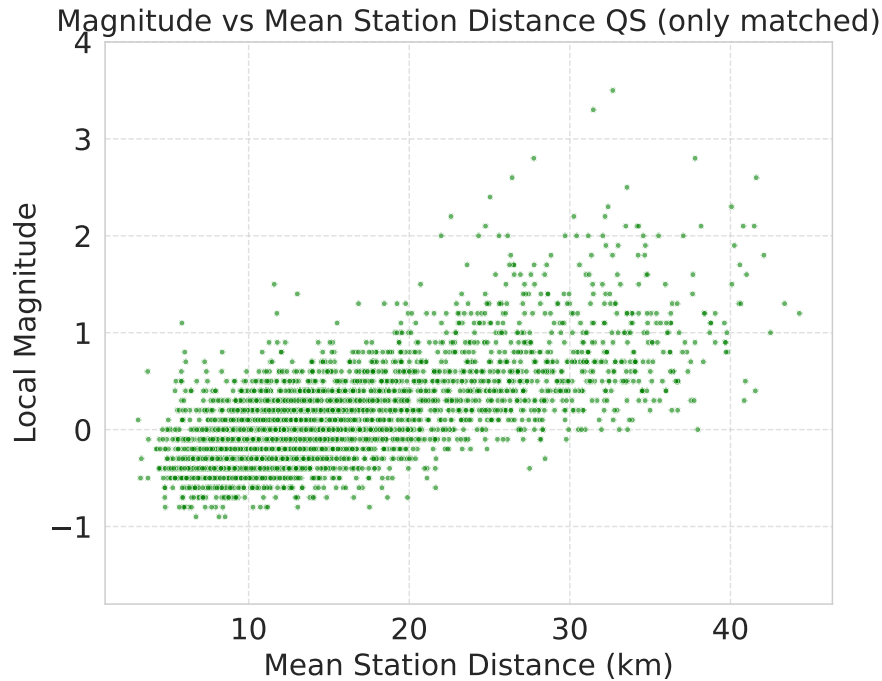
Figure 4.6: Magnitude-Semblance matrix calculated for the QS catalog. The color scale represents the ratio of green dots in each bin.

Based on Figures 4.5 and 4.6, I propose the following empirical magnitude-semblance relation to filter false positives: $M = 26 \times S - 2.5$ (shown in the figure as a blue dashed line) where M is the magnitude and S is the semblance. The two coefficients were determined by grid search, minimizing the proportion of green events to the left of the line and the proportion of red events to its right. A higher weight (0.85) is assigned to the first proportion, as my primary objective is to maximize the recall from the reference catalog. Furthermore, red events are not necessarily false positives, but may also be new detections.

Following an approach similar to Adinolfi et al., (2018)[88], I analyze the relationships between magnitude and average station-to-epicenter distance. Due to the attenuation experienced by the seismic waves with distance, low-magnitude earthquakes are more likely to be recorded by stations nearby. Therefore, detecting small-magnitude events with large average station-to-epicenter distances indicates false detections. Similarly to what I do in Figure 4.5, I determine two empirical relations to cut off false positives in external regions of Figure 4.7a. While the general approach can also be used for other datasets, the obtained empirical relations are specific to the case study.



(a)



(b)

Figure 4.7: a) Magnitude vs average station distance from the epicenter for the QS catalog first output. Green dots indicate the earthquakes found also in the reference CMS catalog, red dots are those that do not match in the reference catalog. The two lines' coefficients were determined by grid search, minimizing the proportion of green events outside the region identified by the lines and the proportion of red events inside. b) same plot but with green dots only for a clearer visualization.

After applying the magnitude–semblance filter, the catalog size was reduced to 11,399 events. This significant reduction, while discarding nearly half of the initial detections, is expected to improve the overall quality of the catalog by removing events with low signal coherence or insufficient magnitude. I repeat the visual check on 100 randomly selected samples. At this stage 78 events are confirmed, giving an estimated precision of 78%. The filtering step was therefore effective in balancing completeness and reliability, as confirmed by the substantial decrease in likely false positives and the retention of a large number of high-quality detections. The final value reflects the trade-off I chose between completeness and precision.

I applied the same analysis to the PRN catalog for consistency. Both red and green dots are clustered within the region of the graph where the mean interstation distance correlates with the magnitude (Figure 4.8).

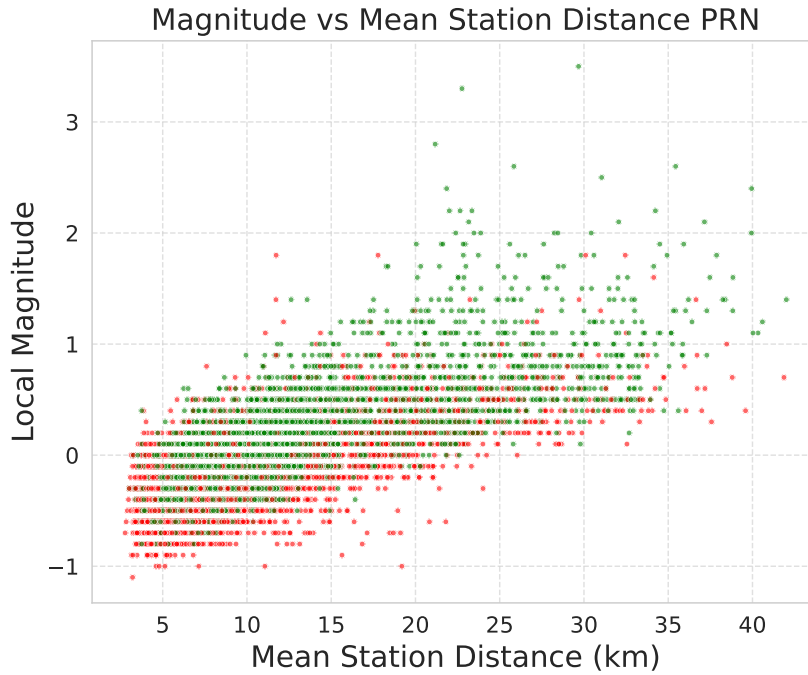


Figure 4.8: Magnitude vs average station distance from the epicenter for the PRN catalog. Green dots indicate the earthquakes found both in PRN and in the reference CMS catalog, red dots are those that do not match in the reference catalog.

Figures 4.9 show examples of waveforms corresponding to low-magnitude earth-

quakes that passed the filters.

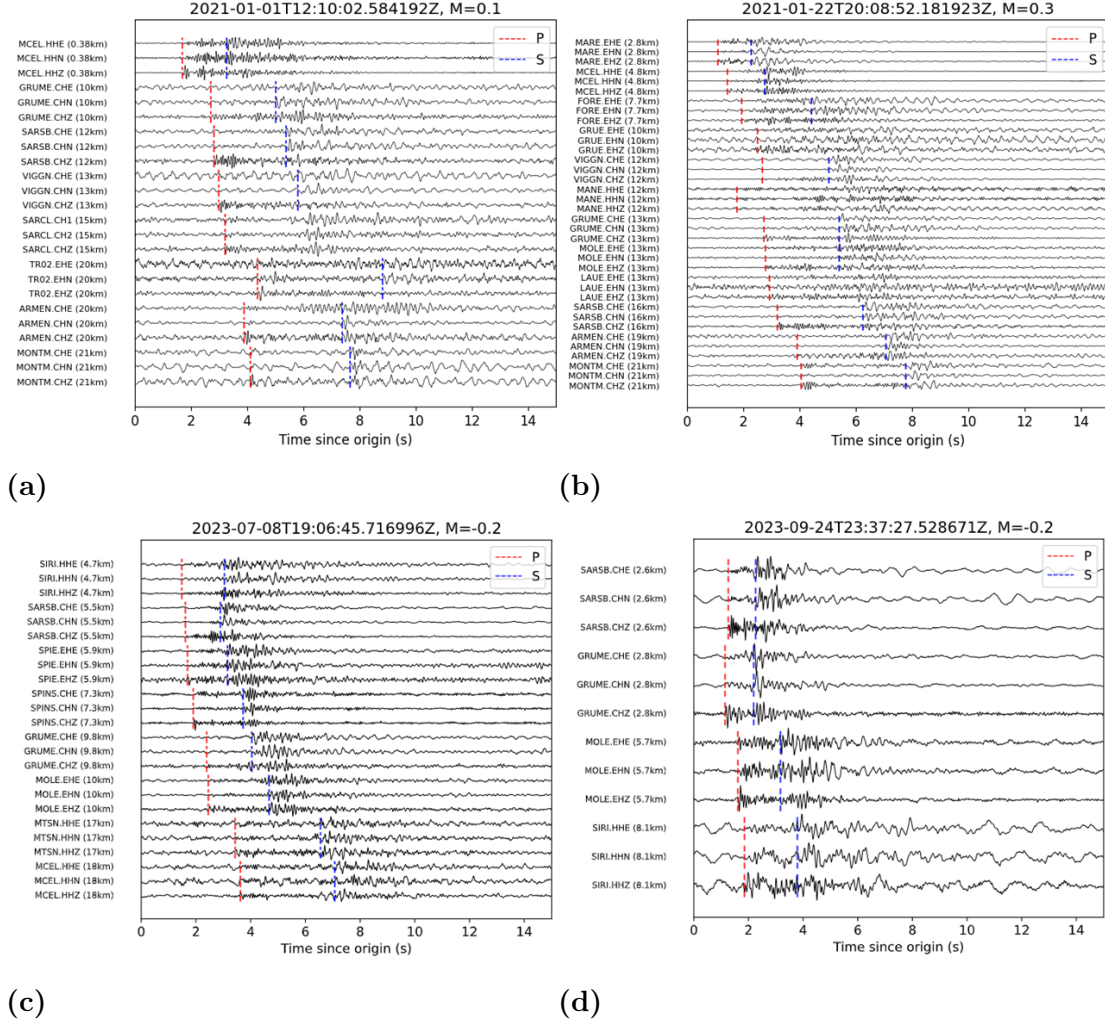


Figure 4.9: Waveform examples of low-magnitude earthquakes that passed the magnitude vs semblance and magnitude vs interstation-distance filter in the QS workflow described in the Methods section.

4.4 Results

The PRN catalog consists of 9,618 events, with a recall rate of 93%. The QS catalog consists of 11,399 earthquakes, with a recall rate of 77%. Figure 4.10 shows the size of the intersections (recall) between the catalogs and the new detections.

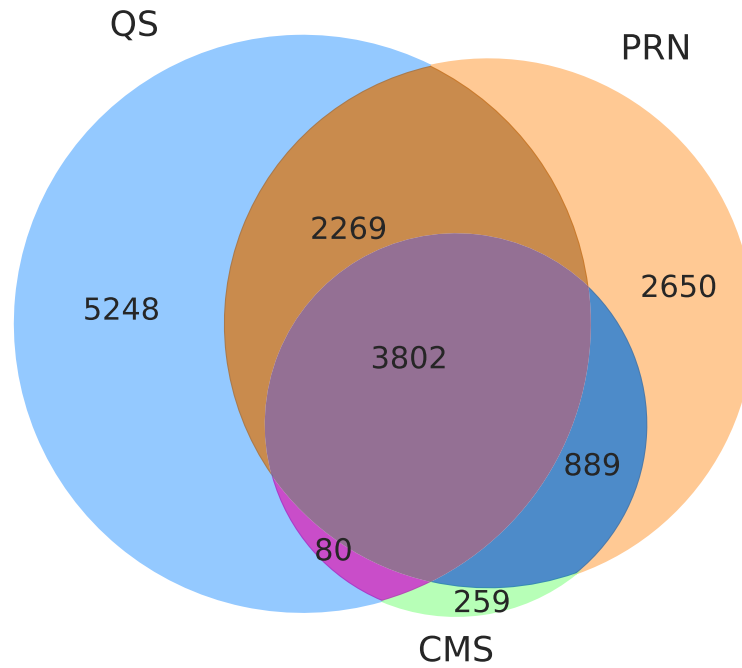


Figure 4.10: Venn diagram of the three catalogs and their intersections. Green: reference CMS catalog, orange: PRN catalog, blue: QS catalog.

To assess the consistency between the catalogs, I analyze the time and locations of detected earthquakes, identifying matching patterns of epicenter clustering. Table

4.1 shows the median of the differences in origin time (Δt_o), epicentral location (Δloc_{epi}), hypocentral depth ($\Delta depth$), and local magnitude (ΔM_L) distributions, computed for each intersection between the catalogs.

Table 4.1: Median difference in origin time (Δt_o), epicentral location distance (Δloc_{epi}), hypocentral depth ($\Delta depth$), and local magnitude (ΔM_L) for same-event determinations in all pairs of catalogs.

<i>Catalog</i>	Δt_o	Δloc_{epi}	$\Delta depth$	ΔM_L
PRN-CMS	0.546 s	1.01 km	-0.18 km	-0.1
QS-CMS	0.515 s	1.34 km	-0.05 km	-0.1
QS-PRN	-0.016 s	1.01 km	0.09 km	-0.1

To assess location quality, I computed the median uncertainty and the 90th percentile for horizontal and vertical errors in each catalog. For the horizontal location error, the CMS catalog has a median of 0.4 km, with 90% of events located with uncertainty < 1.1 km; PRN has a median of 1.1 km (90% < 3.2 km), and QS a median of 0.2 km (90% < 0.4 km). Regarding the vertical uncertainty, CMS and QS both show a median of 0.5 km (90% < 1.1 km both), while PRN has a median of 0.9 km (90% < 2.1 km). The comparison between catalogs is also illustrated by the distribution of hypocenters with depth in the cross-sections of Figures 4.1, 4.11a, 4.11b, while figure 4.10 shows the size of the intersections (recall) between the catalogs and the new detections. The PRN catalog allows the identification of structures where seismicity is clustered, due to the high quality of NonLinLoc locations. Most earthquakes are concentrated at the southwestern side of the basin (Monti della Maddalena Fault System) within the upper 10 km of the crust. At the same time, fewer events occur along the northeastern side, with greater depths.

The roughly oval shape of the epicentral distribution delimits the region where the network sensitivity is optimal (Morelli et al., 2025 [82]), and the domain of attention of approximately 3800 km², where the detection and location exercise has been limited (distributed seismicity extends further out). In agreement with the setup of the reference CMS catalog, I also exclude earthquakes with hypocenter depths larger than 23 km to focus on the shallow crustal domain of interest. New locations are primarily found in the same active region, where the new catalogs achieve detection of weaker earthquakes (Figure 4.12).

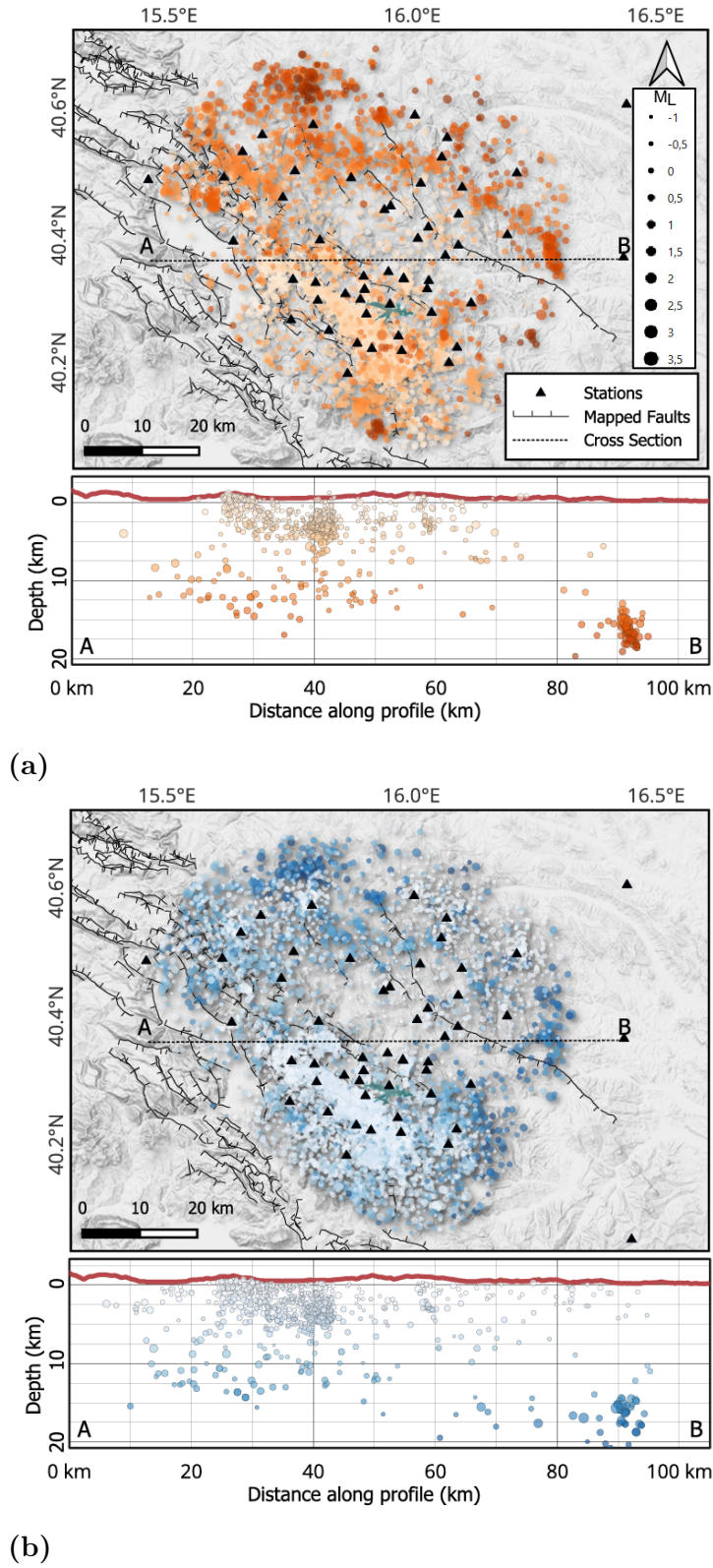


Figure 4.11: Distribution of epicenters and relative vertical cross sections for a) PRN catalog, b) QS catalog. Vertical cross sections represent a sliver with a thickness of 5 km on each side. The dot size is proportional to Magnitude, while the color scales with the depth. Fault traces from the dataset by Lavecchia et al., 2023 [70].

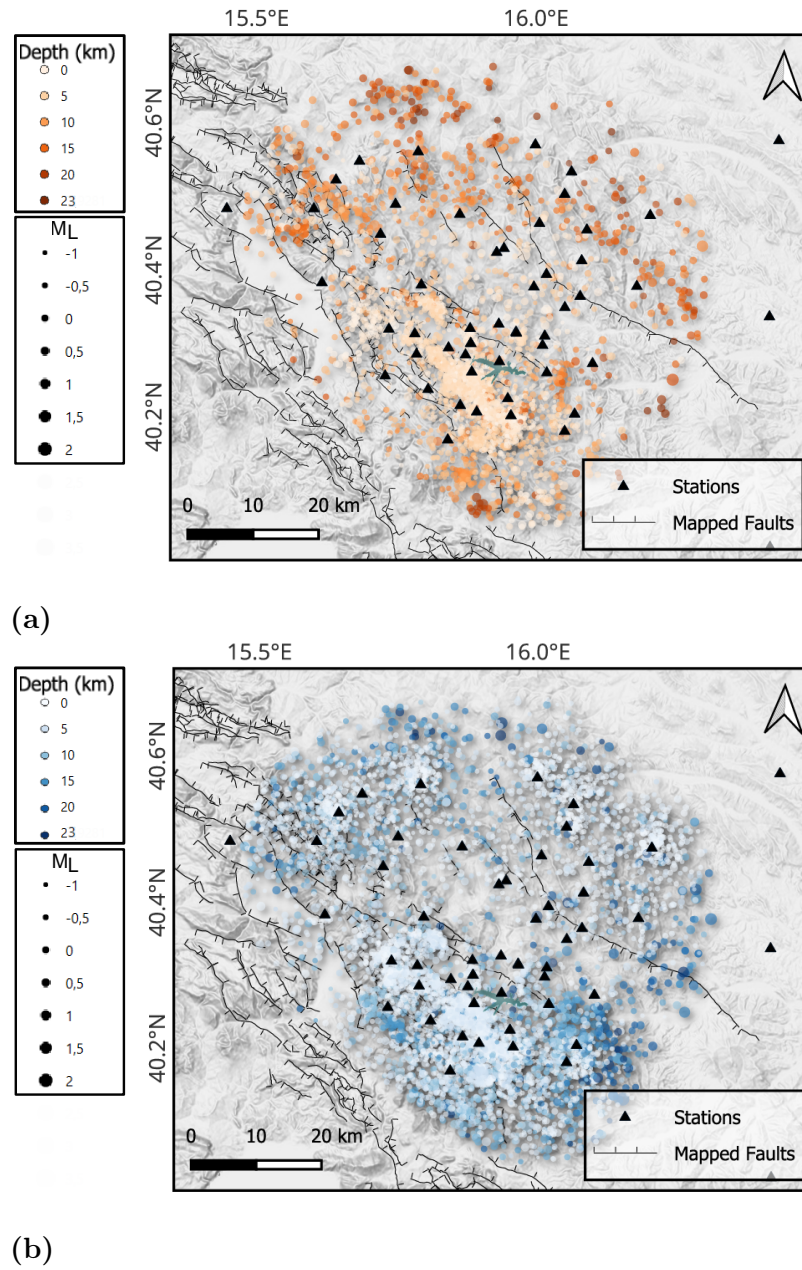


Figure 4.12: a) Epicenter distribution of earthquakes only present in the new PRN catalog and not in the reference CMS bulletin. b) Epicenter distribution of earthquakes only present in the new QS catalog and not in the reference CMS bulletin.

The normalized cumulative number of earthquakes in Figure 4.13 shows similar temporal evolutions of the three catalogs I compare. QS and PRN catalogs have comparable trends throughout the period considered, reflecting natural fluctuations

in time of the seismic occurrence rate (note that cumulative distributions are all normalized to 1). The reference catalog's different time evolution (slightly flatter than QS and PRN at around 2021/09 and steeper at around 2023/05) might be the possible effect of parameter adjustments and optimization in the supervised monitoring routine.

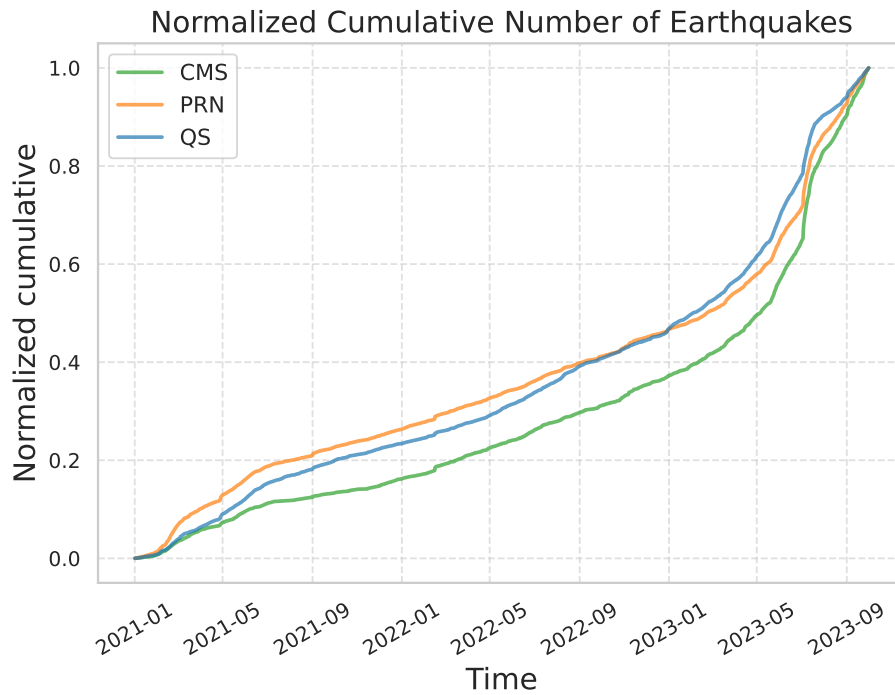


Figure 4.13: Normalized cumulative number of events as a function of time, compared for the three catalogs. Green: reference CMS catalog, orange: PRN catalog, blue: QS catalog.

Figure 4.14 shows the timeline of detected events and their magnitudes across the three catalogs. Overall, the magnitude estimates and their temporal evolution are consistent among the catalogs. For larger events, the magnitude estimates remain in good agreement. The main differences appear at lower magnitudes, where both QS and PRN detect additional smaller events, effectively lowering the detection threshold. This difference becomes less significant over time, as a change in operational procedures has improved the detection threshold in the reference CMS catalog.

Figure 4.14 shows that the rise in the number of events observed in 2023, along

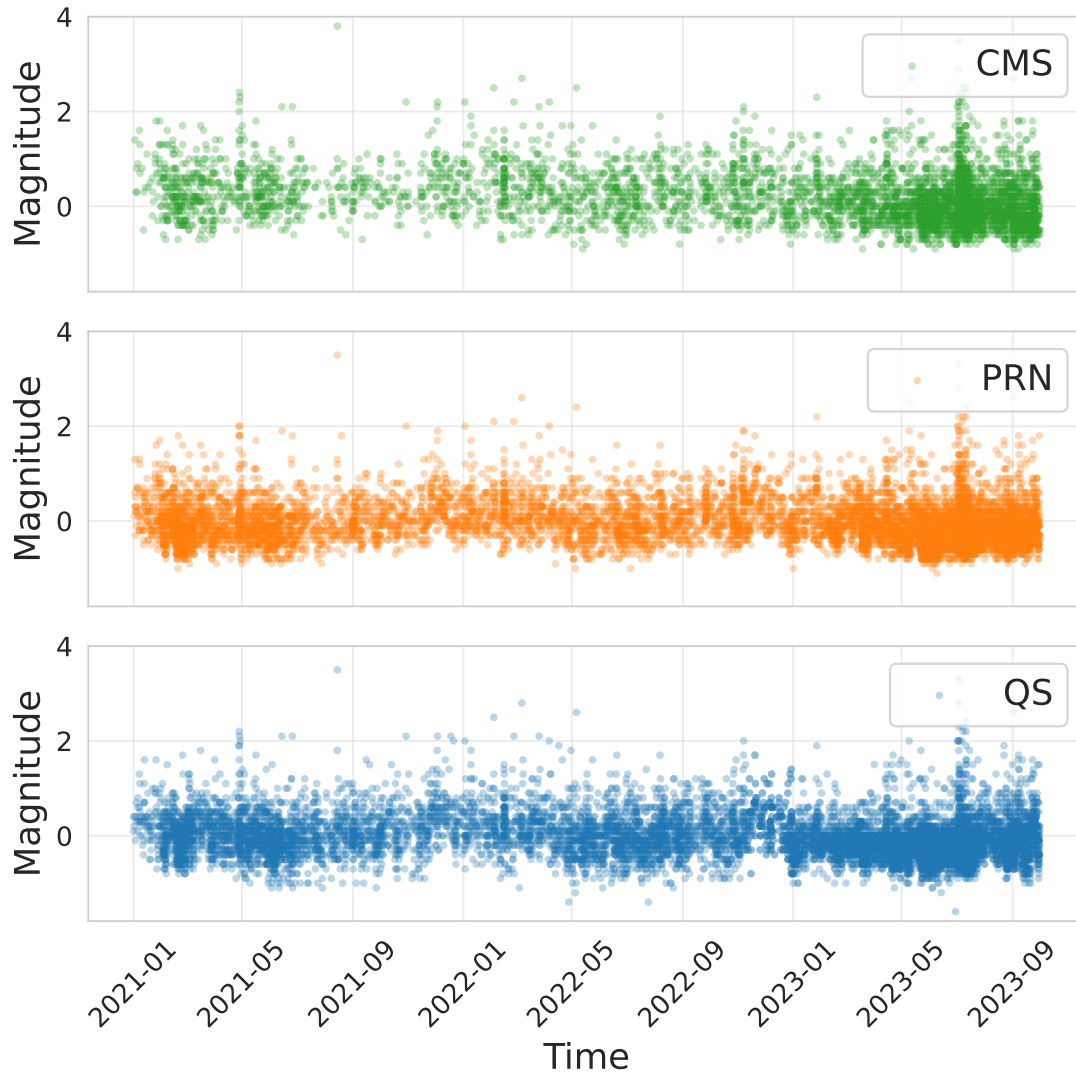


Figure 4.14: Distribution of magnitude of detected events in the three earthquake catalogs as a function of time. Green: reference CMS catalog, orange: PRN catalog, blue: QS catalog.

with larger magnitudes, reflects a general increase in seismic activity consistently captured by all three catalogs.

The frequency histogram shown in Figure 4.15 shows the distribution of local magnitude for the three catalogs. Similar to Figure 4.14, this demonstrates the improved performance of both new catalogs in the magnitude range -1 to 0.7. The discrepancy in few unit counts at larger magnitude is due to (slightly) different magnitude estimates (as pointed out in the next Section) and to few events

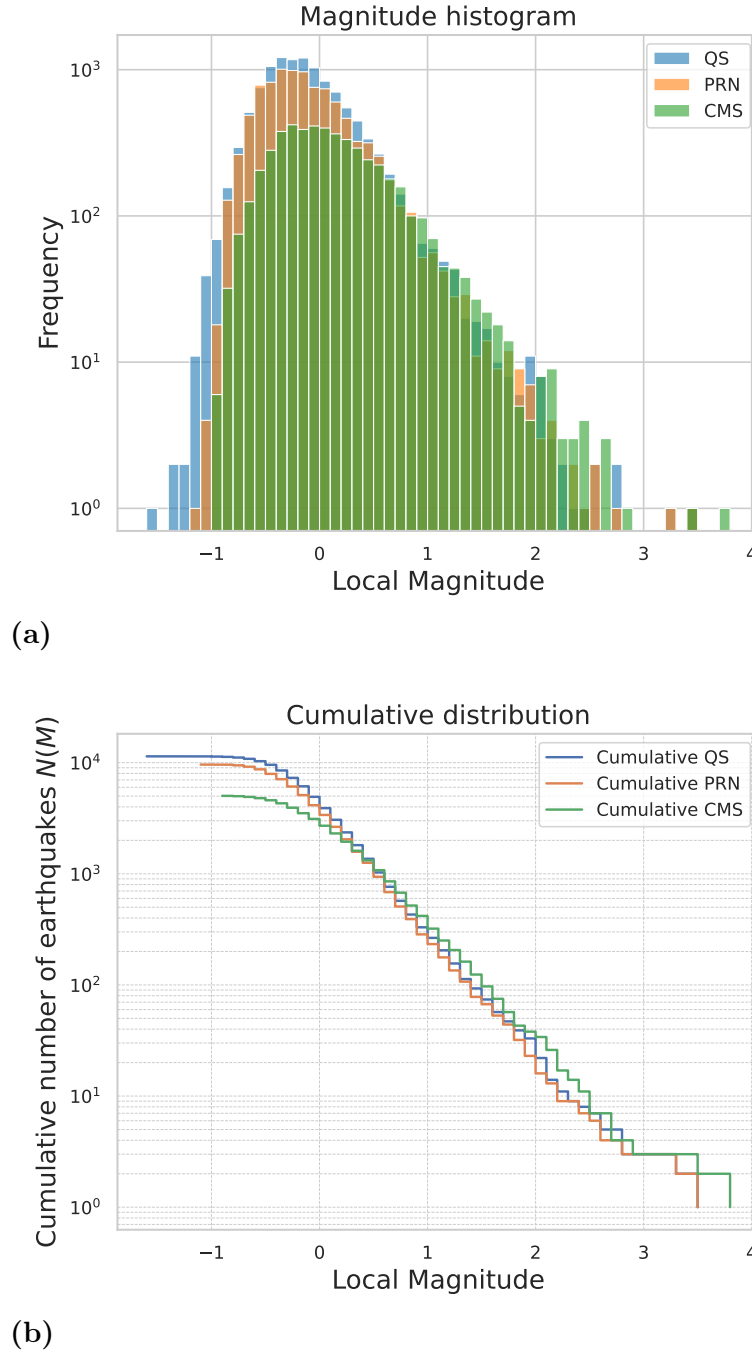


Figure 4.15: a) Frequency histogram and b) cumulative distributions of the magnitude estimations for the three catalogs. Green: reference CMS catalog, orange: PRN catalog, blue: QS catalog.

occurring near the border of the study area, which in different workflows may be located inside or outside the study area (and be therefore excluded in one or another of the catalogs because of slightly different locations).

4.5 Discussion

4.5.1 Comparison of the workflows' performance

Both machine learning-based methods exhibit remarkable efficiency in automatically monitoring the study area. The PRN method exhibits higher recall (93%) from the reference CMS catalog with accurate earthquake locations, resulting in better delineated seismic clusters and lineaments, as evidenced in map plots and cross sections (Figures 4.11 and 4.12). Although the absolute uncertainties from NonLinLoc are slightly larger, they provide a more reliable estimate of location accuracy due to the probabilistic nature of the algorithm, even when using a 1D velocity model. In contrast, the CMS catalog truncates hypocentral coordinates to fewer decimal places, reducing spatial resolution and limiting the ability to resolve small-scale structures illuminated by seismicity. It is worth noting that part of the location uncertainty is inherently related to the adopted velocity model. In this study, a 1D velocity model was used to speed up the automatic processing. As a consequence, uncertainties in the velocity structure — particularly unmodeled lateral heterogeneities — may affect absolute hypocentral parameters, especially event depth. However, these limitations do not compromise the primary goal of the workflow, which is to enhance earthquake detection and catalog completeness. The resulting catalog provides a reliable basis for subsequent refinement through dedicated relocation procedures. For this reason, high-precision relocation and detailed structural interpretation are addressed in later chapters, where the improved catalog is used as input. The normalized cumulative event plots (Fig. 4.13) show similar trends. However, during the first period (before September 2021), the PRN and QS catalogs had a more rapid increase in the earthquake cumulative number than observed in the reference CMS detections. As the processing parameters for PRN and QS do not change with time, the elevated event rates before September 2021 may be due to increased small-magnitude seismicity that went unnoticed

by the standard processing. This is shown in Figure 4.14, where many minor earthquakes are missing in the reference CMS catalog. Over time, this difference has become less pronounced due to a change in operational procedures, resulting in an improved detection threshold for the CMS. The supervised CMS routine has a sharp increase in event rate in August 2023 (Figure 4.14). This highlights an additional advantage of employing an efficient, automatic, and stable method for re-processing an entire period, ensuring consistency and reducing potential biases introduced by procedural changes over time.

I calculate local magnitudes using the same attenuation model as in the CMS catalog, but some small differences in the estimation remain. This explains the minor discrepancies in Figure 4.15. The same event can fall into two different histogram bins according to the catalog we are considering. This is hardly noticeable in more populated bins, but accounts for differences in bins with population on the scale of units. I attribute this to variance in the number of phases detected (and hence used for magnitude computation) for the same event in the three catalogs: PRN and QS events have more detected phases. This can be due to PhaseNet’s higher sensitivity to finding more arrivals with lower amplitude than the original method. However, a more detailed comparison in magnitude estimation is outside the scope of this study. Regardless of these minor discrepancies in the histogram bins, we observe a higher number of earthquakes than the original catalog at low magnitude (between M_L -1 and 0.7), while preserving similarity between the two distributions at higher magnitudes. This outcome aligns precisely with my objectives: enhancing the monitoring sensitivity to lower magnitudes, expecting similar results at higher magnitudes, thereby confirming the method’s reliability.

4.5.2 Noise and performance of the detection

The distribution in time of the magnitude of earthquakes (Figure 4.14) shows coherent detection of the largest earthquakes as expected, and similar cyclic seasonal variations of the minimum magnitude detected in all the earthquake catalogs. This cyclicity is likely due to noise levels’ seasonality, which can increase or reduce the SNR. The periods of lowest magnitude detected coincide with the summers, during which we expect lower noise levels related to calmer weather

conditions. Working in the frequency range 1 - 45 *Hz*, we expect the contribution of sea noise and wind, which is likely a major contributor to noise at the highest frequencies. While in the range 1 - 10 *Hz* we expect the main contribution of anthropogenic noise. Both natural and anthropogenic noise sources, therefore, appear relevant in limiting seismic detection completeness. To analyse the possible correlation between the lowest magnitude detected and seismic noise, I calculate monthly PPSDs (Probabilistic Power Spectral Densities; McNamara & Boaz, 2005 [89]) at all stations. For each PPSD, I consider the median value in different frequency ranges. Many stations show a clear seasonality correlated with the variation of magnitude completeness, especially towards the lower end of the investigated frequency band. Figure 4.16 shows a station (GE.MARCO) exhibiting significant correlation between the PPSD median, computed in the frequency range 1 - 2 *Hz*, and the seasonal variation of the minimum magnitude in the PRN catalog (Pearson coefficient (r) ~ 0.7 , p-value $\ll 0.05$), QS catalog ($r \sim 0.8$, p-value $\ll 0.05$). CMS catalog exhibits a moderate correlation ($r \sim 0.4$, p-value = 0.018), remarking that the detection performance was not determined exclusively by the seasonal noise level.

A similar trend is found when I focus on the hourly variation. To investigate the influence of anthropogenic noise on this dataset I study the variation of the average minimum magnitude detected at different hours of the day: I found a diurnal oscillation with the highest value of such minimum detected magnitude (i.e., worst sensitivity, due to higher noise levels) at about 12 PM, and lowest values (best detection performance) detected around midnight. This diurnal oscillation well correlates with one of the median PPSD as shown in Figure 4.17 for the same example station GE.MARCO (PRN: $r \sim 0.97$, p-value $\ll 0.05$; QS: $r \sim 0.91$, p-value $\ll 0.05$; CMS: $r \sim 0.95$, p-value $\ll 0.05$). The hourly median PPSD value is calculated by averaging the values over almost three years. In this case, all the catalogs are highly impacted by the noise diurnal oscillation.

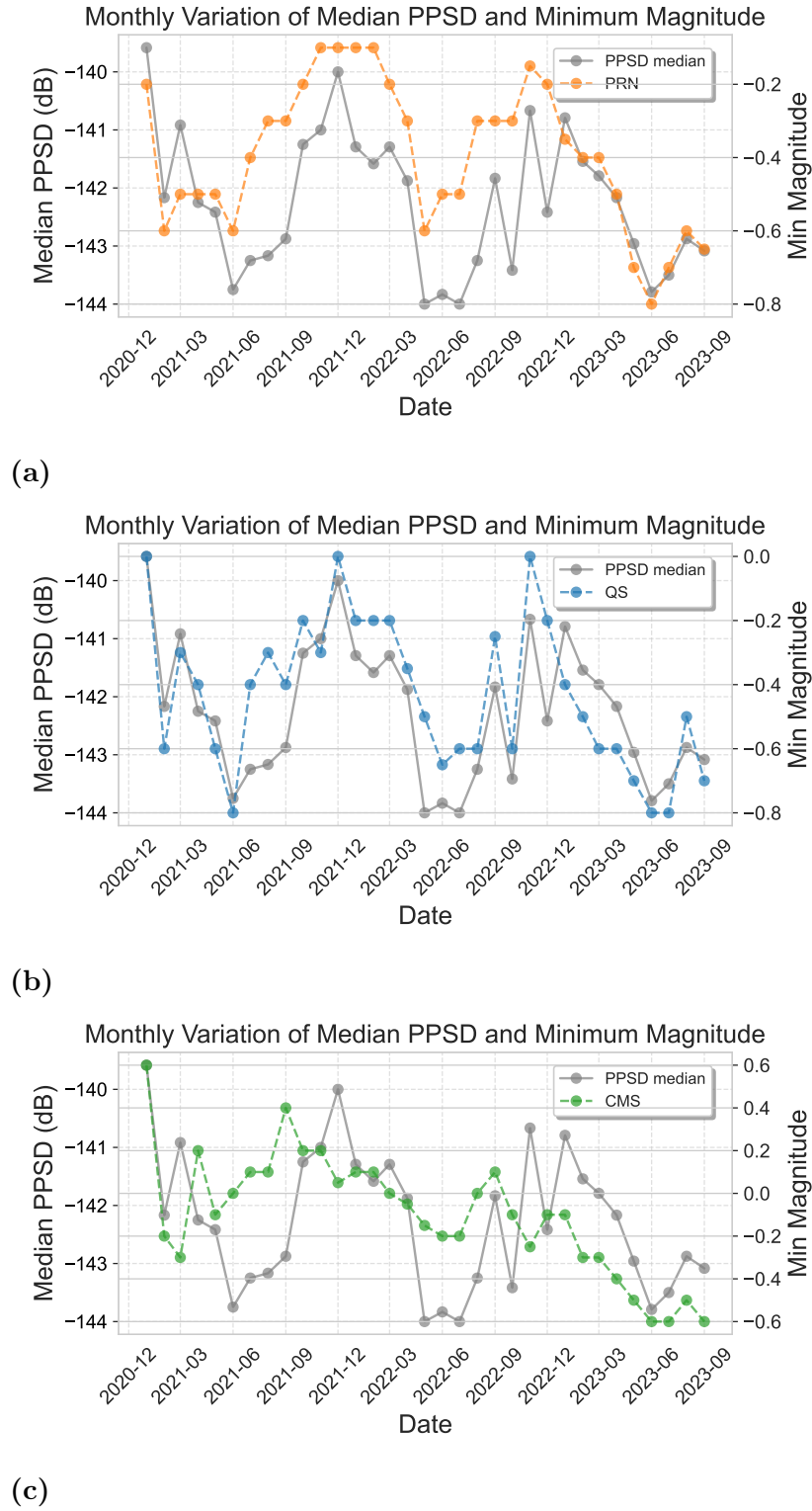


Figure 4.16: a) Monthly and PPSD 50th percentile on a sample station (GE.MARCO) averaged over the three components (E, N, Z) vs minimum magnitude detected day by day in a) PRN catalog, b) QS catalog, c) reference CMS catalog. It is averaged over each month.

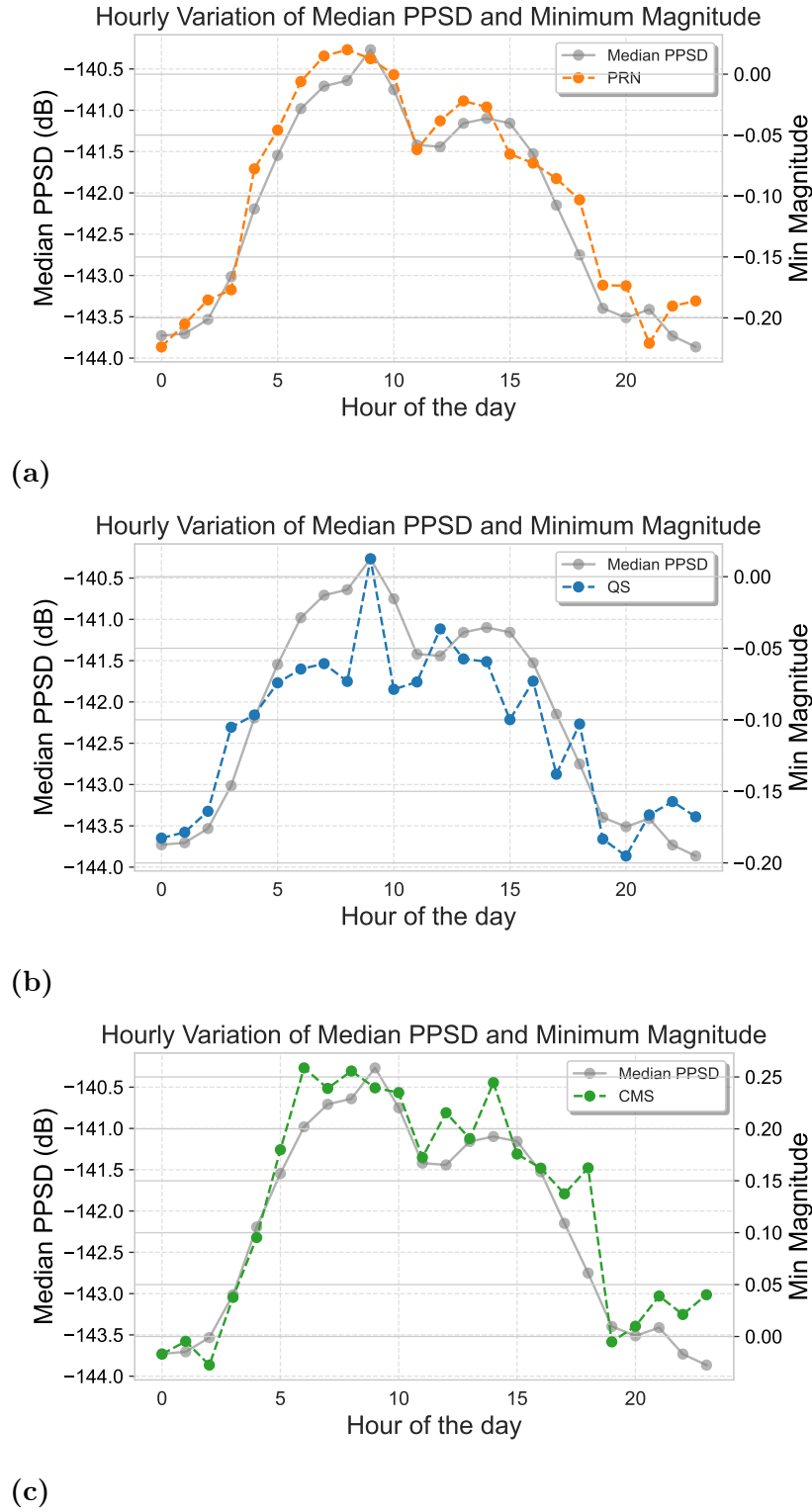


Figure 4.17: Hourly PPSD 50th percentile on a sample station (GE.MARCO) averaged over the three components (E, N, Z) vs minimum magnitude detected day by day in a) PRN catalog, b) QS catalog, c) CMS reference catalog. It is averaged over each hour.

In the case of the CMS catalog, changes in detection parameters over time have influenced the performance of the seismic detection system, affecting the completeness level. Consequently, the correlation between the minimum magnitude and the median PPSD is moderate ($r \sim 0.4$) when considering seasonal variations over the three-year period, while it is significantly higher ($r \sim 0.95$) when computed on the average hourly scale. This suggests that the hourly analysis, being based on a global average, smooths out temporal inconsistencies in detection performance, whereas the seasonal analysis captures their impact more clearly.

4.5.3 Pertusillo water level and seismicity rate

With almost three years of data available, we may address the possible long-term modulation of the seismicity rate, in response to the seasonal changes of the water level at the Pertusillo artificial reservoir (Valoroso et al., 2009 [31]; Stabile et al., 2014 [32]). The reservoir has an average water depth of ~ 95 m spanning an area of ~ 75 km², experiencing significant and rapid seasonal water-level fluctuations (Valoroso et al., 2009 [31]). In such reservoirs, changing water levels change the stress within the surrounding rocks. During the loading phase, the fluid pressure within rock pores rises as the water level increases, which reduces the effective normal stress acting on fractures and faults. This stress can promote fault slip and increased seismicity. Conversely, during the unloading phase, when the water level drops, fluid pressure decreases, effective stress increases, and the seismicity rate tends to decline (e.g., Talwani, 1997 [33]; Gupta et al., 1972 [90]). Berbellini et al., (2021) [91] found a significant correlation between relative seismic velocity variations and the hydrometric level of the nearby Agri river feeding the Pertusillo reservoir – considered a proxy for the total water storage in the shallow aquifers. The poroelastic effect likely governs the observed variations in seismicity rates, as fluctuations in water levels are mostly positively correlated with seismicity. Valoroso et al., (2023) [15] points to such cyclicity over 13 years using a coarse reference catalog, while their local 13-month-long dataset does not allow for studying yearly cycles. Telesca et al., (2025) [92] analyzed the temporal clustering of seismicity and its correlation with reservoir water level variations in the Pertusillo reservoir area using fractal and spectral analysis methods. Their findings suggest a delayed response of approximately 1 month, attributed to pore-pressure diffusion,

supporting the hypothesis of protracted reservoir-triggered seismicity.

To verify such seasonal modulation in this dataset, I investigate a subset of the catalogs, limited to the region of interest closest to the lake (longitude 15.82-16.00, latitude 40.2-40.31), and compute the daily seismicity rate. The selected area is indicated by the black rectangle in Figure 4.1. To exclude noise seasonality effects on the seismicity rate, I consider only the events above the CMS magnitude of completeness $M_c = -0.2$ (the highest one), calculated with MaxCurvature method. In Figure 4.18 I compare such seismicity rate with the water level in the reservoir (downloaded from the district basin authority website: <http://www.adb.basilicata.it>).

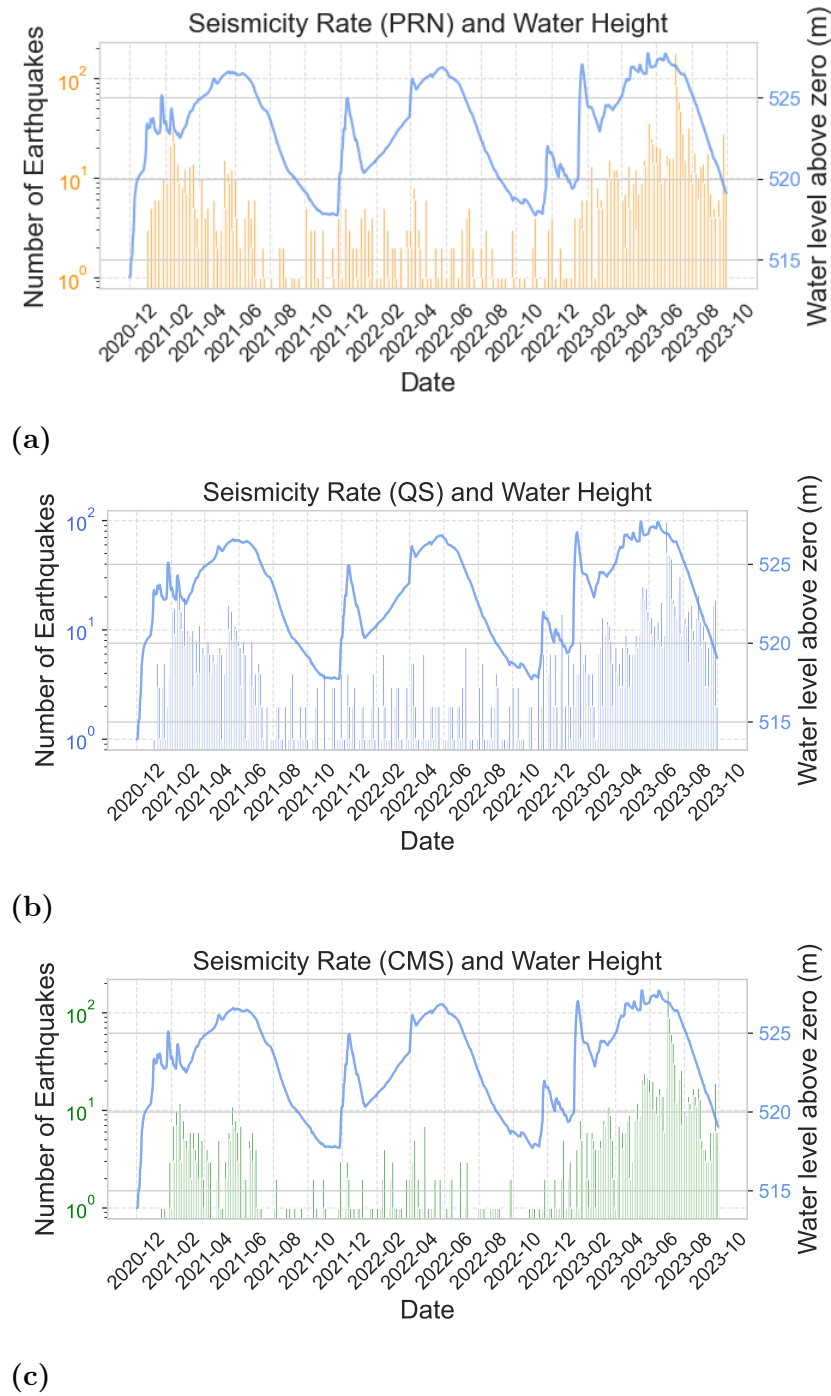


Figure 4.18: Pertusillo lake water level over time (blue) vs number of earthquakes every 3 days (i.e., seismicity rate) in the a) PRN catalog (orange), b) QS catalog (blue), c) CMS catalog (green). Water level data starts in December 2020, considering the delayed effect of seismicity.

We observe a local increase in seismicity when the reservoir water level rises, and a corresponding decrease in seismicity when the water level falls, which is more notable in 2021 and 2023. The seasonal increase in seismicity occurs with a delay of some weeks. Notably, in 2023, both the highest water levels and the highest seismicity rate were recorded. In contrast, in 2022, we noticed a slight increase in seismicity. Still, the number of earthquakes in the selected area remained generally low, nearly comparable to the background seismicity, even when the reservoir was full. In conclusion, the analysis of these (DL-based) more complete seismic catalogs confirms the seasonal evolution of the local seismicity hypothesized by Valoroso et al., (2023) [15].

4.6 Conclusions

Seismology is an excellent and promising field for the application of machine-learning (ML) methods for several purposes (e.g., Mousavi & Beroza, 2022 [93]; Mousavi & Beroza, 2023 [94]). Many applications have shown the ability to carry out complex tasks effectively and efficiently, so favourable operational developments are to be expected. Here, I have used the PhaseNet deep neural network seismic arrival time picking method (Zhu & Beroza, 2019[2]), combined into two different workflows to analyse micro-seismicity recorded by a heterogeneous, integrated seismograph network in the tectonically active Val d’Agri region (Southern Italy), also seat of oil and gas extraction as well as artificial water impoundment. Compared to a more customary reference practice based on an STA/LTA detector and manual operator revision, the ML-based workflows produced rich earthquake catalogs with about twice the number of events and twice the detected arrivals per event (a larger number of phases to be used in the hypocentral location is highly beneficial for its quality). As illustrated in the Methods section, I completed the workflows using an efficient method to discard false positives.

I chose a rather conservative detector setup that reproduces most of the events in the reference catalog (PRN: recall rate 93%; QS: 77%) while discarding many false positives and uncertain events.

The benefit of the automatically generated catalogs is apparent. Once the significant parameters (such as the minimum number of detected phases to allow the definition

of an event, and similar criteria) are set, the workflow can be applied automatically to real-time data. In a real-time/near real-time implementation, continuous data streams could be provided through real time data transmission systems (e.g., a SeedLink server), which provide the miniSEED data. This would enable the automatic picking stage to run on streaming waveforms with minimal latency due to the PhaseNet processing window (30 s) and by the additional computational time needed for event association and localization. Besides the apparent advantage of a high saving of operator time, using an automatic workflow also contributes to the homogeneity of results, as with the wish of any different parameter value, an entire catalog can be recomputed on the fly.

By increasing the sensitivity of the event detection process, we may more easily see the influence of time variations on the population of earthquakes. Furthermore, we can detect the impact of the seasonal variations of high-frequency background seismic noise connected with weather, and we distinctly see a correlation of micro-earthquake occurrence with the charge-discharge cycles of the Pertusillo water reservoir. Although these observations suggest a correlation, further investigation is required to fully understand the relationship between water level fluctuations and seismicity rates. Using the PRN catalog, which demonstrates the best performances, future studies will focus on a more detailed analysis of the temporal and spatial patterns of seismic activity in response to changes in water levels, particularly considering poroelastic effects and delayed responses. This upcoming research will aim to quantify these correlations more rigorously and provide deeper insights into the mechanisms driving the observed seismic behavior.

Data and code availability

I used open seismic data from the following seismic networks: IV (Istituto Nazionale di Geofisica e Vulcanologia – INGV, 2005, [doi:10.13127/SD/X0FXNH7QFY](https://doi.org/10.13127/SD/X0FXNH7QFY)), VD (CNR IMAA Consiglio Nazionale delle Ricerche - Italy, 2019, <https://doi.org/10.7914/SN/VD>, Stabile et al., 2020), IX (Irpinia Seismic Network – ISNet, 2005) and GE (GEOFON Data Centre, 1993, <https://doi.org/10.14470/TR560404>). Waveforms and metadata can be accessed through the Federation of Digital Seismograph Networks Web services at GEOFON (<https://geofon.gfz.de/waveform/>

[webservices/](#)) and INGV (https://terremoti.ingv.it/webservices_and_software).

Seismic data from networks VA (ENI SpA, 2001, <https://doi.org/10.7914/SN/VA>) and TP (TOTAL E&P Italia Spa, 2018, <https://www.fdsn.org/networks/detail/TP/>) must be requested from the operators.

Software PhaseNet is published on GitHub <https://github.com/AI4EPS/PhaseNet>.

Software REAL is available at <https://github.com/Dal-mzhang/REAL>. Soft-

ware NonLinLoc is available at <https://github.com/ut-beg-texnet/NonLinLoc>.

Qseek is published on GitHub <https://github.com/pyrocko/qseek>.

I used Obspy [95] and Pyrocko [96] libraries for data analysis. Plots have been generated using Matplotlib [97], and maps using QGIS.

The PRN and QS catalogs are available on Zenodo.org <https://doi.org/10.5281/zenodo.17123383> [98].

Chapter 5

Magnitude of completeness and b-value

5.1 Introduction

Quantitative analysis of seismicity requires a careful assessment of catalog completeness and magnitude-frequency distribution. The empirical law for the magnitude frequency distribution (FMD) suggests an exponential relation:

$$Prob(Magnitude > M) \propto e^{-\beta M} \quad (5.1)$$

which is often represented by the formulation expressed by Gutenberg and Richter (1944) [99]:

$$\log_{10} N(M) = a - bM \quad (5.2)$$

where $N(M)$ is the number of earthquakes with magnitude above M and the b coefficient is known as *b-value*. It corresponds to the slope of the $\log_{10} N(M)$ vs. M plot and quantifies the relative proportion of small to large earthquakes, offering insights into the stress regime, faulting style, and possible temporal variations preceding large events. Another fundamental parameter in this context is the magnitude of completeness (M_c). It defines the lowest magnitude above which the earthquake catalog can be considered statistically complete, meaning that all occurring earthquakes above that magnitude are detected.

Over the past decades, numerous methods have been developed to estimate M_c , including visual inspection of magnitude-frequency distributions, goodness-of-fit tests, and statistical approaches such as the Maximum Curvature (MAXC) and the Entire-Magnitude-Range (EMR) methods (e.g., Woessner & Wiemer, 2005 [100]). The Max Curvature (MaxC) technique is an empirical method that identifies M_c as the magnitude at which the frequency–magnitude distribution histogram exhibits its maximum slope (the point of highest curvature) (Wiemer & Wyss, 2000) [101]. This approach is straightforward and does not require assumptions about detection probabilities. However, it tends to underestimate M_c , as the curvature maximum often occurs at slightly lower magnitudes than the true completeness level, due to the partial detection of smaller events. The choice of method can significantly affect the resulting M_c value, especially in heterogeneous or evolving datasets, where completeness may vary in time and space due to changes in network configuration, detection capability, or noise conditions.

Similarly, the estimation of the b -value requires robust statistical techniques, as it is sensitive both to quality and completeness of the data. The most commonly used method was introduced by Aki (1965) [102] as a maximum likelihood method. This approach, commonly used in statistics, estimates the parameters of a distribution by maximizing the likelihood function to make the observed data as probable as possible under the assumed statistical model (Rossi, 2018 [103]). Assuming the Gutenberg and Richter distribution for the magnitude, the b -value maximum likelihood estimation (MLE) and its standard deviation σ_b can be written as (Aki, 1965 [102]):

$$b = \frac{1}{\ln(10)(\bar{M} - M_c)} \quad (5.3)$$

$$\sigma_b = \frac{b}{\sqrt{N}} \quad (5.4)$$

where $\bar{M}$ is the average magnitude, and N is the total number of magnitude data in the sample.

The classical maximum likelihood approach of Aki (1965) provides a robust framework for estimating the b -value under the assumption of a complete catalog above a fixed completeness magnitude M_c . Still, it does not explicitly account for the gradual transition between incomplete and complete detection at lower magnitudes.

To overcome this limitation, Ogata & Katsura (1993) [104] proposed a more flexible statistical framework that jointly models the Gutenberg–Richter distribution and the probability of detection. In their formulation, the theoretical intensity of earthquakes at magnitude M is given by the exponential law $\lambda(M) \propto e^{-\beta M}$, while the detection probability $q(M)$ is expressed as the cumulative distribution function of a normal distribution:

$$q(M) = \text{erf}\{M|\mu, \sigma\} = \int_{-\infty}^M \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \quad (5.5)$$

where μ is the magnitude corresponding to a 50% detection probability, and σ controls the width of the transition from partial to complete detection. The observed magnitude frequency distribution is then obtained by *thinning* the theoretical intensity with this detection probability:

$$\lambda_{\text{obs}}(M) = \lambda(M) \cdot q(M) \quad (5.6)$$

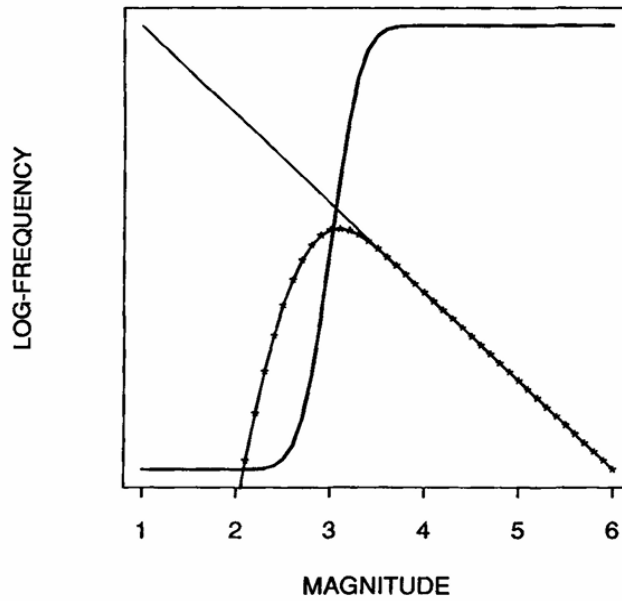


Figure 5.1: The schematic diagram illustrates the probabilistic framework of the detected earthquake frequency distribution. The straight line represents the Gutenberg–Richter relation in logarithmic scale, while the thick solid curve depicts the detection probability of events at each magnitude, ranging from 0 to 1 in linear scale. The line connecting the asterisks shows the histogram of the detected events, plotted in logarithmic scale. Image taken from Ogata & Katsura (1993) [104]

In this chapter, I present a detailed analysis of the magnitude of completeness and b -value associated with the seismicity detected in the Val d’Agri region, following the Ogata & Katsura (1993) [104] model, implemented through a Python script provided by Sebastian Hainzl (personal communication, 2025). I compare results obtained from the reference catalog with those derived from enhanced detections using deep learning workflows.

5.2 Analysis

To validate the robustness and stability of the b -value I performed a stability analysis by computing the b -value for increasing values of cutoff magnitude M_{cut} using the Maximum Likelihood method (Aki, 1965). Results are shown in Figure 5.2. The blue dots represent the b -value computed for events with $M \geq M_{cut}$, with vertical bars denoting the standard deviation (σ_b) derived from bootstrap resampling. The b -value shows a monotonic increase with M_{cut} up to $M_{cut} \approx 0.3$ for PRN and QS, after which it stabilizes in the range $b \approx 1.1$ – 1.2 , indicating that the effects of magnitude incompleteness are negligible above this threshold. For the CMS catalog the b -value stabilizes around $M_{cut} \approx 0.5$, with a value ≈ 1 – 1.1 . The gray stepped line shows the number of events available above each M_{cut} value (right axis).

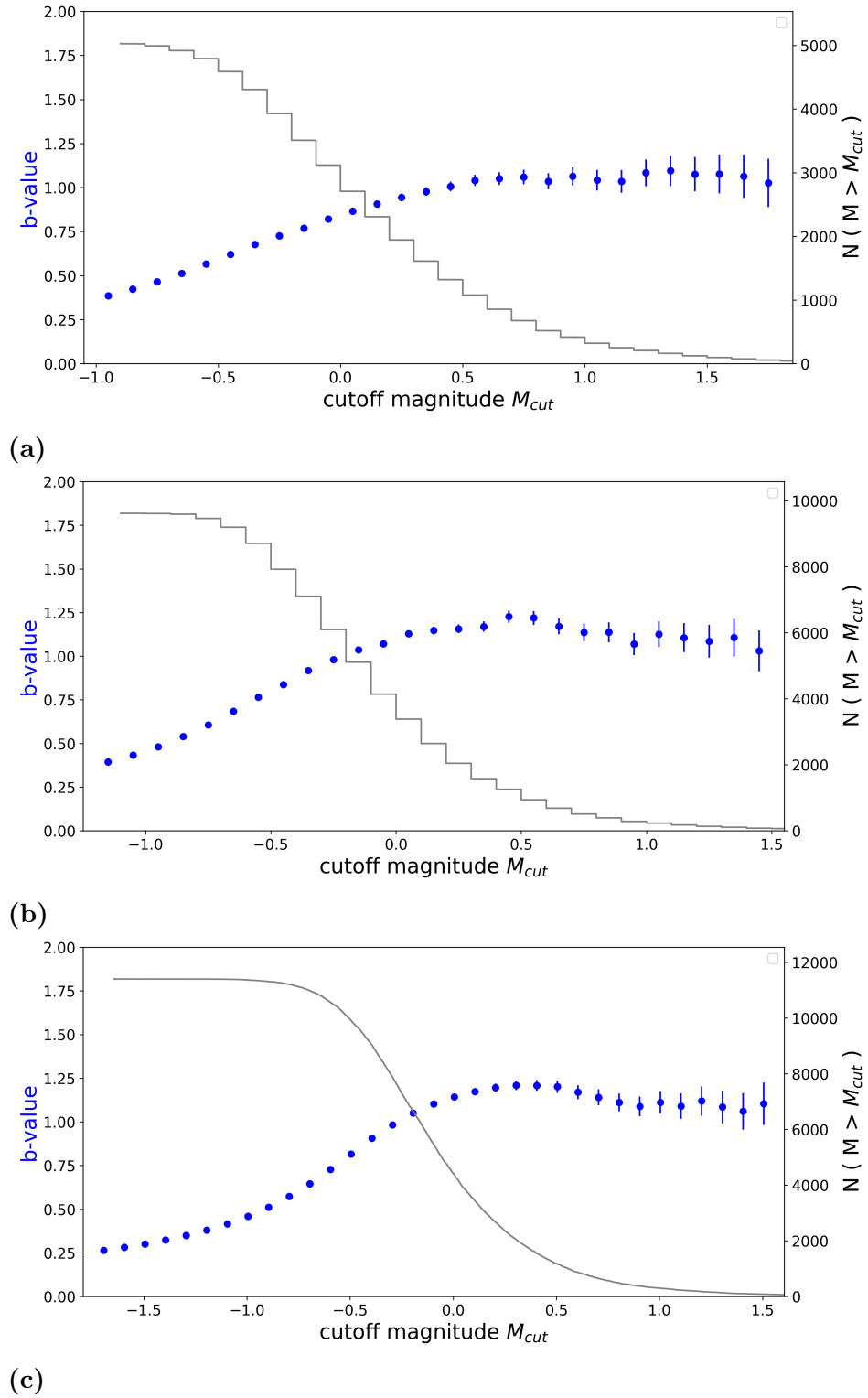


Figure 5.2: b-value stability analysis for a) CMS, b) PRN and c) QS catalogs. b-values for increasing values of cutoff magnitude (M_{cut}) are calculated using the Maximum Likelihood method.

To assess the magnitude of completeness (M_c) and the statistical properties of the magnitude distribution, I applied the Ogata & Katsura (1993) approach [104], which models the observed frequency–magnitude distribution (FMD) as the convolution of a Gutenberg–Richter (G–R) law with a cumulative normal distribution representing the detection probability. This method allows for simultaneous estimation of the b -value and the detection probability, accounting for the gradual detection drop-off at lower magnitudes.

Figure 5.3 illustrates the observed frequency-magnitude distribution (grey histogram, logarithmic scale), the fitted Ogata–Katsura model (red curve), and the inferred detection probability (blue curve). The detection function is parameterized by a Gaussian error function with parameters $b=1.09$, $\mu=-0.32$, and $\sigma = 0.23$ (PRN), $b=1.19$, $\mu=-0.23$, and $\sigma = 0.26$ (QS), $b=0.93$, $\mu=-0.11$, and $\sigma = 0.29$ (CMS). These values indicate a smooth transition in detection capability around μ . I will refer to the magnitude value with the 90% and 99% of detection probability respectively as M90 and M99. 90% and 99% detection thresholds (blue dashed lines), respectively occur at M90 = -0.1 and M99 = 0.2 (PRN), M90= 0.1 and M99= 0.3 (QS), M90 = 0.3 and M99 = 0.6 (CMS). For comparison, the maximum curvature (MaxC) method (green dashed line), often used as a quick empirical estimate of M_c , yields significantly lower values of $M_c = -0.3$ for PRN, $M_c = -0.3$ for QS, and $M_c=-0.2$ for CMS. Moreover, the three panels in 5.3 clearly demonstrate the lower sensitivity of the completeness magnitude estimation obtained with the MaxC method. The M_c value derived from this approach appears remarkably stable across the three catalogs, despite the evident decrease in completeness of the CMS catalog at lower magnitudes. This behavior is further supported by the detection probability curves, which exhibit significantly higher M90 and M99 values for the CMS catalog compared to the two deep-learning-based catalogs, confirming the reduced detection capability of the traditional workflow at smaller magnitudes.

5.3 Discussion

The comparison among the three catalogs provides important insights into the relative quality and representativeness of the datasets. The CMS catalog is

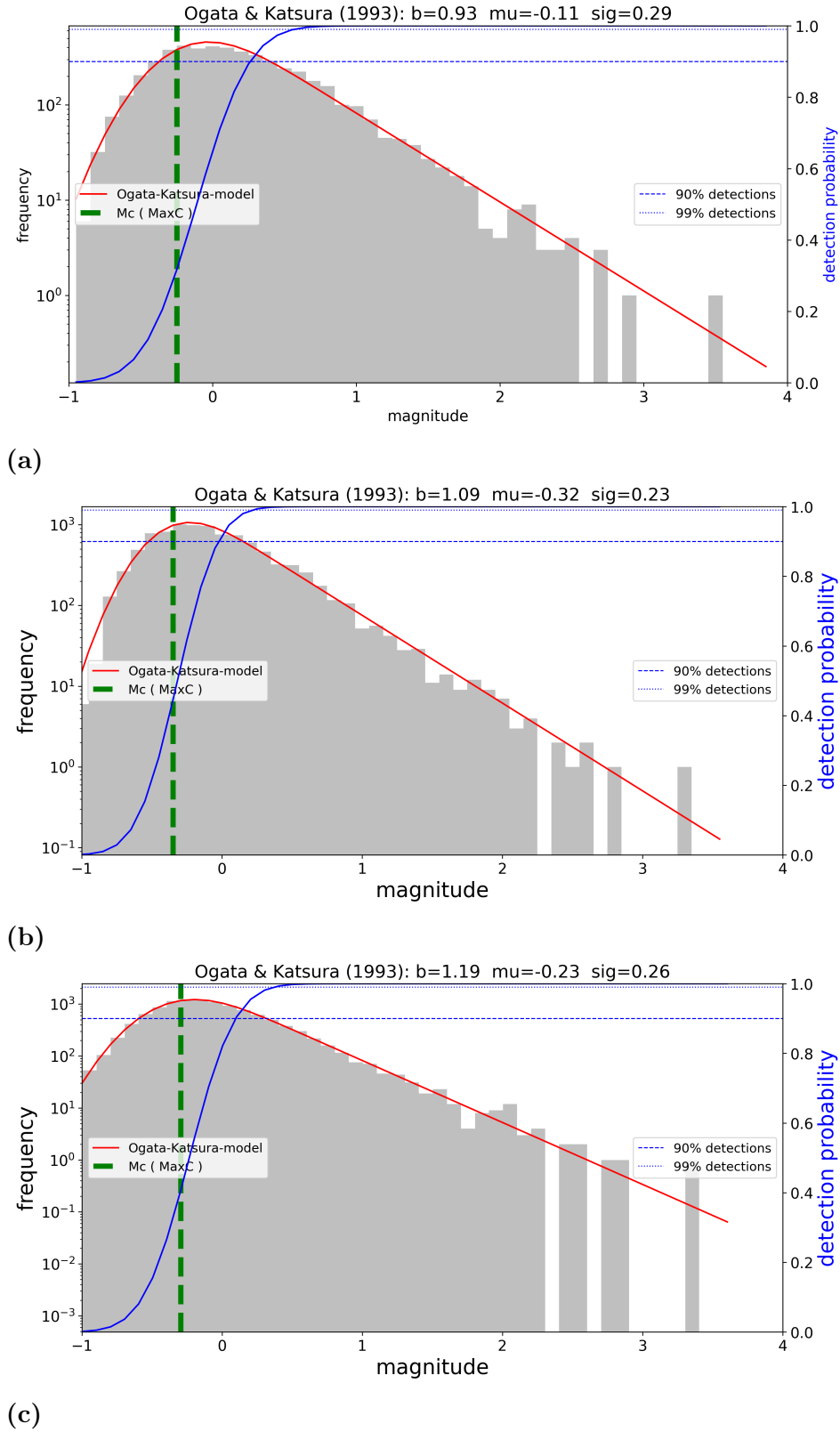


Figure 5.3: the panels illustrate the observed frequency-magnitude distribution (grey histogram, logarithmic scale), the fitted Ogata-Katsura model (red curve), and the inferred detection probability (blue curve) for each catalog: a) CMS, b) PRN and c) QS). The green dashed line shows the magnitude of completeness obtained with the MaxCurvature method.

characterized by a relatively high magnitude of completeness ($M_c \sim 0.5\text{--}1.0$) and a low *b*-value (~ 0.93). This combination suggests that the catalog underrepresents small-magnitude events, which may be related to method limitations or periods of poorer detection capability. As a result, the inferred size distribution of earthquakes is biased toward larger magnitudes, potentially affecting seismicity rate estimates and hazard assessment.

In contrast, the PRN and QS catalogs achieve a substantially lower M_c and yields higher and more stable *b*-values ($\sim 1.1\text{--}1.2$). This indicates a more reliable recording of small earthquakes and a dataset that conforms more closely to the Gutenberg–Richter distribution. The higher *b*-value reflects a richer population of small-magnitude events, which may be linked to differences in local seismicity patterns or improved detection efficiency of the monitoring system. The ability to reliably capture the smallest events provides a more accurate view of the seismic process, particularly for studies focused on earthquake triggering, stress release, and scaling relationships.

Overall, these results highlight the critical role of catalog completeness in seismic analysis and the benefits of a more sensitive method for earthquake detection.

This analysis reveals systematic differences not only in the magnitude of completeness (M_c) but also in the stability of the estimated *b*-values. In the CMS catalog, the *b*-value exhibits a pronounced variability as a function of the cutoff magnitude: it increases steeply at low M_{cut} and only stabilizes at values close to unity for $M > 1$. This instability reflects the limited detection of small-magnitude earthquakes, which biases the frequency–magnitude distribution and produces a less reliable estimate of the scaling exponent.

By contrast, the PRN and QS catalogs show a more consistent behavior: the *b*-value increases smoothly with M_{cut} and converges toward stable values ($\sim 1.1\text{--}1.2$) already at lower magnitudes than CMS. This suggests that the catalogs include a sufficient number of small events to constrain the slope of the Gutenberg–Richter relation, even though some fluctuations remain visible at the lowest magnitudes due to partial detection. This robustness indicates both completeness and a reliable representation of the seismicity.

Taken together, these observations highlight that the variability of the *b*-value

with respect to M_c is a sensitive evaluation of catalog quality.

In Figure 5.4, I compare only the events above a cutoff magnitude of the CMS catalog. To provide a more robust magnitude threshold for comparing the catalogs, I use the detection probability function derived by the Ogata & Katsura (1993) [105] model. From this, I take the magnitude corresponding to a 99% probability of detection in the CMS catalog ($M_{99}=0.6$), which I used as a threshold to filter the events in the three maps. This approach provides a more objective and statistically efficient estimate of the magnitude of completeness than commonly used empirical methods such as the Maximum Curvature, which tend to underestimate the completeness level (e.g., Woessner & Wiemer, 2005 [100]). Limiting the CMS catalog to $M > 0.6$ the recall increases to 94% (PRN) and 81% (QS). This outcome aligns precisely with my objectives: enhancing the monitoring sensitivity to lower magnitudes, expecting similar results at higher magnitudes, thereby confirming the method's reliability.

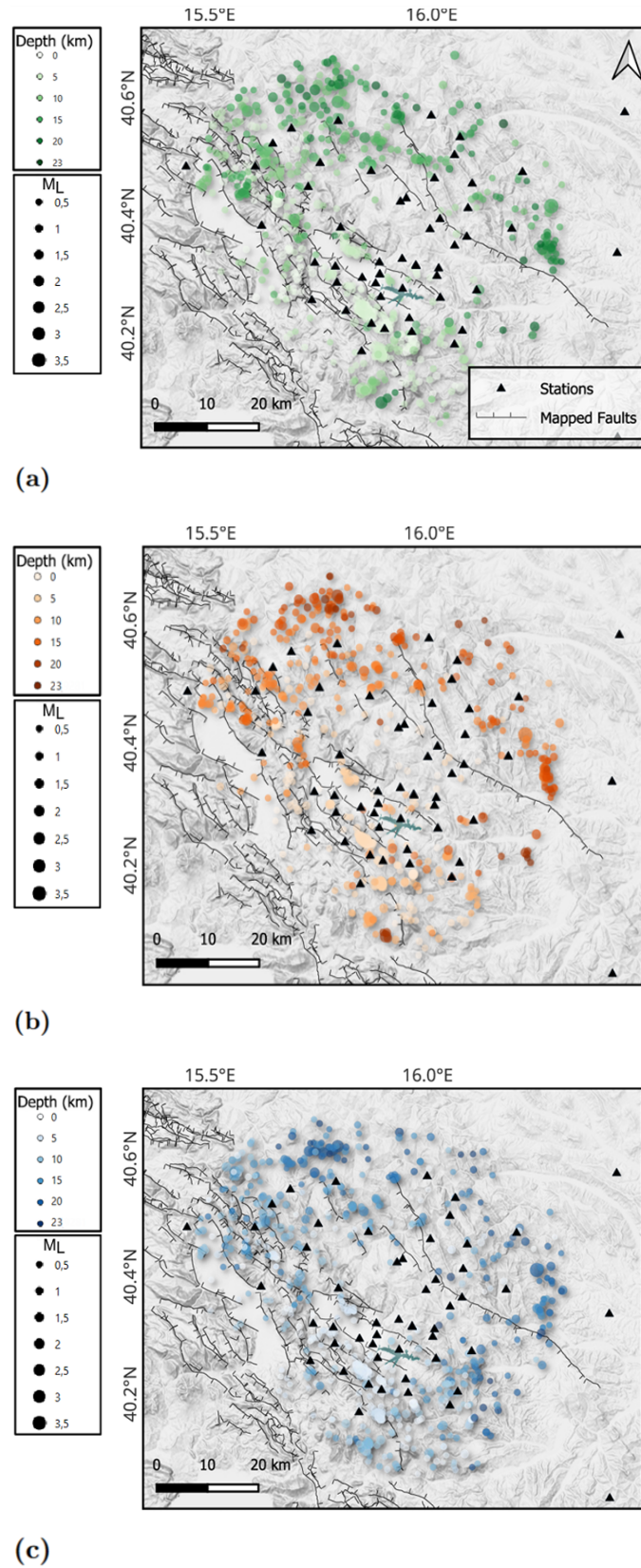


Figure 5.4: Distribution of epicenters of only the events above the $M=0.6$ of the CMS catalog for a) CMS catalog, b) PRN catalog, c) QS catalog. The dot size is proportional to the Magnitude, while the color scales with the depth.

Chapter 6

From moment tensors and waveform similarity to hazard implications

6.1 Introduction

Understanding the geometry and kinematics of active faults in seismically active regions is fundamental for improving seismic hazard assessment. In areas with complex fault systems, such as the Val d’Agri region in southern Italy, high-resolution analysis of microseismicity provides crucial insights into fault segmentation, rupture processes, and stress field evolution.

In this chapter¹, I take advantage of the deep learning based seismic catalog obtained in Chapter 4 for the Val d’Agri region from 2021 to 2023. Enhanced catalogs enable detailed imaging of fault geometries and provide valuable insights into the spatial and temporal patterns of seismicity. I employ a waveform similarity-based clustering technique to identify groups of earthquakes generated by repeated ruptures along common fault segments or closely spaced sources.

Waveform similarity clustering is a seismological technique aimed at grouping earthquakes based on the similarity of their recorded waveforms, rather than solely

¹This Section is taken from: Caredda et al., Geometry and kinematics of active seismogenic structures from clustered microseismicity in the Val d’Agri (Southern Italy), Tectonophysics, under review

on source parameters such as magnitude or location. The underlying assumption is that events with highly correlated waveforms share similar source mechanisms, focal depths, and propagation paths (Geller & Mueller, 1980 [106]). This approach exploits the fact that, for events occurring in close proximity and recorded at the same stations, the path and site effects are nearly identical, making the waveform shape primarily sensitive to the source characteristics. The method is based on the computation of the cross-correlations between the waveforms recorded at the seismic stations for the selected events. When similar waveforms are observed consistently at several stations, they suggest comparable focal mechanisms, source depths, and ray paths (e.g., Maurer & Deichmann, 1995 [107]). Consequently, grouping earthquakes into clusters of similar events can provide valuable insights into fault geometry and faulting style. Under favorable conditions, this method can be used to identify and map faults with high spatial detail (e.g., Tsujiura, 1983 [108]; Shearer et al., 2003 [109]), and its main advantage is that it does not rely on precise hypocentral locations, making it suitable even in cases where location uncertainties are large. Events displaying very high correlation coefficients (typically > 0.9 – 0.95) are often referred to as *repeaters* (e.g., Geller & Mueller, 1980 [106]; Igarashi et al., 2003 [110]).

In this study, waveform similarity clustering is particularly advantageous because it allows us to extract geometric and source-related information even for microseismic events, for which calculating focal mechanisms is often challenging due to their low magnitude and limited signal-to-noise ratio. By combining waveform similarity analysis with focal mechanism determination for larger events, I can extend the source characterization to a broader portion of the seismic catalogue. Furthermore, the similarity-based clustering provides the foundation for the subsequent double-difference relocation, which directly exploits the waveform cross-correlation measurements to achieve higher relative location accuracy. By focusing on clusters of similar events, cross-correlation techniques can significantly enhance location accuracy compared to standard picking methods, by providing highly precise P- and S-wave arrival times (e.g. Nakamura, 1978 [111]; Menke, 1999 [112]; Walhauser & Ellsworth, 2000 [113]; Shearer et al., 2005 [114]).

To refine the event locations and better resolve the geometry of the seismicity, I apply double-difference relocation (Waldhauser & Ellsworth, 2000 [113]). Together,

cluster identification and accurate hypocentral locations can be used to image the geometry of distinct fault structures. The spatial and temporal evolution of seismicity allow to track the progressive activation of these fault structures.

Focal mechanisms, also known as fault-plane solutions, are fundamental tools in seismology for interpreting the source processes of earthquakes. They describe the orientation of the fault plane that slipped and the slip direction, providing insights into the tectonic environment and stress regime at the earthquake source. Focal mechanisms solutions are obtained through the moment tensor inversion. The moment tensor is a second-order tensor that represents the distribution of forces within a seismic source. It can be decomposed into three fundamental components: Isotropic component (ISO) that represents a uniform volume change, such as an explosion or implosion, Double-Couple component (DC) which corresponds to shear faulting on a plane, characterized by two equal and opposite forces acting along a fault plane, and Compensated Linear Vector Dipole component (CLVD) that accounts for non-double-couple sources, such as tensile faulting or crack opening. Mathematically, the moment tensor M can be expressed as the sum of these three components:

$$M = M_{ISO} + M_{DC} + M_{CLVD} \quad (6.1)$$

with $M_{ISO} = \frac{1}{3}Tr(M)I$. $Tr(M)$ is the trace of the moment tensor and I the identity matrix. The decomposition allows for the classification and physical interpretation of seismic sources, aiding in the understanding of faulting mechanisms and stress fields (Vavryčuk, 2015 [115]; Jost & Herrmann, 1989 [116]). To interpret the Moment Tensor components one should remember that high DC component indicates shear faulting, typical of tectonic earthquakes; high ISO component suggests explosive sources, such as volcanic activity or underground explosions; while high CLVD component implies non-DC sources, which can result from processes like fluid injection or mining activities (Julian et al., 1998 [117]; Miller et al., 1998 [118]).

Understanding the balance between these components is crucial for interpreting the nature of seismic events and their implications for hazard assessment.

In regions with complex tectonic settings, such as the Southern Apennines, analyzing the moment tensor solutions of seismic events can reveal detailed infor-

mation about the fault geometry and stress regime. Even a limited number of well-constrained moment tensor solutions can provide valuable insights into the seismotectonic framework of an area (Vavryčuk, 2015 [115]).

Decreasing the target of moment tensor inversion down to $M \sim 2$ in some cases, which is achieved by modeling full and bodywaves waveforms using a region-specific 1D velocity model, provides new information on rupture processes in this region and supports the fault plane identification, further inferring information on their rupture modes.

My integrated approach, combining waveform similarity clustering, high-precision relocation, and source mechanism inversion, provides a comprehensive view of the active fault system in the Val d'Agri and supports ongoing efforts in seismic hazard characterization, particularly in a region of high societal and industrial relevance.

6.2 Data and Methods

6.2.1 Waveform similarity-based clustering

I analyze the enhanced PRN earthquake catalog from Chapter 4 obtained with machine learning picking (see Figure 4.11a). This catalog includes 9,661 earthquakes recorded from 2021/01/01 to 2023/10/01, with magnitudes ranging from -1.1 to 3.5. The data were acquired from a combination of seismic networks (IV, GE, VA, VD, IX, TP), accounting for a total of 56 stations.

To compute the waveform similarity, I apply Clusty, a tool by Petersen et al. (2020)[119]. I consider only earthquakes with $M \geq 0$ (4,141 events). The pre-processing stage involves bandpass filtering in the frequency range of interest. I test three frequency bands for the cross-correlations: higher frequency bands 1-3 Hz and 0.5-2 Hz better resolve bodywave phases, while a lower frequency band 0.05-0.2 Hz mostly depicts surface waves for shallow earthquakes; this last frequency band was also used for the focal mechanism inversion. I consider event pairs with a minimum SNR of 1.5 and a maximum inter-event distance of 40 km. If the waveforms satisfy the predefined thresholds for signal-to-noise ratio (SNR),

inter-event distance, and station-event distance, cross-correlations are computed for event pairs at the various seismic stations.

For each selected pair, I require a minimum cross-correlation coefficient of 0.65 at at least three stations with an azimuthal coverage of at least 45° . A weighted sum of the cross-correlation coefficients across the different stations is then used to calculate the network similarity for each event pair (with values ranging from 0 to 1, where 1 indicates maximum correlation).

The resulting network similarity values are then used as the input metric—in which a value of 0 corresponds to identical events—for the density-based DBSCAN clustering algorithm (Ester et al., 1996 [120]), whose outcome depends on the selection of two clustering parameters, namely ϵ and $N_{\min}$, whose choice depends on the objective. ϵ controls the required grade of similarity: smaller ϵ values highlight repeaters along the same structure, yielding many small clusters and numerous unclustered events, while larger ϵ values are more suitable for identifying and delineating broader fault structures. $N_{\min}$ is the minimum number of earthquakes required to declare a cluster: smaller values allow for smaller clusters to be detected, while higher values ignore them. The workflow is schematized in Figure 6.1.

6.2.2 Double-difference relocation

I relocate each cluster individually using the hypoDD software (Waldhauser & Ellsworth, 2000 [113]), which implements the double-difference algorithm to refine earthquake locations by minimizing the effects of an inaccurate velocity model. The double-difference (DD) method, as implemented in HypoDD (Waldhauser & Ellsworth, 2000 [113]), is designed to improve relative earthquake locations by minimizing the residuals of differential travel times between pairs of nearby events recorded at common stations. The core assumption is that, for two closely spaced earthquakes i and j recorded at the same station k , the ray paths from the hypocenters to the station are nearly identical, and thus most of the path-dependent velocity model errors are canceled when considering their travel-time difference. The method seeks to minimize the residuals between observed and

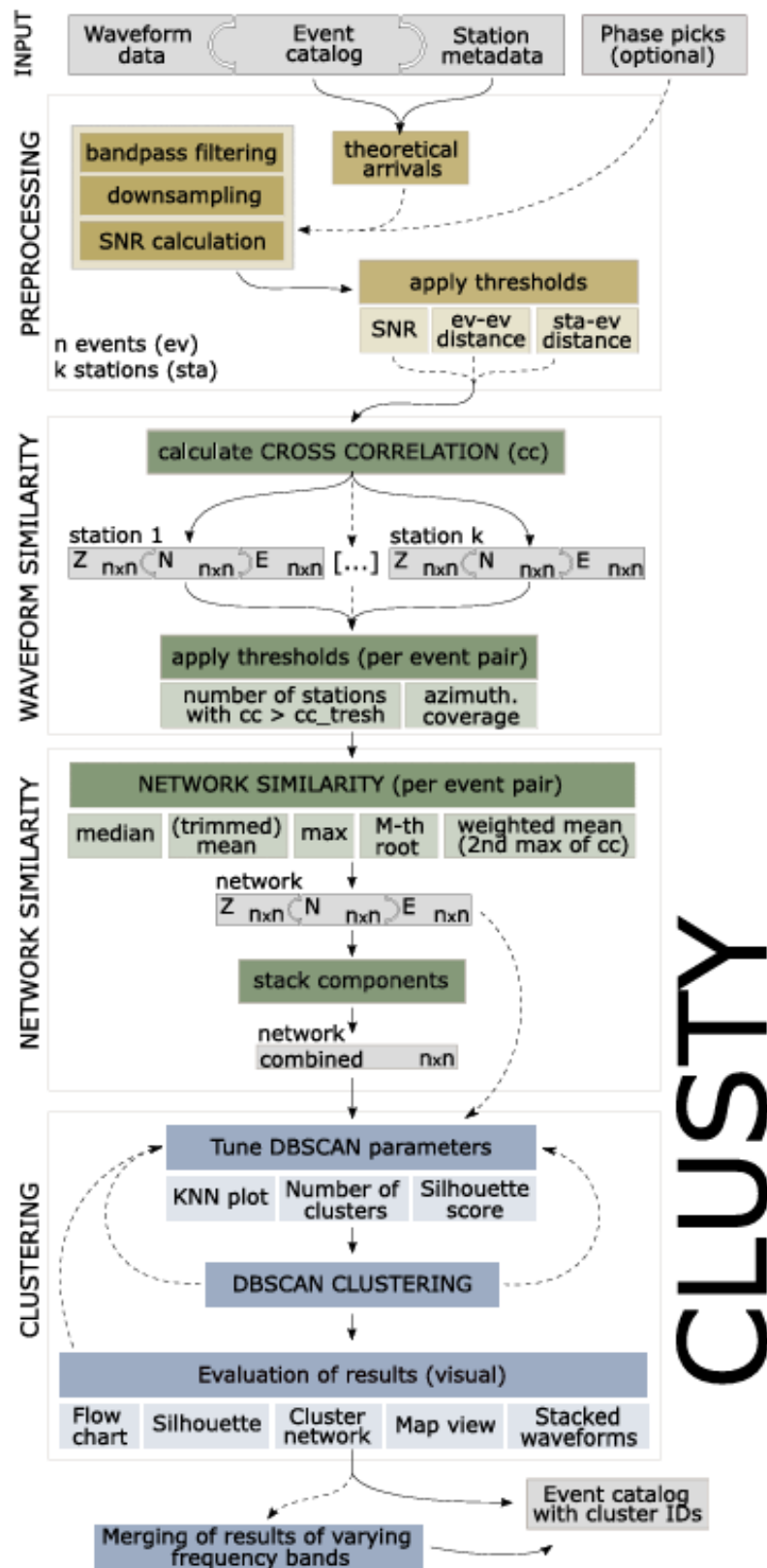


Figure 6.1: Clusty workflow. Image taken from Petersen et al., 2021.

calculated differential times over all event pairs and stations:

$$r_k^{ij} = (t_k^i - t_k^j)^{obs} - (t_k^i - t_k^j)^{cal} \quad (6.2)$$

The equation 6.2 defines the double-difference. The problem is linearized by expanding the travel-time function around the current hypocentral estimates, leading to a system of equations of the form:

$$\mathbf{W}\mathbf{G}\mathbf{m} = \mathbf{W}\mathbf{r} \quad (6.3)$$

where $\mathbf{G}$ is a matrix containing the partial derivatives of travel times with respect to source coordinates and origin time, $\mathbf{m}$ is the change in the hypocenter coordinates we want to determine, $\mathbf{r}$ is the data vector containing the double-differences, and $\mathbf{W}$ is a matrix containing the weights. By solving this system in a least-squares sense, HypoDD produces highly accurate relative locations for clustered events, typically reducing location errors from kilometers to tens of meters when high-quality cross-correlation data are used.

I performed the relocation using a combined dataset, allowing both catalog (P and S picks) and cross-correlation data – only retained if the correlation coefficient exceeds 0.7 – to constrain the solution. I include both P and S phases, and limit the maximum station-event distance to 50 km to ensure high-quality differential times. Relocations are controlled through four steps of four iterations using the LSQR inversion method (Paige & Saunders, 1982 [121]). In the first two steps I assign more weight to catalog picks than cross-correlation differential times, to constrain the cluster absolute location. In the last two steps, I progressively increase the weight of cross-correlation differential times and therefore sharpen the internal relative geometry of the cluster. I use a 1D velocity model representative of the Val d’Agri region from (author?) [38], the same employed for the absolute locations in the original catalog. To estimate hypocentral location uncertainties I employ a bootstrap approach that perturbs differential travel times and cross-correlations with Gaussian noise and repeats the HypoDD relocation 100 times for each cluster. I estimate the uncertainties as the standard deviation of the relocated hypocentral coordinates obtained from the bootstrap solutions.

6.2.3 Moment tensor inversion

To compute selected source mechanisms, I use Grond, an open-source software for robust earthquake source characterization (Heimann et al., 2018 [122]). Grond performs a probabilistic joint inversion of moment tensors and finite fault parameters using seismic waveforms and waveform-derived attributes. The inversion minimizes an objective function or misfit value between the observed and predicted data. The misfit between observed and synthetic data is defined as a weighted norm of sample differences, normalized by the observations (Heimann, 2018 [122]):

$$M = \frac{(\sum (w_i \cdot |d_{obs,i} - d_{synth,i}|)^p)^{\frac{1}{p}}}{(\sum (w_i \cdot |d_{obs,i}|)^p)^{\frac{1}{p}}} \quad (6.4)$$

where w_i are the weighting factors and p is typically 2 and corresponds to the L^2 -norm. Waveform data are preprocessed by tapering and bandpass filtering before misfit computation. For waveform data, Grond also allows misfit based on the maximum cross-correlation between observed and synthetic data:

$$M_{CC} = 1 - \max(CC) \quad (6.5)$$

The inversion relies on a Bayesian bootstrap approach (Rubin, 1981 [123]), which generates multiple resampled datasets with random weights assigned to each observation. This method allows full exploration of the model space and provides meaningful uncertainty estimates while accounting for trade-offs between parameters. Forward modeling is based on pre-computed Green's function databases, enabling rapid generation of synthetic seismograms and near-field displacements. I computed the Green's functions using the 1D velocity model from Improta et al. (2017)[38], with a maximum frequency of 20 Hz. I invert full waveform data in the 0.03–0.2 Hz frequency range to obtain moment tensor solutions for selected $M > 2$ events. For a group of chosen earthquakes, I retrieve the waveforms from the INGV web service (https://terremoti.ingv.it/webservices_and_software) to extend the distance range, including more seismic stations. For the largest earthquakes ($M \geq 2.5$), I use two group targets: frequency band 0.05–0.2 Hz for stations at distance ≤ 30 km and 0.03–0.08 Hz for stations at distance between 30 and 65 km. I reduce the distance radius and move to higher frequency ranges

for smaller earthquakes. I performed 100,000 iterations to ensure full exploration of the model space and to obtain robust posterior distributions for the source parameters, while naturally accounting for trade-offs between parameters.

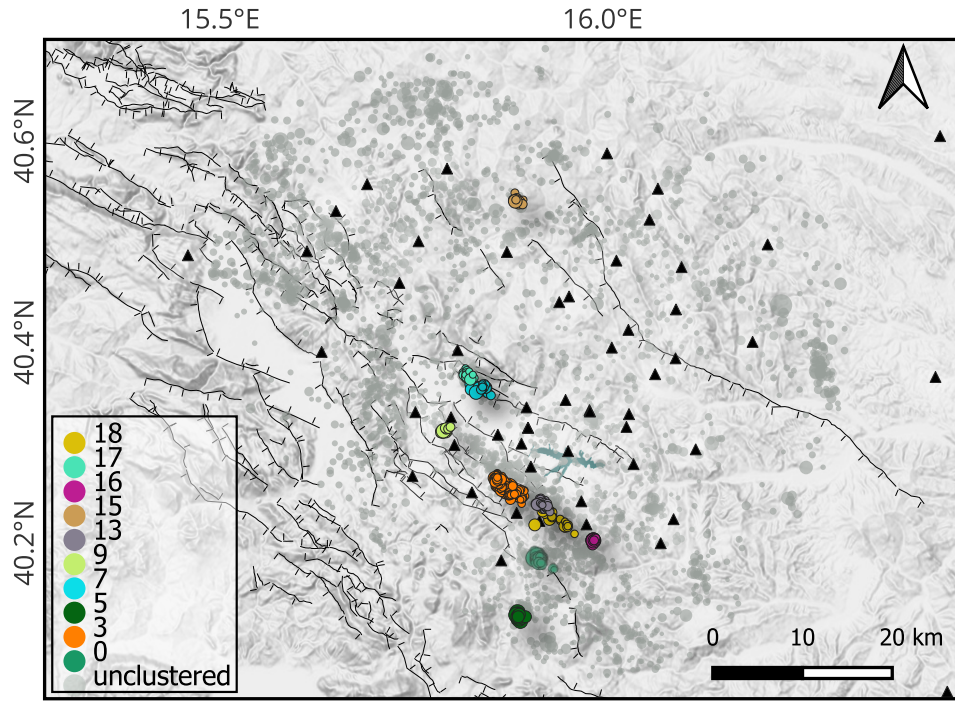
6.3 Results

Coherent spatial distribution and clustering results are found for the higher-frequency bands ($1 \div 3$ Hz in Figure 6.2a).

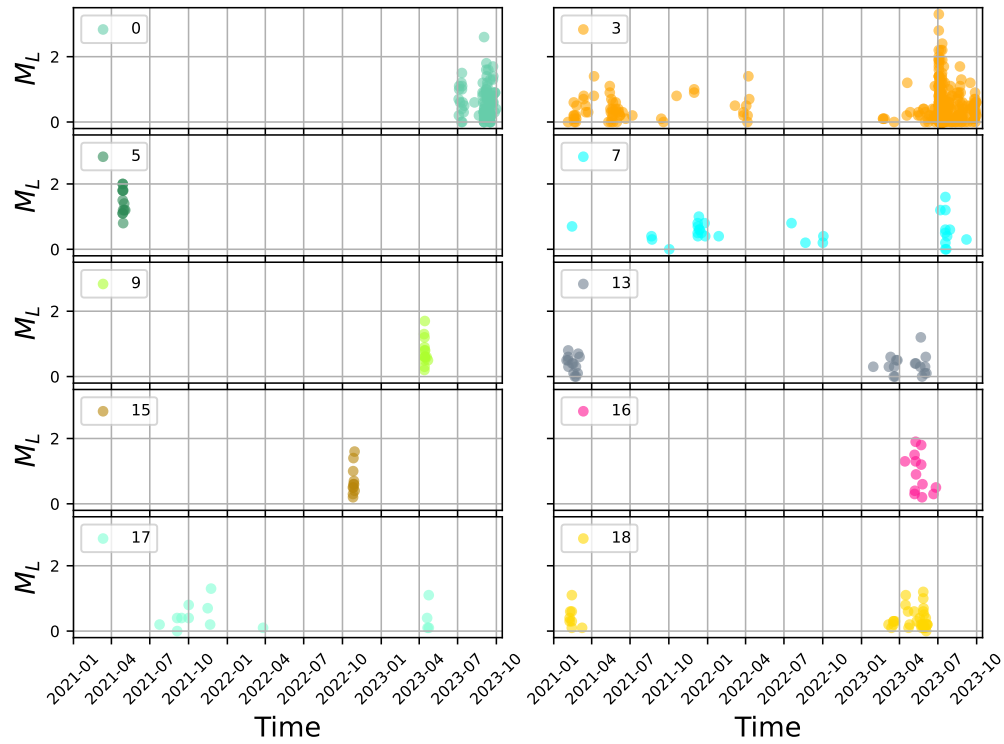
I tested the sensitivity of the clustering results to variations in the frequency range used to compute waveform similarity, as well as to the DBSCAN parameters ϵ and N_{min} . The results are shown in Figures A.4-A.6 in the Supplementary Information. At lower frequencies ($0.05 \div 0.2$ Hz), only one main cluster is visible (see Figure A.4a), corresponding to the group for which most focal mechanisms were obtained. Increasing the frequency (Figure A.4b) allows the identification of more clusters, although some remain sparsely populated. Similarly, smaller ϵ values tend to split clusters that likely belong to the same structure, whereas larger values merge geographically scattered events (Figure A.5). For N_{min} , very small values produce numerous but less reliable clusters, while larger values result in fewer, more robust clusters (Figure A.6).

After testing different values for the ϵ and N_{min} parameters, I select the setup with $\epsilon = 0.34$ and $N_{min} = 10$, which enables the identification of 10 distinct fault structures on the map. The total number of clustered earthquakes is 607.

Most clusters are located near the Pertusillo Lake. These clusters likely represent groups of earthquakes originating from spatially close sources or single fault segments. Some clusters delineate structures that have been reactivated multiple times between 2021 and 2023 (e.g., the orange and light-blue clusters, number 3 and 7 in Figure 6.2), while others are active only for shorter periods. The most populated is cluster 3 (orange), where we distinguish a small seismic sequence started in July 2023, with the mainshock having $M_L = 3.3$ (that we will discuss in more detail further on). Except for clusters 0 and 3, all the clusters are characterized exclusively by micro-earthquakes with local Magnitude $M_L < 2$.



(a)



(b)

Figure 6.2: a) Map view of waveform similarity-based clusters obtained with $\epsilon=0.34$ and $N_{min}=10$, frequency range: 1-3 Hz. Each cluster has a reference color shown in the legend. Unclustered earthquakes are shown in light grey. b) Magnitude-time distribution for each cluster in the map.

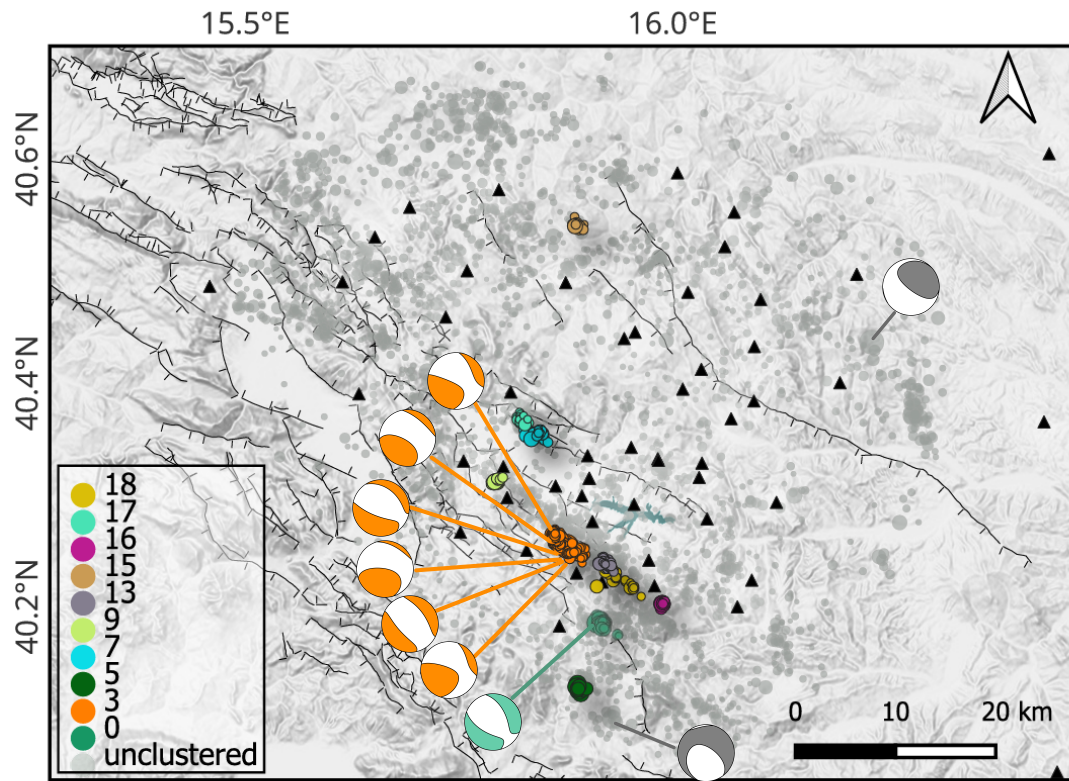


Figure 6.3: Map view of waveform similarity-based clusters obtained with $\epsilon=0.34$ and $N_{\min}=10$, frequency range: 1-3 Hz. Each cluster has a reference color shown in the legend. Unclustered earthquakes are shown in light grey. Calculated focal mechanisms are shown with color paired to the corresponding cluster color.

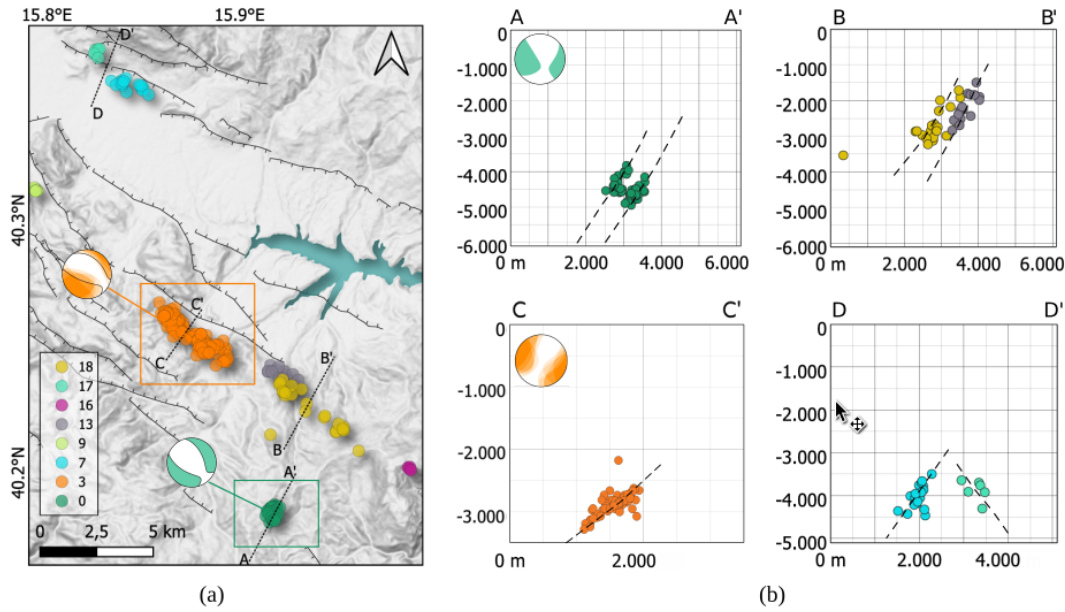


Figure 6.4: Relocated clusters in the area of interest in a) map view and b) vertical cross-sections. Reference focal mechanisms for clusters 0 and 3 are plotted, both in map view and vertical cross-section.

Figure 6.3 displays the calculated focal mechanisms. Six solutions are associated with cluster 3 (orange), showing predominantly extensional features with strike-slip components. One is in cluster 0, with a focal mechanism similar to the ones in cluster 3. I also retrieved two solutions corresponding to unclustered earthquakes. Both show predominantly reverse components. They correspond to deeper seismicity (15–20 km) compared to the clusters near the Pertusillo lake. The complete list of focal mechanism parameters — including origin time, hypocentral location, magnitude, and nodal plane orientations — is reported in Table A.1 in the Supplementary Information.

Figure 6.4 presents the map view of the hypoDD relocations of the most populated clusters (0, 3, 7, 13, 17, and 18), along with their corresponding cross-sections. For most clusters, at least 50% of the original earthquakes have successfully been relocated using the double-difference technique, yielding a total of 318 events. Final horizontal and vertical location uncertainties are of the order of tens of meters.

Cross-section A–A' (cluster 0) highlights two distinct structures between approxi-

mately 3.5 and 5 km depth. In cross-section B–B', clusters 13 and 18 are shown to be spatially close, both in location and depth (1.5–3.5 km). Cluster 3 (orange), the most populated, delineates a structure approximately 3 km in length, extending between 2 and 4 km depth.

The reference focal mechanisms for clusters 0 and 3 indicate normal faulting. I can select the actual nodal plane from the geometry inferred from cross-sections A–A' and C–C'. Cluster 7 also displays a dip direction compatible with the nearby mapped fault trace, as shown in section D–D'.

All cross-sections reveal structures dipping toward the southwest, except for cluster 17, which dips to the northeast. However, the limited number of events in this cluster precludes a robust structural interpretation.

6.4 Discussion

6.4.1 Waveform similarity and focal mechanisms interpretation

Waveform similarity is robust and has allowed to identify significant features in the study area, whose significance we will shortly discuss in the following. The fact that some clusters are concentrated near the Pertusillo Lake suggests the presence of pre-existing fault zones and accounts for their mechanical heterogeneities.

The variability observed across different frequency bands indicates that the choice of frequency range should be tailored to the magnitude and distance characteristics of the specific case study. Lower-magnitude earthquakes tend to radiate more energy at higher frequencies, making it difficult to group them based on waveform similarity in the lower-frequency bands that are more representative for larger events.

The selection of DBSCAN parameters $\epsilon = 0.34$ and $N_{min} = 10$ effectively isolates coherent seismicity patterns that likely correspond to individual fault segments or fault zones. The elongated shapes and spatial continuity of some clusters support the interpretation of repeated ruptures along well-defined structures. In particular, clusters 3 and 7 not only delineate fault geometries but also show temporal persistence, being active at multiple times between 2021 and 2023. Moreover, several clusters were activated around the same time during spring 2023.

Most clusters are composed solely of micro-earthquakes ($M < 2$). These results highlight the importance of analyzing low-magnitude seismicity to illuminate the finer-scale architecture of the fault system, which might be invisible to traditional catalog-based approaches.

The identification of discrete clusters allows for targeted focal mechanism and relocation analysis, that could further explain the tectonic regime and the mechanical behavior of individual structures.

The relocated seismicity and focal mechanisms in Figure 6.4 suggest that the activated structures are steeply dipping normal faults (around 60°), with the dip

direction generally toward the southwest. This structural orientation is consistent with the regional extensional regime and in agreement with the findings of Valoroso et al. (2023) [15], who proposed that active structures in the area may have a strike direction opposite to that of the major mapped faults. The focal mechanisms in this study are predominantly normal faulting, supporting the tectonic setting dominated by NE–SW extension.

In most cases, the inferred fault dips do not match the orientations of the surface fault traces. I therefore attribute this seismicity to a system of pre-existing blind reverse faults, as shown in Valoroso et al. (2023) [15]. The hypocentral depths correspond to the Apulian carbonate platform. This fault system developed during a previous compressional regime associated with the formation of the Apennine chain. At present, however, these structures exhibit kinematics consistent with the current extensional stress field of the area, suggesting a reactivation and kinematic inversion of the system.

In contrast, for cluster 7, the dip of the structure imaged along the DD' cross-section aligns with the southwestward dip of the mapped surface fault. This observation suggests that the seismicity in cluster 7 is likely related to the main Eastern Agri fault system.

The focal mechanisms in our study are predominantly normal faulting, supporting the tectonic setting dominated by NE–SW extension. I assess the quality of the focal mechanism solutions by considering the number of contributing stations and the azimuthal gap. Mechanisms are classified as quality A when they exceed the thresholds of a minimum of 10 stations, and at least 270° on azimuthal coverage; and as quality B otherwise (see Table A.1 in the Supplementary Information). Within cluster 3, some quality B solutions show slightly different strike angles compared to the higher quality mechanisms, yet remain broadly consistent with the dominant faulting style. Given the high waveform similarity within this cluster, I infer that the focal mechanisms for lower-quality events are likely to share the same geometry as those with robust solutions.

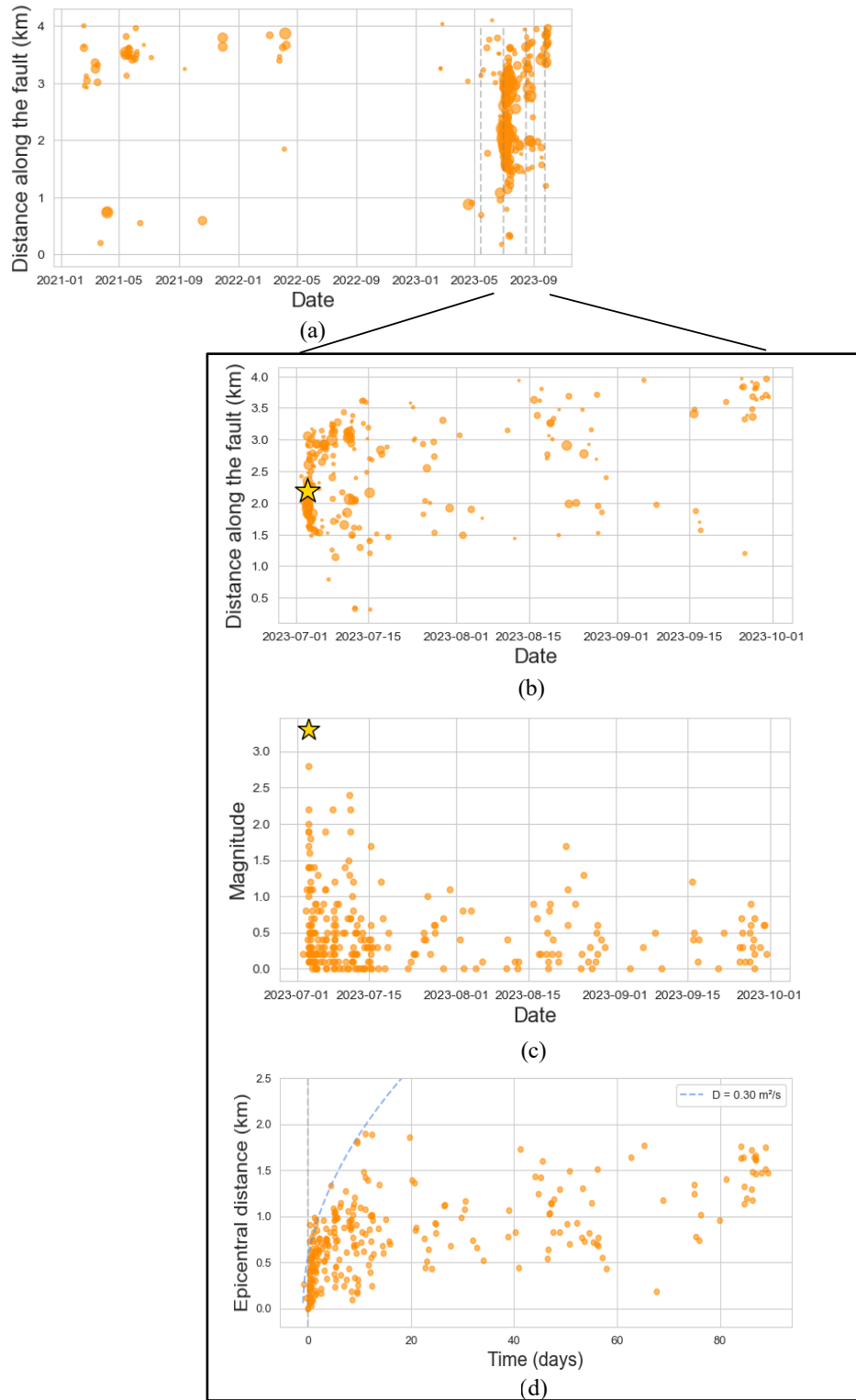


Figure 6.5: a) Distribution of epicenters in time along the fault structure illuminated by cluster 3. b) Zoom on the time window after the seismic sequence starts. c) Magnitude-time distribution during the same timespan as b). d) Epicentral distance from the starting point of the seismic sequence and possible diffusivity coefficient.

6.4.2 July 2023 seismic sequence

A relevant seismic sequence began in July 2023, marked by a mainshock of $M_w=3.4$, as derived from the moment tensor inversion performed with Grond (the local magnitude reported in the reference catalog is $M_L = 3.3$). This sequence appears to have activated the entire structure delineated by cluster 3 defined by waveform similarity-based clustering (Figure 6.5 a,b).

Following the mainshock, we observe a migration of seismicity toward the north-western end of the cluster, along with a relative quiescence in the central part of the structure. This spatial pattern may reflect the location of the rupture area of the mainshock, suggesting that stress was released in the central segment and subsequently transferred toward the unbroken edges of the structure. Such migration patterns have been interpreted in other contexts (e.g. Parotidis et al., 2003 [124]; Miller et al., 2004 [125]; Hainzl & Ogata, 2005 [126]) as indicative of aftershock triggering driven by post-seismic stress redistribution or fluid migration. I investigate whether this pattern could be compatible with fluid migration with a diffusivity coefficient of $0.3 \text{ m}^2/\text{s}$ (figure 6.5d), which is a reasonable value for the diffusivity (Hainzl and Ogata, 2005 [126]; Lei et al., 2008 [127]; Deng et al., 2010 [128]). The observed spatio-temporal evolution appears consistent with this diffusion model, suggesting that pore pressure propagation may have played a role in driving the sequence.

6.4.3 Hazard Scenarios

To assess the potential ground shaking associated with different rupture scenarios of the fault segment imaged by the relocation and clustering analysis, I performed 3-D finite-fault ground-motion simulations with the Pyrocko Seismosizer framework (Heimann et al., 2017 [96]). This physics-based approach allows us to explore both the recorded M_w 3.4 event and a set of hypothetical, larger ruptures to quantify their implications for local seismic hazard.

These simulations are intended to provide a first-order estimate of potential ground motions associated with the complete rupture of the fault segment illuminated by cluster 3, offering a physics-based complement to probabilistic hazard maps.

For the finite-fault ground-motion simulations, I pre-computed a database of Green's functions using the QSeis 2006b code (Wang, 1999 [129]), using the same 1D velocity model from Improta et al. (2017) [38] employed before. Green's functions are calculated for source depths from 1 to 4 km (step 0.1 km) and epicentral distances from 0 to 40 km (step 0.1 km), to ensure enough detail over the simulations' area.

I impose a rectangular source to simulate waveforms on a grid of synthetic stations of 20 km x 20 km (equally spaced at 0.25 km) around the hypoDD solution for the epicenter which corresponds to the distance range where peak ground velocities (PGVs) are expected to be most significant for earthquakes of this magnitude. A triangular source time function (STF) of duration 0.2 s was chosen to represent the slip-rate history, as it provides a simple yet physically realistic approximation of the moment release for moderate-magnitude events, avoiding the artificial high-frequency content of a boxcar STF.

To validate the model, I test it on the actual $M_w = 3.4$ earthquake and compare it to the PGV values measured at the real stations (Figure 6.7). For the original $M_w = 3.4$ event, I test the nodal plane solution derived from the source mechanism inversion (see Figure 6.6) and the geometry evidence (strike 107° , dip 63° towards SW direction) as a rupture scenario.

I subsequently hypothesize the rupture of the whole seismically active structure. Based on the empirical scaling relations proposed by Wells and Coppersmith (1994) [130], the expected moment magnitude for a complete rupture is estimated at $M_w = 4.7$ (Figure 6.8). The rupture velocity has been set to ~ 0.8 times the shear wave velocity at that depth, a typical value for normal faulting events (e.g. Graves & Pitarka, 2010 [131]; Broberg, 1996 [132]).

The resulting PGV distributions (Figures 6.7 and 6.8) show that the $M_w = 4.7$ scenario produces ground motion amplitudes up to 40 times larger than the $M_w = 3.4$ event. The maximum simulated PGV for the hypothetical $M_w = 4.7$ rupture reaches approximately 0.13 m/s (Figure 6.8a). Using empirical PGV–intensity relationships (e.g., Wald et al., 1999 [133]), this corresponds to a macroseismic intensity of roughly VI on the MMI (Modified Mercalli Intensity) scale, indicating moderate shaking with potential minor damage to poorly constructed buildings.

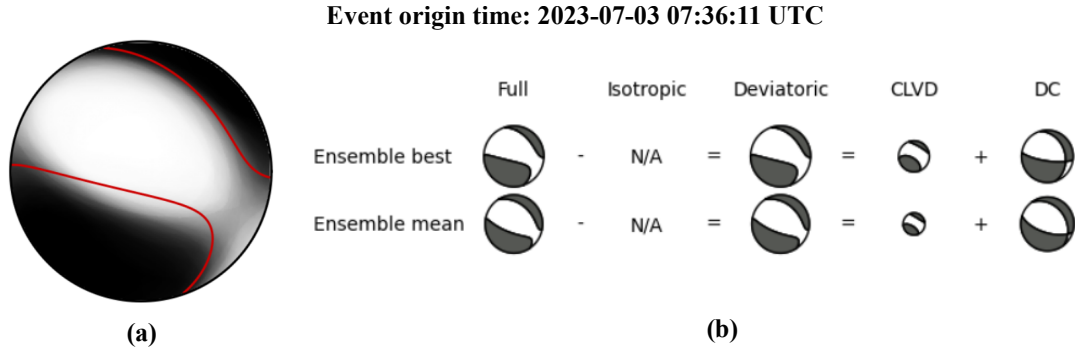


Figure 6.6: Moment Tensor solution for the $M_W = 3.4$ event. a) Fuzzy MT: for each ensemble solution, the strength of the P-wave radiation pattern is computed and stacked across all ray spokes. The resulting projection represents the cumulative (stacked) radiation pattern. When the variability among the ensemble solutions is low, the fuzzy plot exhibits well-defined black and white regions that align with the nodal lines of the global best-fitting solution (shown in red). b) Decomposition of the best-fitting moment tensor into its deviatoric and optimal double-couple components. The symbol size indicates the relative strength of the components.

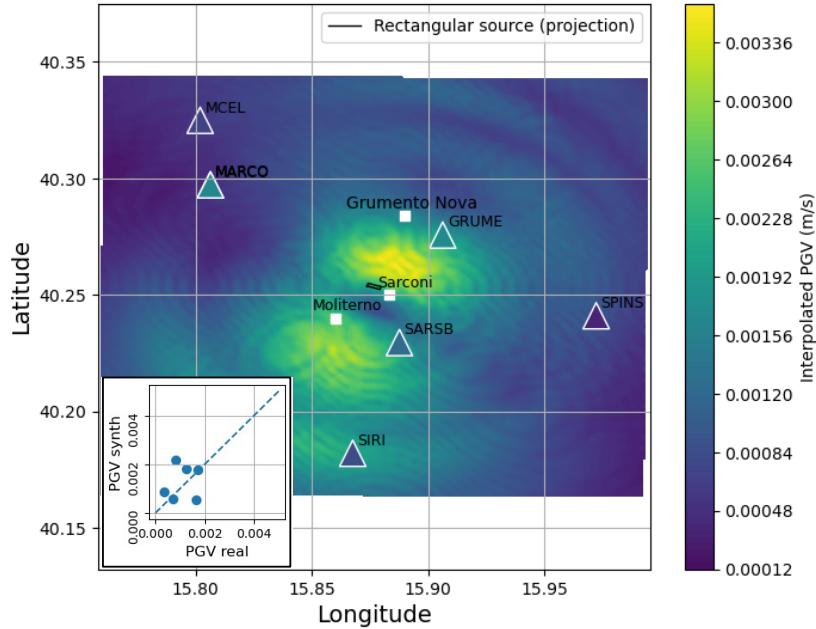


Figure 6.7: PGV simulation for the $M_W=3.4$ earthquake using the inverted focal mechanism and nucleation at the hypoDD location. Triangles indicate real seismic stations, colored by observed PGV for comparison with the synthetic results (color scale in the background).

The asymmetric PGV pattern arises from the oblique rake, which adds a horizontal slip component to the rupture and produces an uneven radiation pattern, focusing shear energy toward one side of the fault. When the rake is set to -90° , representing pure normal faulting, this horizontal component disappears and the PGV field becomes symmetric (Figure 6.9). In addition, the nucleation point strongly modulates shaking through rupture directivity: initiation at the northwestern edge channels energy toward the southeast, while a southeastern nucleation point enhances shaking toward the northwest. These effects imply that, for the same magnitude, the spatial distribution of ground motion—and the exposure of nearby towns—depends sensitively on both rake and nucleation position.

Using the empirical relationships from Wells & Coppersmith (1994) [130], a $M_w = 7.1$ earthquake would require a rupture length of approximately 46.8 km. In contrast, the seismically active fault segment illuminated in this study extends only about 3 km. This difference indicates that the analyzed segment is far too short to accommodate an earthquake of such magnitude. While the historical $M_w = 7.1$ event may have involved a broader fault system, potentially linking multiple segments, the cluster we focus on is clearly insufficient to account for such a large rupture.

6.5 Conclusions

I investigated the spatiotemporal patterns and source characteristics of clustered seismicity in the Val d’Agri region from 2021 to 2023 by integrating waveform similarity based clustering, double-difference relocation, focal mechanism analysis, and ground motion simulations.

The waveform similarity-based clustering revealed distinct groups of earthquakes, most of which align along steeply dipping normal faults consistent with the regional NE–SW extensional stress regime. The most active and coherent structure was delineated by cluster 3, located near Pertusillo Lake in the Monti della Maddalena Fault System.

The focal mechanisms confirm the dominance of normal faulting and suggest a southwestward dip direction. Despite some variability in strike for lower-quality

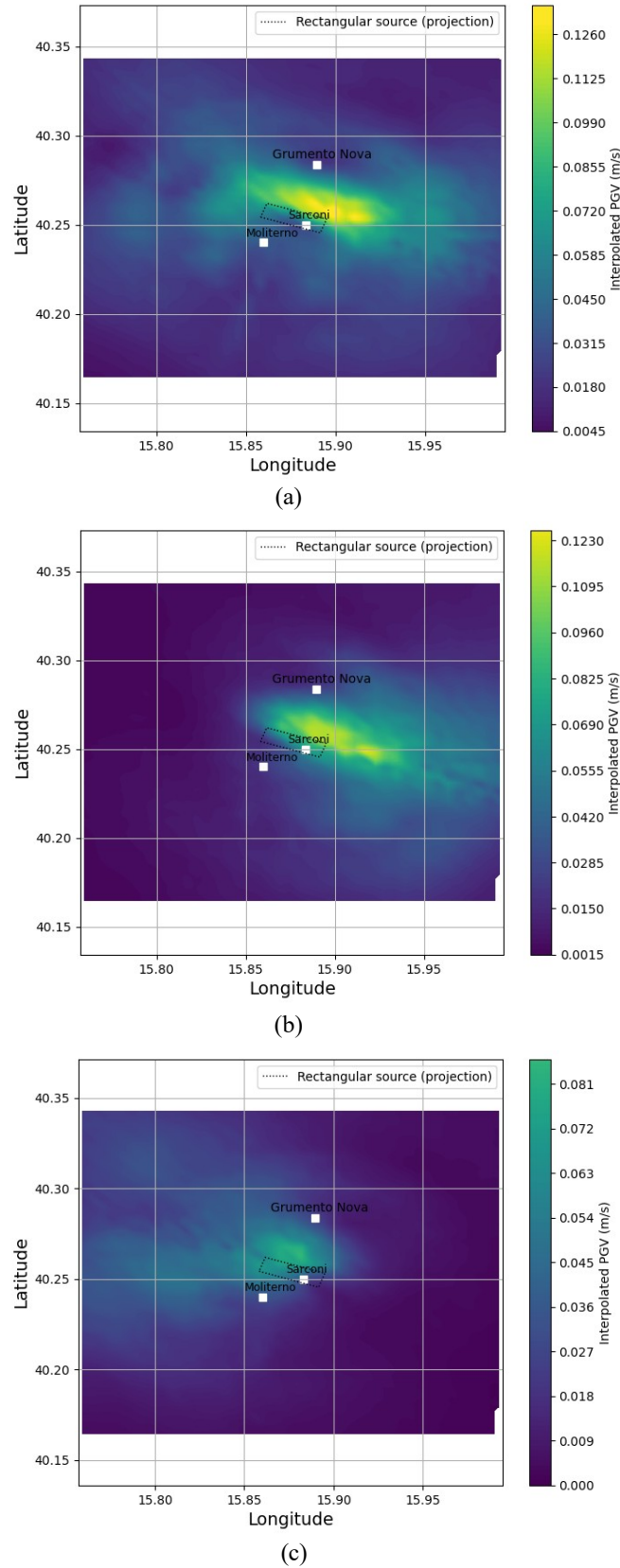


Figure 6.8: Hypothetical $M_w = 4.7$ scenarios using the same source mechanism and extended rupture over the full structure illuminated by cluster 3. Nucleation is placed at the center (a), northwestern (b), and southeastern (c) edge of the fault. The asymmetric PGV patterns reflect the influence of rupture directivity and the rake angle.

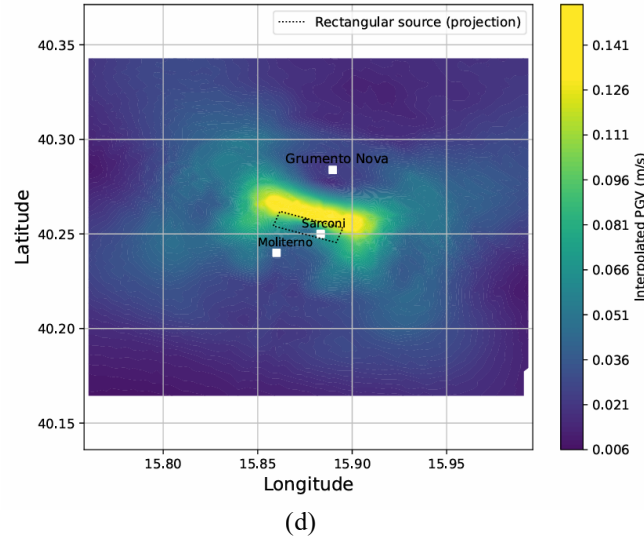


Figure 6.9: Hypothetical $M_w = 4.7$ scenarios using the same source mechanism and extended rupture over the full structure illuminated by cluster 3, but with rake = -90° (vertical), representing pure normal faulting. Nucleation is placed at the center.

mechanisms, the overall faulting style is consistent within each cluster, especially in cluster 3.

A well-defined seismic sequence began in July 2023 within cluster 3, with a mainshock of Mw 3.4. The sequence shows spatial migration of epicenters after the mainshock, possibly due to stress redistribution or fluid diffusion.

Ground motion simulations using a finite-fault model and Pyrocko Seismosizer reproduce observed PGV patterns and highlight the influence of rupture geometry. A hypothetical Mw 4.7 rupture of the full structure produces significantly larger ground motions than the 2023 event. These results provide new insights into the geometry and behavior of active faults in the Val d’Agri region and highlight the value of integrating high-resolution seismological analysis with physical modeling for seismic hazard assessment.

Chapter 7

Extending the DL-based detection workflow to other Italian Regions

7.1 Introduction

I have demonstrated that a notable strength of machine-learning-based seismic detection lies in its capacity to identify micro-earthquakes that would otherwise be overlooked by conventional approaches. Consequently, it appears highly pertinent to investigate regions characterized by low seismicity. This chapter provides a concise overview of preliminary results obtained by employing an automated workflow for earthquake detection on two novel seismic datasets located in Italy. The objective of this analysis is to underscore the adaptability and resilience of Deep Learning-based workflows for monitoring microseismicity in regions exhibiting markedly diverse tectonic and seismic characteristics. This analysis further emphasizes the potential for efficient, automated earthquake detection even in areas with minimal background seismicity.

The first study area is Sardinia, a region characterized by intrinsically low seismicity and sparse historical earthquakes (Meletti et al., 2021 [134]). The application of the detection workflow here provides insights into background seismic activity and enables the identification of previously uncataloged events, offering a systematic approach to detect and cluster microseismicity. The analysis allows a direct

comparison with historical and instrumental seismicity datasets, providing a first validation of the workflow in a quiescent tectonic setting.

The second area is the Ivrea-Verbano Zone (IVZ) in northern Italy, monitored by the dedicated DIVEnet seismic network within the framework of the ICDP DIVE project (Drilling the Ivrea-Verbano zone, www.dive2ivrea.org, <https://www.icdp-online.org/projects/by-continent/europe/dive-italy/>). This region is characterized by low local seismicity, but also by significant tectonic complexity associated with the Western Alps and the exposed Ivrea Geophysical Body. The application to the DIVEnet dataset aims to provide an automated, reproducible workflow capable of detecting events down to low magnitudes and characterizing their spatial distribution, further demonstrating the generalizability of the method to different geological contexts.

These preliminary analyses aim to demonstrate the potential of an automatic detection workflow not only as a research tool for seismic catalog enrichment but also as a foundation for near-real-time monitoring workflows.

7.2 Sardinian microplate

7.2.1 Introduction

Unlike other regions of the Italian peninsula, Sardinia is characterized by very limited and sporadic seismic activity (Meletti et al., 2021 [134]). The island is generally regarded as one of the least seismically active areas of the central Mediterranean, with only a handful of moderate historical earthquakes documented in both historical and instrumental catalogs (e.g. CPTI15 catalog from Rovida et al., 2022 [135], <http://terremoti.ingv.it>). This apparent quiescence has led to relatively limited seismic monitoring in the region: the permanent seismic networks deployed by INGV and other institutions provide only sparse coverage, insufficient to capture the full spectrum of microseismicity that might be occurring.

As a consequence, no comprehensive or well-constrained earthquake catalog currently exists for Sardinia. Available datasets, such as CPTI15 for historical

earthquakes [135] and the INGV instrumental catalog, are reliable for moderate to large events but suffer from poor completeness at lower magnitudes. This lack of high-resolution information hinders our ability to characterize the current level of seismicity and to identify potentially active seismogenic structures within the Sardinian microplate.

The aim of this analysis is therefore twofold. First, we seek to exploit the broadband seismic recordings from the temporary *LiSard* experiment (mid-2016 to late 2018; see Magrini et al., 2020 [136]; Cammarano et al., 2025 [137]), which operated in Sardinia and provided the densest seismic coverage available during this time window, to produce an earthquake catalog with significantly improved completeness compared to existing compilations. Second, by applying advanced detection algorithms—specifically, Deep-Learning-based phase picking through PhaseNet—we aim to systematically identify and characterize even the smallest seismic events, which are crucial for resolving active fault geometries and constraining the geodynamic behavior of the region.

Beyond the local implications, this is a preliminary analysis aimed at testing the performance of the automatic detection workflow in a low-seismicity region, where conventional catalog-based approaches are typically insufficient. By integrating dense temporary seismic observations with modern machine-learning methods, we can bridge the gap between historical macroseismic evidence and the modern instrumental era, opening the way to a better understanding of the role of the Sardinian block within the broader geodynamic setting of the central Mediterranean.

7.2.2 Data and Methods

The temporary *LiSard* seismic experiment operated between July 2016 and September 2018 and consisted of 10 broadband stations (network code UT) uniformly distributed across Sardinia. To enhance spatial coverage, we additionally incorporated data from two permanent stations of the Italian National Seismic Network (IV) and from the Mediterranean network (MN).

The analysis period has been divided into two subintervals due to the relocation of station UT.001, which has been moved from La Maddalena (Northern part of

Sardinia) to the vicinity of Villacidro (South). The two time windows are therefore defined as: 1 July 2016 – 17 November 2017 and 18 November 2017 – 1 September 2018. All stations recorded three-component broadband signals on HH channels.

I apply the *PRN* workflow described in ???. I preprocess the waveforms with a band-pass filter in the frequency range 0.5–30 Hz. Compared to the Val d’Agri dataset, a lower frequency range is selected here, reflecting the longer source-to-receiver distances and the consequent reduction in high-frequency signal content. Phase picking is performed using the original PhaseNet neural network model (Zhu & Beroza, 2019 [2]). For location procedures, a dedicated 1D velocity model is derived from a regional 3D reference model, provided by Henrique Berger Roisenberg (personal communication).

A major challenge of this case study is the lack of a reference earthquake catalog for the investigated time window. Consequently, the interpretation of the results relies on the integration with the instrumental catalog of INGV (<https://terremoti.ingv.it/>), the historical seismicity catalog CPTI15 (Rovida et al., 2020 [138] [135]), and historical seismicity mapped by Meletti et al. (2021) [134], which provide information on seismogenic areas and background seismicity patterns in Sardinia.

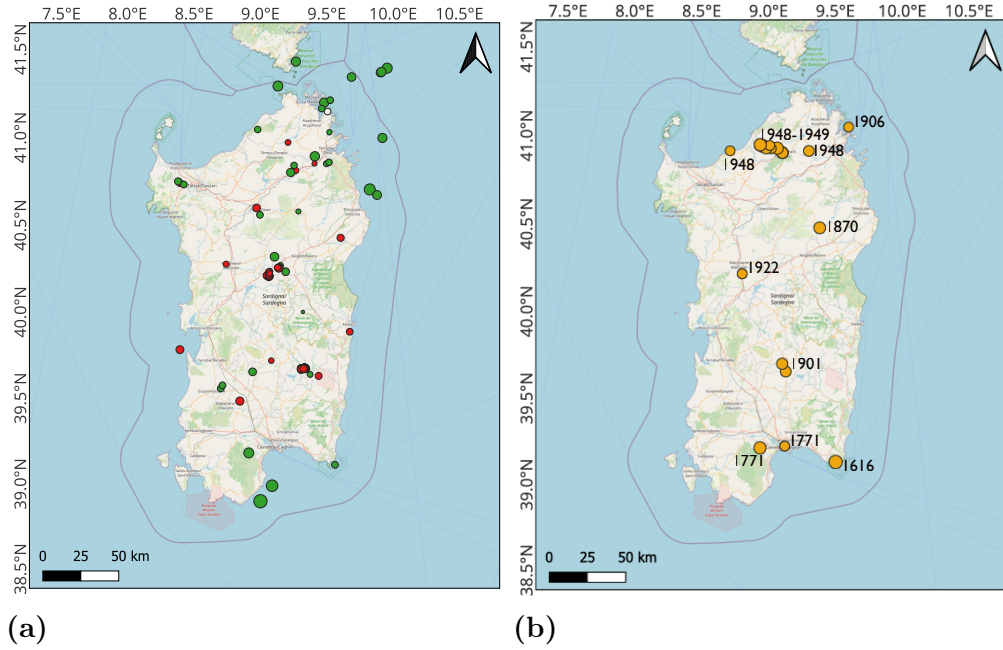


Figure 7.1: a) Map of epicenter distribution in Sardinia from the INGV instrumental catalog in the timespan 2000-2025. Green: earthquakes, red: quarry blasts. b) Map of historical earthquakes epicenters from Meletti et al. (2021) [134].

Figure 7.1a shows the spatial distribution of epicenters from the INGV instrumental catalog between 2000 and 2025. Events classified as earthquakes are shown in green, while quarry blasts are indicated in red. Figure 7.1b displays the historical seismicity mapped by Meletti et al. (2021) [134], which already includes earthquakes within the study area as reported in the CPTI15 catalog.

7.2.3 Results

The preliminary results of this analysis are shown in Figure 7.2.

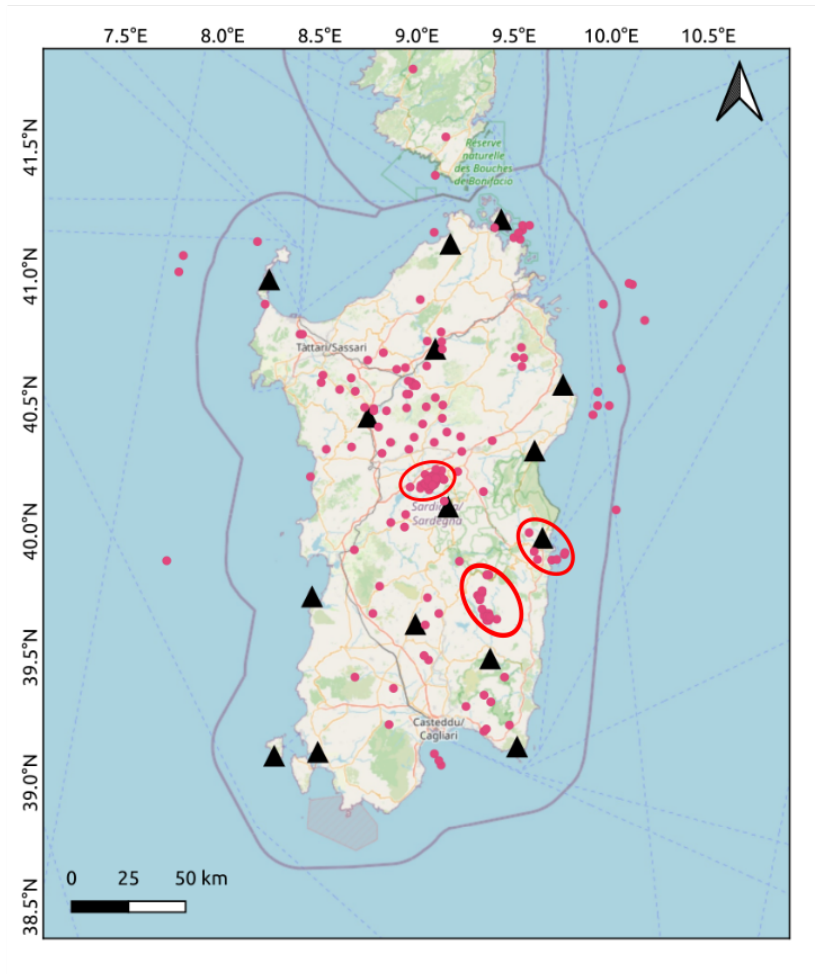


Figure 7.2: Pink dots represent the epicenter distribution resulting from this analysis. The timespan is Jul 2016 - Sep 2018. Black triangles represent the seismic stations used for this analysis, both from the UT temporary network (Lisard experiment) and from the permanent networks (IV, MN, TV). Red circles identify spatial clusters that are commented in the text.

The analysis identifies 181 earthquakes, with a minimum of six picks per event (3 P and 3 S). The spatial distribution of the detected clusters closely resembles that shown in Figure 7.1. At this stage, quarry blasts have not yet been explicitly identified; however, based on the INGV catalog shown in Figure 7.1a, it can be reasonably inferred that the spatial clusters highlighted by red circles in the map in Figure 7.2 are mainly composed of quarry blast events.

7.2.4 Discussion and Conclusions

This analysis allows the identification of approximately two years of previously uncatalogued regional seismicity. The main challenge of this work lies in applying a fully automatic detection workflow to an area lacking any reference catalogued earthquake for the period of interest, and characterized by a generally low seismicity level, with limited pre-existing information on earthquake occurrence. Indeed, the INGV instrumental catalog does not report any earthquakes within the study area and time window. To obtain reference points on potentially seismogenic zones, the catalog generated through this automatic approach is compared with both the historical seismicity and the INGV instrumental catalog covering different time spans.

The results show that the spatial distribution of seismicity detected by the automatic method are consistent with those identified in the INGV instrumental catalog. The two northeastern clusters overlap with seismically active areas present in both the INGV instrumental and historical catalogs, supporting the plausibility of our results. In the southern portion of the island, a few small spatial clusters and scattered seismicity are also observed. Similar patterns are evident in the INGV catalog, and the historical catalog includes a few epicenters in the same southern region. The method, however, cannot automatically discriminate events that are likely quarry blasts. Therefore, the next step will be to inspect the signals and identify quarry blasts based on the P-to-S amplitude ratio, which is higher for quarry explosions than for natural earthquakes (Wang et al., 2021 [139]). Nevertheless, we can reasonably assume that the clusters corresponding to quarry blasts reported by INGV—such as those located in the central and southeastern parts of the island (red circles in Figure 7.2)—are also quarry blasts in our catalog.

Assuming that the central clusters mainly correspond to quarry blast activity, the seismicity identified through this analysis appears to be primarily concentrated offshore, as found by Meletti et al., (2021) [134] in the historical catalog. Within the island, instead, the events are more scattered and sparse, suggesting a lower level of background seismicity or limited detection capability due to the current station coverage.

Although the results are promising, this analysis should be considered preliminary.

Further work will include the identification of quarry blasts and the integration of data from the recently installed 4P network seismic stations to extend the temporal coverage. These steps will enable a more comprehensive and temporally consistent analysis of the regional seismicity and its potential correlations with local geological structures.

7.3 The Ivrea Verbano Zone

7.3.1 Introduction

The Ivrea–Verbano Zone (IVZ) represents one of the most remarkable geological environments of the Alps, amounting to a portion of the continental lower crust uplifted and exposed at the surface. Located along the southwestern margin of the Alps, it is characterized by the juxtaposition of deep crustal and upper mantle rocks uplifted during the Alpine orogenic cycle. The IVZ forms the exposed part of the Ivrea Geophysical Body, a prominent gravity and seismic velocity anomaly that has been a natural laboratory for studying crust–mantle transition processes and the rheological structure of the continental lithosphere (e.g., Quick et al., 1992 [140]; Schmid et al., 1987 [141]; Pistone et al., 2020 [142]).

Within the framework of the ICDP DIVE project (Drilling the Ivrea-Verbano zone, www.dive2ivrea.org, <https://www.icdp-online.org/projects/by-continent/europe/dive-italy/>), two fully cored boreholes were drilled between 2022 and 2024 at Ornavasso and Megolo, approximately 7 km apart, to directly sample and characterize this crustal section. To monitor both the regional seismicity and any potential microseismicity induced by drilling operations, a dedicated temporary network (DIVEnet) consisting of 11 broadband and short-period stations was deployed around the boreholes. The network provides high-quality continuous waveform data in an area otherwise characterized by low natural seismicity and limited permanent network coverage.

This section presents the first results of applying a fully automated, DL-based earthquake detection workflow to the DIVEnet recordings. The analysis aims to evaluate the performance of this methodology in a low-seismicity context and to

assess its ability to improve detection completeness compared to the reference catalog. In particular, this work aims to investigate the spatial distribution of newly detected events and their relationship to known geological structures, such as the Insubric Line, providing preliminary insights into the microseismicity of the Ivrea–Verbano region.

7.3.2 Data and Methods

I analyse continuous seismic waveforms recorded from November 2021 to April 2024 for all stations of the DIVE network, along with two stations from the Swiss (CH) network and one from the Italian national network (IV) operating within the study area. Depending on availability, both HH and EH channels were used.

The analysis is performed using the PRN workflow, applying a bandpass filter between 1 and 45 Hz during preprocessing. The events have been relocated using hypoDD [113].

7.3.3 Results

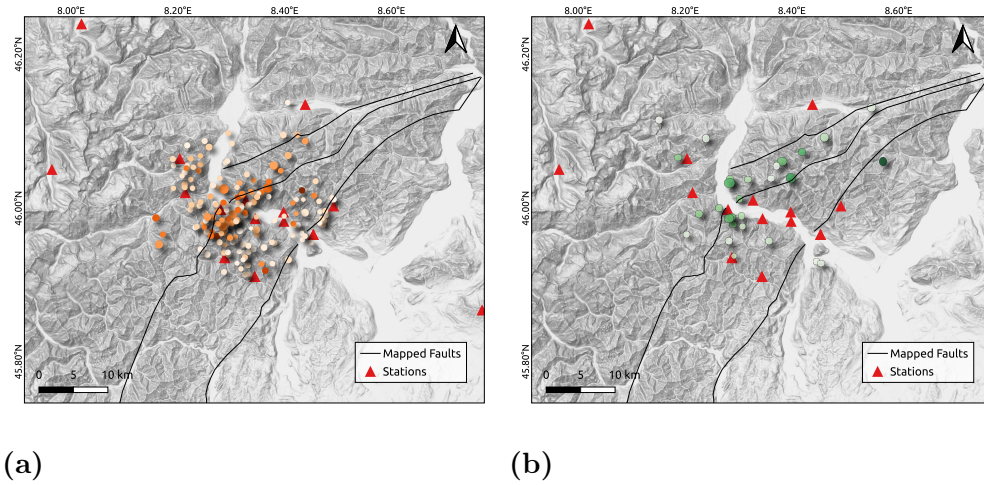


Figure 7.3: a) Orange dots represent the epicenter distribution resulting from the analysis made with the PRN workflow. Black triangles represent the seismic stations used for this analysis, both from the DIVE temporary network and from the permanent networks (IV, CH). b) Map of epicenters distribution in the same area from the reference catalog [143] in the same time window.

Figure 7.3 shows the spatial distribution of epicenters detected in this study, together with those reported in the reference catalog for the same time window. The PRN workflows detects 182 earthquakes inside the DIVEnet area (I consider latitude between 45.9 and 46.2, longitude between 8.15 and 8.6), 28 of which correspond to events already listed in the reference catalog from Confal et al., (2025) [143], while the remaining are newly detected events.

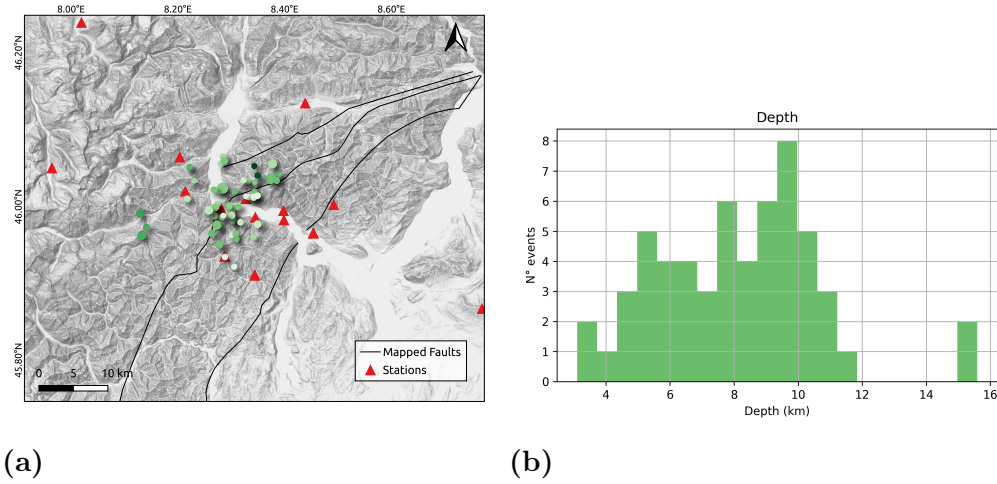


Figure 7.4: a) Green dots represent the epicenter distribution resulting from the hypoDD relocations. Black triangles represent the seismic stations used for this analysis, both from the DIVE temporary network and from the permanent networks (IV, CH). b) Depth histogram of the hypocenters calculated with hypoDD.

HypoDD relocates 57 events, mostly distributed in the proximity of the Insubric Line 7.4.

The newly detected earthquakes are generally of small magnitude and distributed across the same seismogenic volume as the reference events, indicating consistent performance of the detection and location procedures. Some little variations in the magnitude estimation can be due to the different picks set.

7.3.4 Discussion and Conclusions

The overall seismicity rate in the investigated area is low, consistent with previous observations. However, this analysis identifies approximately twice as many events

as reported in the reference catalog, demonstrating the enhanced sensitivity of the workflow adopted here.

Interestingly, several of the newly detected events are located near the Insubric Line, a major tectonic boundary currently considered geologically inactive. While these detections do not imply reactivation, they highlight the potential of automated detection methods to reveal microseismicity that might otherwise remain undetected.

The histogram of hypocentral depths provides further insight: in particular, the depth at which seismicity tapers off may indicate the approximate location of the brittle–ductile transition in the crust. A quantitative comparison with thermal and rheological models is left for future work, but the observed cutoff depth offers a first-order indication of the depth range of brittle deformation in the study area.

Overall, the results confirm and enhance those obtained from the reference catalog, suggesting that the detection method employed here represents a robust and efficient approach for improving detection completeness and phase-picking accuracy in low-seismicity regions.

This study demonstrates the potential of applying modern, data-driven detection and phase-picking methods to improve the completeness and accuracy of local seismic catalogs.

Future work will extend this analysis to the entire available time span of continuous data (more than 3 years), which will allow the detection and relocation of a significantly larger number of events. By combining these detections with high-precision relocation techniques, such as the double-difference method, it will be possible to better constrain the geometry of active or potentially reactivated structures. This will ultimately contribute to a more detailed understanding of the seismotectonic framework of the study area and its temporal evolution.

Chapter 8

Conclusions

In this doctoral work, I demonstrate how modern machine learning methods can significantly enhance the resolution and interpretability of seismicity patterns in the Val d’Agri region of Southern Italy. By applying the deep neural network PhaseNet (Zhu & Beroza, 2019 [2]) for automatic seismic phase picking, I generate enriched earthquake catalogs characterized by greater event detection rates and a higher number of phase picks per event (Chapter 4). Compared to traditional workflows based on STA/LTA detection and manual revision, the machine-learning-based approach nearly doubled both the number of identified events and the average number of arrivals per event. The magnitude of completeness is also improved in the catalogs obtained from the application of the new workflows (Chapter 5), further confirming the quality of the performance. This increase in data volume and quality is beneficial for the reliability of subsequent hypocentral location and source analysis.

The automated nature of the workflow ensures not only a substantial reduction in analyst time and effort but also greater internal consistency, as the same detection criteria can be applied uniformly over time or revised systematically when needed. The enhanced sensitivity of the approach allows for a more detailed observation of temporal changes in seismicity, including seasonal variations in background seismic noise and possible correlations with anthropogenic influences. In particular, the occurrence of microseismicity appears to be modulated by the seasonal charge and discharge cycles of the Pertusillo water reservoir. While a clear correlation emerges,

further investigation is necessary to fully quantify and model the poroelastic and hydromechanical effects potentially responsible for this behavior.

The machine-learning-based workflow presented herein demonstrates its efficacy in characterizing low-seismicity regions as well, such as the Sardinia and the Ivrea-Verban Zone datasets (see Chapter 7). The workflow enabled the identification of extensive microseismicity previously undetected due to sparse monitoring or limited sensitivity. Its automated and systematic nature ensures consistent detection criteria across different networks and facilitates comparative studies between regions, highlighting the workflow’s scalability and applicability to a variety of seismic settings..

Building upon the enriched catalogs of the tectonically active Val d’Agri, I then explore the spatiotemporal evolution and source characteristics of clustered seismicity through waveform similarity-based clustering, double-difference relocation, focal mechanism analysis, and finite-source ground motion modeling (Chapter 6). The clustering reveals several well-defined groups of events, predominantly aligned along steeply dipping normal faults consistent with the NE–SW extensional stress regime of the region. Among these, a particularly active structure, near the Pertusillo Lake, exhibits coherent geometry and hosted a seismic sequence in July 2023 initiated by a M_w 3.4 mainshock. The spatial migration of epicenters following the mainshock suggests stress transfer or fluid-driven processes at depth.

Focal mechanism solutions indicate predominantly normal faulting with southwestward dip directions, and the coherence of mechanisms among earthquakes belonging to the same cluster supports the reliability of the classification. Ground motion simulations replicate the observed peak ground velocity patterns and highlight the importance of rupture geometry in controlling ground motion distribution. A hypothetical M_w 4.7 rupture encompassing the entire fault segment imaged by Cluster 3 suggests the potential for moderate ground shaking, although empirical scaling relationships indicate that this structure alone is unlikely to host larger earthquakes (e.g., $M_w > 7$).

The results I present in this dissertation open several paths for future research, aimed at deepening our understanding of earthquake processes in the Val d’Agri region and extending the proposed methodologies to broader contexts. Further

improvements in the automatic detection and characterization of seismic events are anticipated. A key direction involves the implementation of a quasi real-time or daily monitoring workflow based on the tested deep learning algorithms. However, the operational requirements of the Val d'Agri monitoring system require an updated catalog on a daily basis rather than immediate event solutions. For this reason, an implementation based on daily processing of the previous 24 hours of data represents a practical and computationally efficient solution.

The implementation of such systems could enhance the responsiveness and consistency of seismic surveillance, particularly in regions of anthropogenic activity or complex tectonic settings such as the Val d'Agri.

A promising future step involves the retraining of the neural network model using a locally curated, high-quality dataset. The availability of densely instrumented networks and well-constrained event catalogs in the Val d'Agri region provides an ideal foundation for this task. A region-specific training set would allow the model to better capture the local noise characteristics and waveform features distinctive of the Southern Apennine environment, and improve both detection accuracy and phase-picking precision.

An additional line of investigation worth pursuing is the quantitative investigation of the relationship between microseismicity and hydrological variations, particularly those associated with the water reservoir loading. While a qualitative correlation with water level fluctuations in the Pertusillo reservoir has been observed, a more rigorous assessment is needed to quantify the effects of poroelastic stress changes and delayed pore pressure diffusion on the timing and distribution of earthquakes. This can be achieved through the coupling of high-resolution seismic catalogs with numerical models of fluid diffusion and stress transfer.

Moreover, the detailed structural insights obtained through waveform clustering and double-difference relocation highlight the importance of high-precision analysis of micro-seismicity to resolve active fault geometries. This approach could be extended to other seismically active regions with dense monitoring networks, offering a valuable tool for improving local seismic hazard models.

Overall, this research underscores the value of integrating machine learning, high-resolution waveform analysis, and physical modeling to improve our understanding

of fault systems and to support seismic hazard assessment in tectonically and anthropogenically active regions.

Appendix A

Supplementary Information

Network	Station	Locat	Channel	Latitude	Longitude	Elevation_m	Depth_m	SensorDescription	Scale	ScaleFreq	ScaleUnits	SampleRate_Hz
GE	MARCO		HH*	40.29724	15.80632	933		1 GFZ-GE1993-ST5-2.5/g=1500	629145000	0.02	M/S	100
IV	MMN		HH*	39.890961	15.990414	921		0 NANOMETRICS TRILLIUM-40S	1179650000	0.2	M/S	100
IV	ACER		HH*	40.7867	15.9427	690		1 NANOMETRICS TRILLIUM-40S	1179650000	0.2	M/S	100
IV	SCHR		EH*	40.19924	16.0759	968		0 LENNARTZ LE3D-5S	503316000		1 M/S	100
IV	MTSN		HH*	40.2663	15.7515	1056		0 NANOMETRICS TRILLIUM-120S	480400000		1 M/S	100
IV	CRAC		EH*	40.3814	16.435	384		0 LENNARTZ LE3D-5S	503316000		1 M/S	100
IV	SIRI		HH*	40.1821	15.8675	1063		0 NANOMETRICS TRILLIUM-40S	1500000000	0.2	M/S	100
IV	ORI		HH*	40.05096	16.450405	375		0 NANOMETRICS TRILLIUM-40S	1179650000	0.2	M/S	100
IV	SLCN		HH*	40.39	15.63277	986		0 NANOMETRICS TRILLIUM-40S	1179650000	0.2	M/S	100
IV	PTRP		HH*	40.52148	16.06124	1077		1 NANOMETRICS TRILLIUM-120S	377802000		1 M/S	100
IV	MCEL		HH*	40.3249	15.8019	960		0 NANOMETRICS TRILLIUM-40S	1500000000	0.2	M/S	100
IV	MIGL		HH*	40.6044	16.441	440		0 NANOMETRICS TRILLIUM-40S	1500000000	0.2	M/S	100
IV	SALB		HH*	39.8772	16.3459	1200		0 NANOMETRICS TRILLIUM-40S	1500000000	0.2	M/S	100
IX	STN3	2	EH*	40.53	15.6515	832		0 S13/208/g=208	22341600000	5	m/s	125
IX	PGN3	2	HH*	40.5722	15.7967	882		0 TRILLIUM 40S/1500/g=1500	2798510000		1 m/s	125
IX	SRN3	2	HH*	40.4861	15.458	1067		0 CMG-40T/800/g=800	1492540000		1 m/s	125
VD	SARCL	0	CH*	40.221487	15.917465	881		14 Trillium Compact	299470000		1 M/S	250
VD	SARSB	0	CH*	40.229581	15.887608	745		6 Trillium Compact	298993000		1 M/S	250
VD	SPINS	0	CH*	40.241083	15.972221	882		0 Trillium Compact	298993000		1 M/S	250
VD	VIGGN	0	CH*	40.34209	15.951559	1043		6 Trillium Compact	298993000		1 M/S	250
VD	ARMEN	0	CH*	40.328033	16.034446	880		6 Trillium Compact	298993000		1 M/S	250
VD	GRUME	0	CH*	40.275704	15.906211	602		6 Trillium Compact	298993000		1 M/S	250
VD	MONTM	0	CH*	40.278168	16.040405	652		6 Trillium Compact	298993000		1 M/S	250
TP	TR01	0	EH*	40.48082	16.018	1200	0		335544320		1 M/S	200
TP	TR02	0	EH*	40.489328	15.8749	1091	0		335544320		1 M/S	200
TP	TR03	0	EH*	40.58723	16.00579	511	0		335544320		1 M/S	200
TP	TR04	0	EH*	40.49684	16.21562	614	0		335544320		1 M/S	200
TP	TR05	0	HH*	40.38358	16.09527	718	0		1258291200	0.1	M/S	200
TP	TR06	0	EH*	40.29292	16.12164	829	0		335544320		1 M/S	200
TP	TR07	0	EH*	40.44513	15.956123	850	0		335544320		1 M/S	200
TP	TR08	0	EH*	40.39426	16.01198	723	0		335544320		1 M/S	200
TP	TR09	0	EH*	40.399955	16.195975	735	0		335544320		1 M/S	200
TP	TR10	0	EH*	40.43226	16.09583	877	0		335544320		1 M/S	200
TP	TR11	0	EH*	40.474294	16.103167	1212	0		335544320		1 M/S	200
TP	TR12	0	EH*	40.411922	16.03365	896	0		335544320		1 M/S	200
VA	PADH	0	EH*	40.3144444	15.9025	470	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	ABRE	0	EH*	40.5	15.75944	1542	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	BRIN	0	EH*	40.49	15.613611	935	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	CASE	0	EH*	40.552222	16.072222	703	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	FORE	0	EH*	40.3075	15.863056	570	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	GRUE	0	EH*	40.298611	15.893889	550	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	LAUE	0	EH*	40.439444	15.943333	1090	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	MARE	0	EH*	40.391667	15.810556	1061	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	MOLE	0	EH*	40.25	15.829167	1153	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	SPIE	0	EH*	40.218056	15.978889	1319	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	TITE	0	EH*	40.556667	15.692222	676	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	VIGE	0	EH*	40.334722	15.900556	816	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	GUPE	0	EH*	40.367777	16.068611	674	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	PADE	1	HH*	40.3144444	15.9025	610	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	MANE	0	HH*	40.458611	15.734444	1188	0	0 Solgeo Datalogger	1258299589	100	M/S	200
VA	ARME	0	HH*	40.315278	16.031111	989	0	0 Solgeo Datalogger	1258299589	100	M/S	200
VA	RAPE	0	EH*	40.223239	16.092717	769	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	LAPE	0	EH*	40.291103	15.954886	568	0		335544320		5 M/S	200
VA	MOME	0	EH*	40.331944	15.982778	1152	0	0 Lennartz LE-3D/5 seismometer	335544320		5 M/S	200
VA	MOMN	0	HH*	40.330594	15.983422	1157	0	0 Solgeo Datalogger	1258299589	100	M/S	200
VA	TRAE	0	EH*	40.329942	15.755306	714	0		335544320		5 M/S	200

Figure A.1: Characteristics of the seismic stations used in this study.

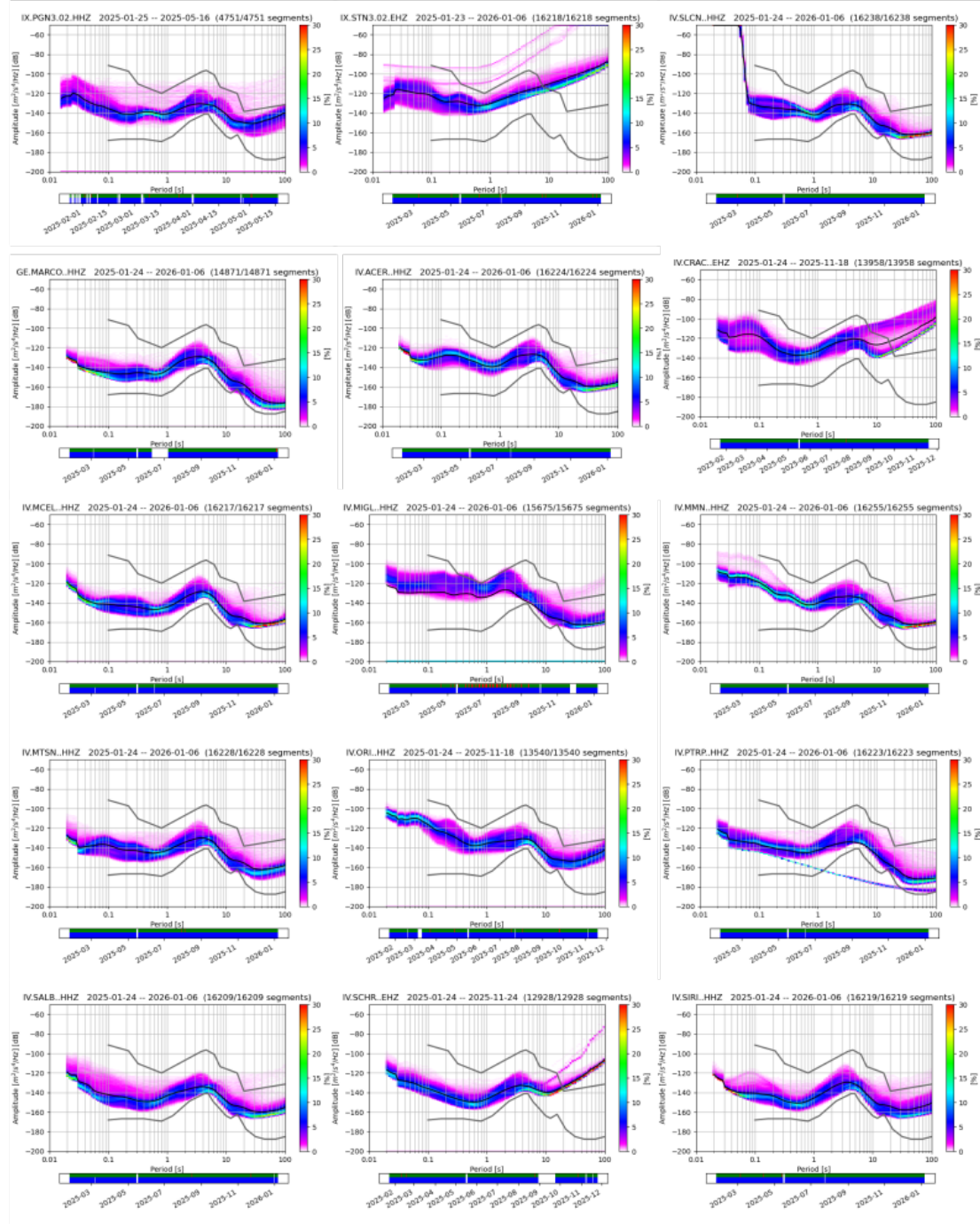
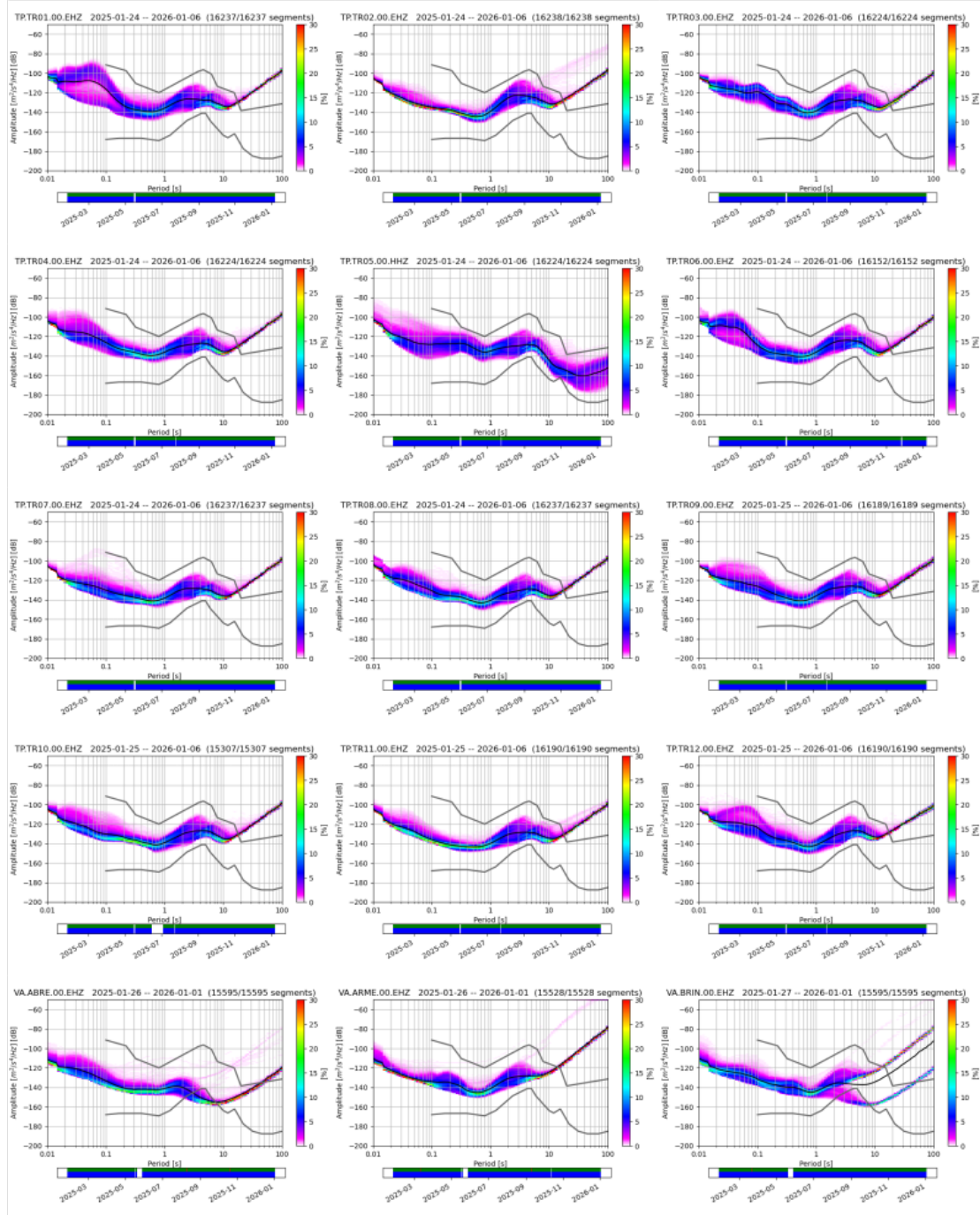
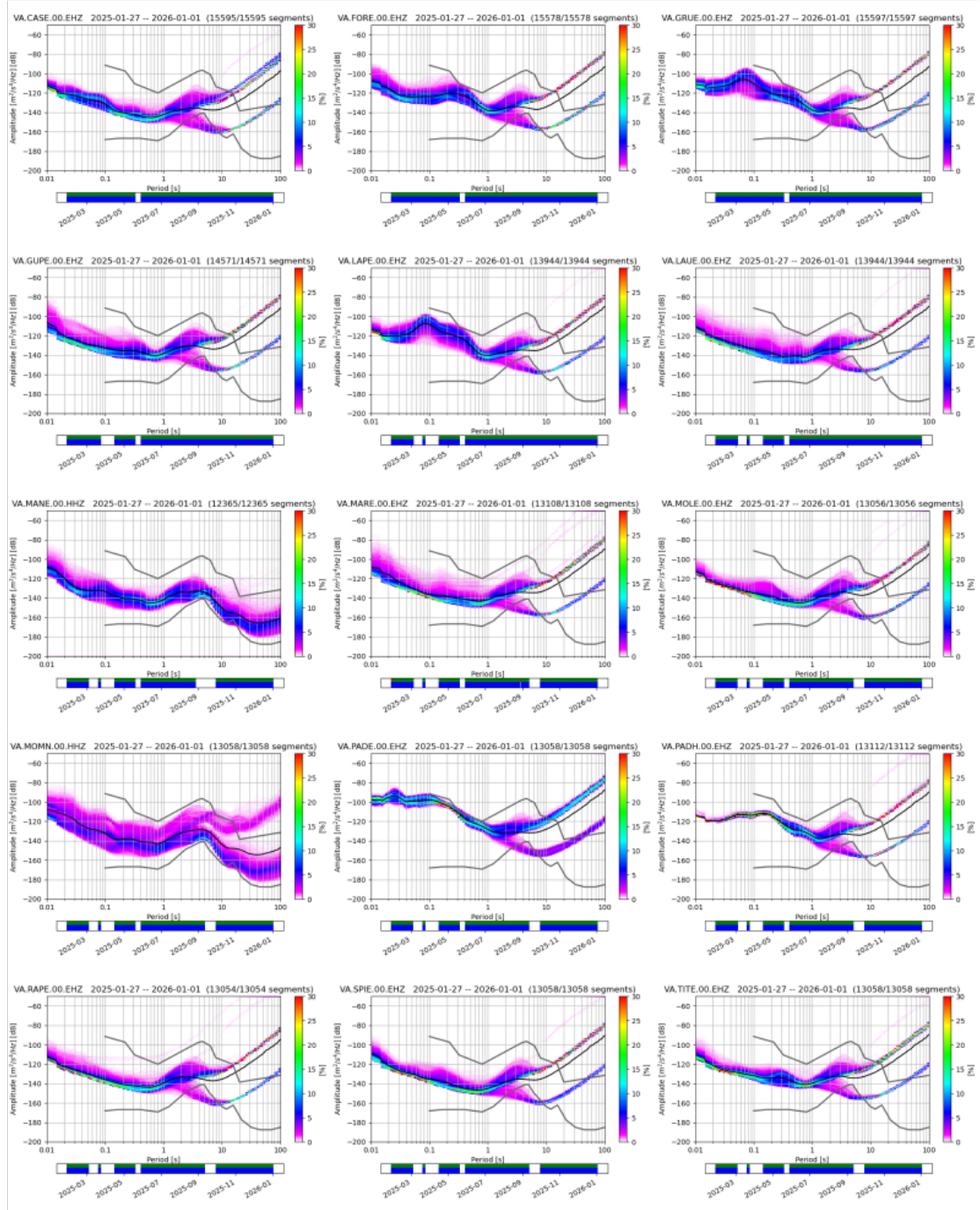


Figure A.2: Probabilistic Power Spectral Density (PPSD) [40] of seismic noise for each station, computed over around one-year. The PPSDs are shown together with the global low- and high-noise models of Peterson (1993) [144]. For some stations, the spectra rise above the high-noise model at long periods (~ 10 s and above); this is due to the limited frequency response of the sensors and electronic noise, rather than actual ground motion. Note that the analysis presented in this study focuses on the frequency range 1–45 Hz, so these long-period features do not affect the results. The bar below the actual PPSD plots is a visualization of the data basis for the PPSD. The top row shows data fed into the PPSD, green patches represent available data. The bottom row in blue shows the single psd measurements that go into the histogram.





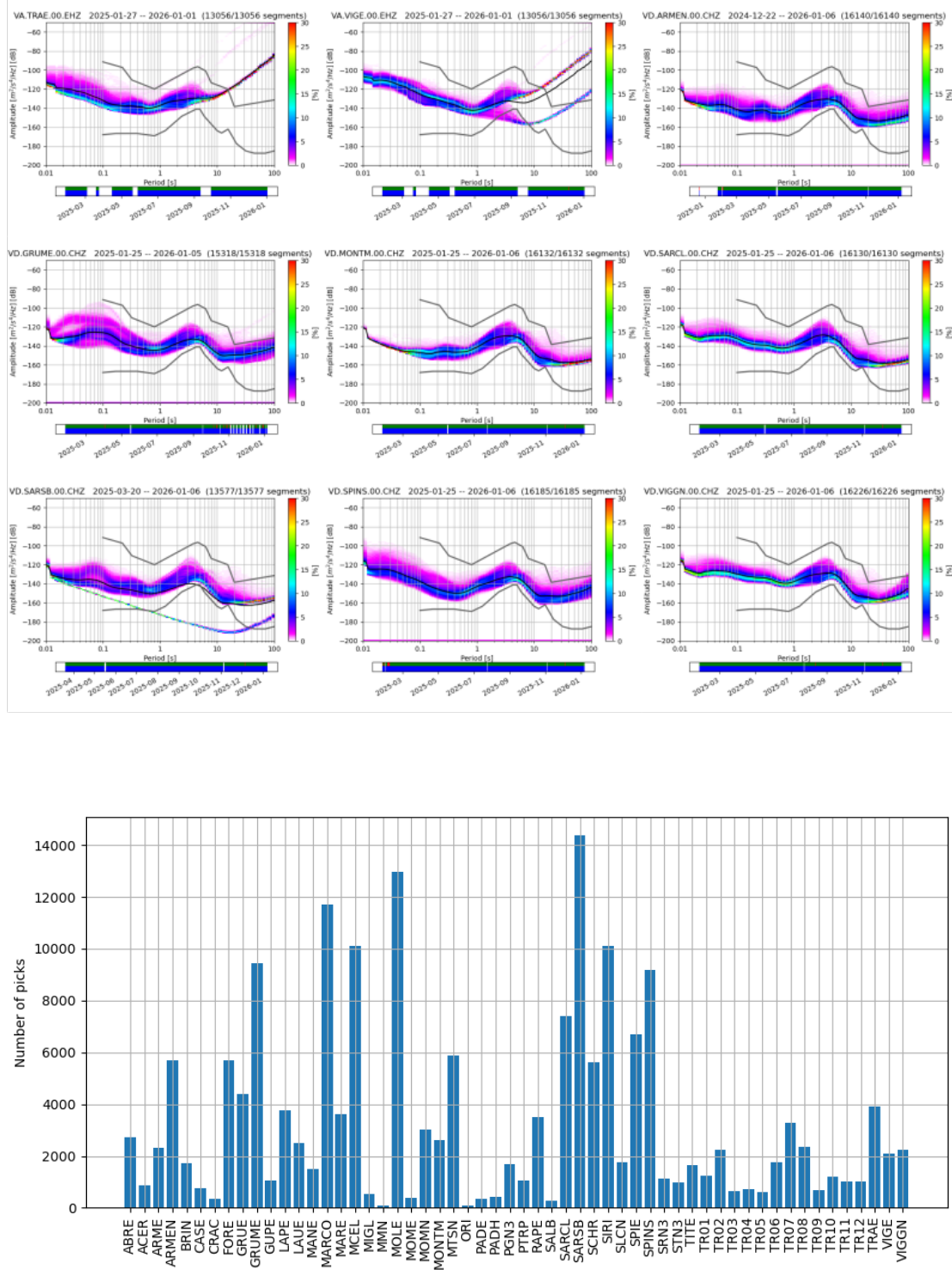


Figure A.3: Number of picks per station in the PRN catalog.

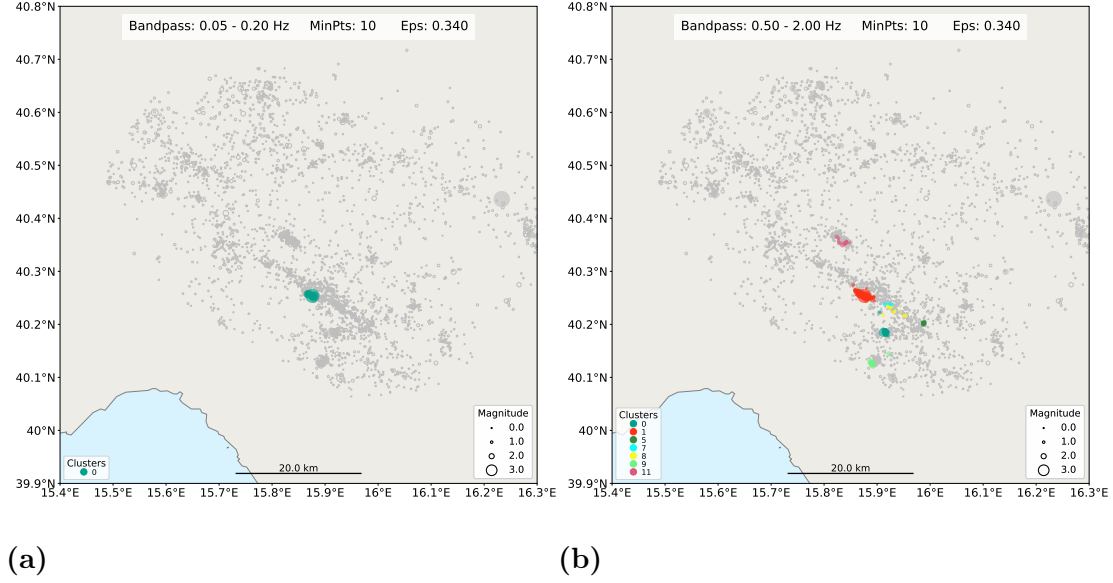


Figure A.4: Maps of clustered earthquakes according to waveform similarity varying the frequency range: a) 0.05-0.2 Hz and b) 0.5-2 Hz.

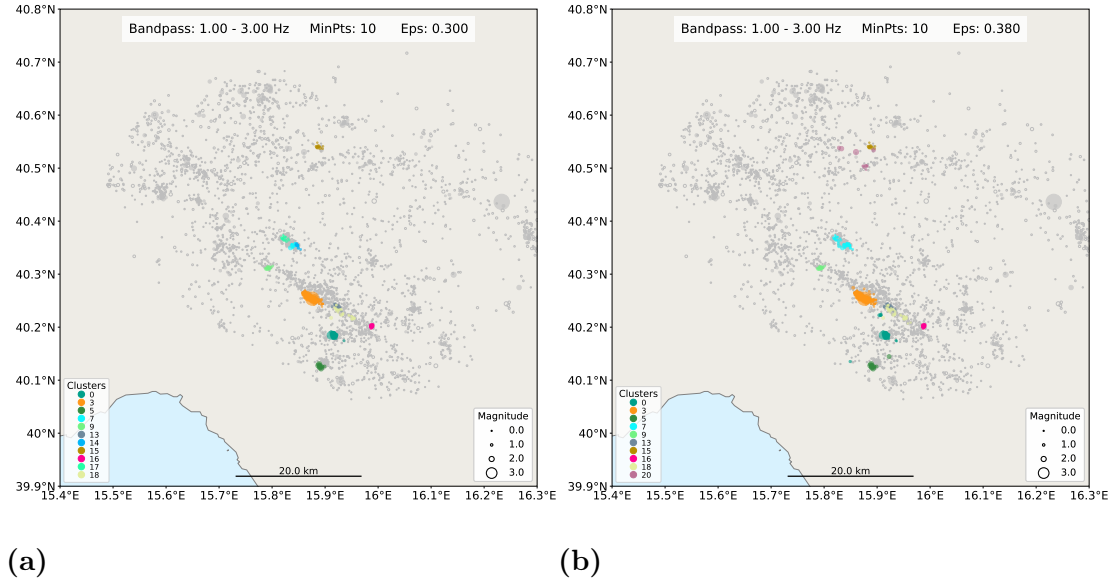


Figure A.5: Maps of clustered earthquakes according to waveform similarity in frequency range 1-3 Hz, varying ϵ as a) 0.30 and b) 0.38.

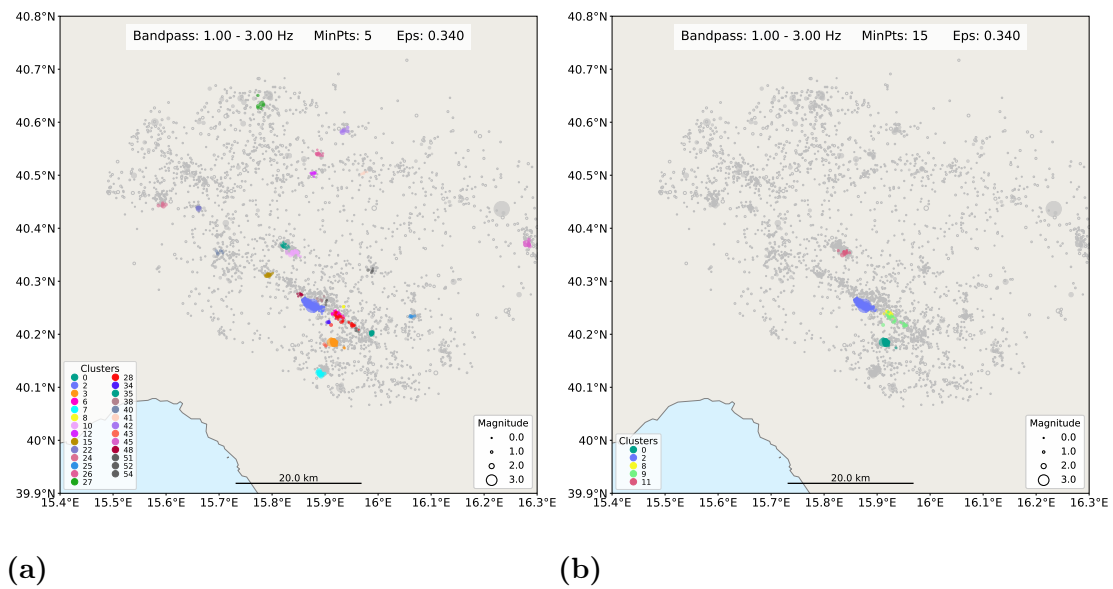


Figure A.6: Maps of clustered earthquakes according to waveform similarity in frequency range 1-3 Hz, varying N_{min} as a) 5 and b) 15.

Origin time (UTC)	Lat	Lon	Depth (km)	M _L	Strike1	Dip1	Rake1	Strike2	Dip2	Rake2	Quality
2021/08/14 09:29:28	40.4363	16.2332	13.3	3.5	125.4	84.8	100.4	241.8	11.6	26.9	A
2023/05/12 01:02:30	40.0987	15.9287	21.0	2.5	311.5	72.8	-119.8	194.3	34.0	-31.8	B
2023/07/03 05:16:27	40.2531	15.8767	3.2	2.2	290.8	80.4	126.0	33.6	37.1	16.0	B
2023/07/03 05:18:43	40.2529	15.8782	3.4	2.8	133.9	52.6	-100.2	330.4	38.6	-76.9	A
2023/07/03 07:36:11	40.2538	15.8759	3.4	3.3	107.3	63.3	-108.0	323.2	31.8	-58.4	A
2023/07/08 00:18:04	40.2581	15.8671	3.4	2.2	104.8	85.8	-113.5	5.3	23.9	-10.3	B
2023/07/11 06:53:56	40.2567	15.8668	3.3	2.4	101.3	80.6	-121.4	356.3	32.6	-17.7	B
2023/07/11 10:01:58	40.2518	15.8789	3.4	2.2	283.2	82.5	135.5	20.5	46.0	10.4	B
2023/09/01 19:47:59	40.1849	15.9116	5.0	2.6	116.4	53.6	-120.8	341.6	46.2	-55.2	A

Table A.1: Focal mechanism solutions with associated parameters.

Bibliography

- [1] Katie M. Keranen and Matthew Weingarten. Induced seismicity. *Annual Review of Earth and Planetary Sciences*, 46(Volume 46, 2018):149–174, 2018.
- [2] W. Zhu and G. C. Beroza. Phasenet: A deep-neural-network-based seismic arrival time picking method. *Geophys. J. Int.*, 216:261–273, 2019.
- [3] Zonghui Dai, Lianqing Zhou, Xuhui Hu, Junhao Qu, and Xia Li. Generalization of phasenet in shandong and its application to the changqing m4. 1 earthquake sequence. *Earthquake Science*, 36(3):212–227, 2023.
- [4] Andrea Morelli, Mario Anselmi, Thomas Braun, Stefania Danesi, Maddalena Errico, Daniela Famiani, Alexander Garcia, Irene Molinari, Maurizio Vassallo, and Lucia Zaccarelli. A state-of-the-art analysis system designed to monitor microseismicity at a local scale. *Annals of Geophysics*, 67(6):S661–S661, 2024.
- [5] Etta Patacca, Paolo Scandone, et al. Geology of the southern apennines. *Bollettino della Società Geologica Italiana*, 7:75–119, 2007.
- [6] Claudio Faccenna, Thorsten W Becker, Francesco Pio Lucente, Laurent Jolivet, and Federico Rossetti. History of subduction and back arc extension in the central mediterranean. *Geophysical Journal International*, 145(3):809–820, 2001.
- [7] Ingrid Hunstad, G Selvaggi, Nicola D’Agostino, Philip England, Peter Clarke, and M Pierozzi. Geodetic strain in peninsular italy between 1875 and 2001. *Geophysical Research Letters*, 30(4), 2003.

- [8] Rob Westaway and James Jackson. Surface faulting in the southern italian campania-basilicata earthquake of 23 november 1980. *Nature*, 312(5993):436–438, 1984.
- [9] Luigi Improta, Giovanni Iannaccone, Paolo Capuano, Aldo Zollo, and P Scandone. Inferences on the upper crustal structure of southern apennines (italy) from seismic refraction investigations and subsurface data. *Tectonophysics*, 317(3-4):273–298, 2000.
- [10] Etta Patacca, Paolo Scandone, et al. Post-tortonian mountain building in the apennines. the role of the passive sinking of a relic lithospheric slab. *Atti dei Convegni Lincei*, 80:157–176, 1989.
- [11] Robert WH Butler, Sveva Corrado, Stefano Mazzoli, Mauro De Donatis, Daniela Di Bucci, Giuseppe Naso, Davide Scrocca, Carlo Nicolai, and Vittorio Zucconi. Time and space variability of «thin-skinned» and «thick-skinned» thrust tectonics in the apennines (italy). *Rendiconti Lincei*, 11(1):5–39, 2000.
- [12] P Shiner, A Beccacini, and Stefano Mazzoli. Thin-skinned versus thick-skinned structural models for apulian carbonate reservoirs: constraints from the val d’agri fields, s apennines, italy. *Marine and Petroleum Geology*, 21(7):805–827, 2004.
- [13] Luigi Ferranti, Mimmo Palano, Flavio Cannavò, Maria Enrica Mazzella, John S Oldow, Erwan Gueguen, Mario Mattia, and Carmelo Monaco. Rates of geodetic deformation across active faults in southern italy. *Tectonophysics*, 621:101–122, 2014.
- [14] Paolo D’Adda, Roberto Longoni, Corrado Magistroni, Marco Meda, Fabrizio Righetti, Cristian Cavozi, Yago Nestola, and Fabrizio Storti. Extensional reactivation of a deep transpressional architecture: Insights from sandbox analogue modeling applied to the val d’agri basin (southern apennines, italy). *Interpretation*, 5(1):SD55–SD66, 2017.
- [15] L. Valoroso, D. Piccinini, L. Improta, S. Gaviano, and C. Giunchi. Characterizing seismogenic fault structures of the lake pertusillo reservoir induced seismicity (southern italy) using a relocated template-matching catalog.

- Journal of Geophysical Research: Solid Earth*, 128, e2022JB025879:488–502, 2023.
- [16] G. Cello, E. Tondi, L. Micarelli, and L. Mattioni. Active tectonics and earthquake sources in the epicentral area of the 1857 basilicata earthquake (southern italy). *Journal of Geodynamics*, 36(1):37–50, 2003. Active Faults: Analysis, Processes and Monitoring.
- [17] L Maschio, Luigi Ferranti, and P Burrato. Active extension in val d’agri area, southern apennines, italy: implications for the geometry of the seismogenic belt. *Geophysical Journal International*, 162(2):591–609, 2005.
- [18] Stefania Candela, Stefano Mazzoli, Antonella Megna, and Stefano Santini. Finite element modelling of stress field perturbations and interseismic crustal deformation in the val d’agri region, southern apennines, italy. *Tectonophysics*, 657:245–259, 2015.
- [19] Luisa Valoroso, Luigi Improta, Pasquale De Gori, and Claudio Chiarabba. Upper crustal structure, seismicity and pore pressure variations in an extensional seismic belt through 3-d and 4-d vp and vp/vs models: The example of the val d’agri area (southern italy). *Journal of Geophysical Research: Solid Earth*, 116(B7), 2011.
- [20] Alessio Lavecchia, Vincenzo Serlenga, Marilena Filippucci, Tony Alfredo Stabile, Giacomo Prosser, and Andrea Tallarico. Fault (re)activation and fluid-induced seismicity: An example from the val d’agri intermontane basin (southern italy). *Journal of Geophysical Research: Solid Earth*, 129(7):e2024JB028710, 2024. e2024JB028710 2024JB028710.
- [21] Giuseppe Cello and Stefano Mazzoli. Apennine tectonics in southern italy: a review. *Journal of Geodynamics*, 27(2):191–211, 1998.
- [22] Carlo Doglioni, Saverio Merlini, and Giuseppe Cantarella. Foredeep geometries at the front of the apennines in the ionian sea (central mediterranean). *Earth and Planetary Science Letters*, 168(3-4):243–254, 1999.
- [23] Davide Scrocca, Eugenio Carminati, and Carlo Doglioni. Deep structure of the southern apennines, italy: Thin-skinned or thick-skinned? *Tectonics*, 24(3), 2005.

- [24] Stefano Mazzoli, S Barkham, Giuseppe Cello, R Gambini, L Mattioni, P Shiner, and Emanuele Tondi. Reconstruction of continental margin architecture deformed by the contraction of the lagonegro basin, southern apennines, italy. *Journal of the Geological Society*, 158(2):309–319, 2001.
- [25] Robert Trice. Application of borehole image logs in constructing 3d static models of productive fracture networks in the apulian platform, southern apennines. *Geological Society, London, Special Publications*, 159(1):155–176, 1999.
- [26] David A Wavrek and Fausto Mosca. Compositional grading in the oil column: advances from a mass balance and quantitative molecular analysis. 2004.
- [27] Enrico Tavarnelli and Giacomo Prosser. The complete apennines orogenic cycle preserved in a transient single outcrop near san fele, lucania, southern italy. *Journal of the Geological Society*, 160(3):429–434, 2003.
- [28] Luigi Improta, Luisa Valoroso, Davide Piccinini, and Claudio Chiarabba. A detailed analysis of wastewater-induced seismicity in the val d’agri oil field (italy). *Geophysical Research Letters*, 42(8):2682–2690, 2015.
- [29] Eni spa. Val d’agri network, 2001.
- [30] T. Braun, S. Danesi, and A. Morelli. Application of monitoring guidelines to induced seismicity in italy. *Journal of Seismology*, pages 1–14, 2020.
- [31] L. Valoroso, L. Improta, L. Chiaraluce, R. Di Stefano, L. Ferranti, A. Govoni, and C. Chiarabba. Active faults and induced seismicity in the val d’agri area (southern apennines, italy). *Geophysical Journal International*, 178(1):488–502, 2009.
- [32] T.A. Stabile, A. Giocoli, V. Lapenna, A. Perrone, S. Piscitelli, and L. Telesca. Evidence of low-magnitude continued reservoir-induced seismicity associated with the pertusillo artificial lake (southern italy). *Bulletin of the Seismological Society of America*, 104(4):1820–1828, 2014.
- [33] P. Talwani. On the nature of reservoir induced seismicity. *Pure and Applied Geophysics*, 150:473–492, 1997.

- [34] Pierfrancesco Burrato and Gianluca Valensise. Rise and fall of a hypothesized seismic gap: Source complexity in the mw 7.0 16 december 1857 southern italy earthquake. *Bulletin of the Seismological Society of America*, 98(1):139–148, 02 2008.
- [35] Simone Bello, Giusy Lavecchia, Carlo Andrenacci, Maurizio Ercoli, Daniele Cirillo, Filippo Carboni, Massimiliano R. Barchi, and Francesco Brozzetti. Complex trans-ridge normal faults controlling large earthquakes. *Scientific Reports*, 12(10676), 2022.
- [36] Luigi Cucci, Silvia Pondrelli, Alberto Frepoli, Maria Teresa Mariucci, and Marco Moro. Local pattern of stress field and seismogenic sources in the pergola–melandro basin and the agri valley (southern italy). *Geophysical Journal International*, 156(3):575–583, 03 2004.
- [37] Alberto Frepoli, Francesca Romana Cinti, Laura Amicucci, Giovanni Battista Cimini, Pasquale De Gori, and Simona Pierdominici. Pattern of seismicity in the lucanian apennines and foredeep (southern italy) from recording by saptex temporary array. *Annals of Geophysics*, 48(6), 2005.
- [38] L. Improta, S. Bagh, P. De Gori, L. Valoroso, M. Pastori, D. Piccinini, C. Chiarabba, M. Anselmi, and M. Buttinelli. Reservoir structure and wastewater-induced seismicity at the val d’agri oilfield (italy) shown by three-dimensional v_p and v_p/v_s local earthquake tomography. *J. Geophys. Res.*, 122(11):9050–9082, 2017.
- [39] T.A. Stabile, V Serlenga, C Satriano, M Romanelli, E Gueguen, M R Gallipoli, E Ripepi, J-M Saurel, S Panebianco, J Bellanova, and E. Enrico Priolo. The insieme seismic network: a research infrastructure for studying induced seismicity in the high agri valley (southern italy). *Earth System Science Data*, 12(1):519–538, 2020.
- [40] Daniel E McNamara and Raymond P Buland. Ambient noise levels in the continental united states. *Bulletin of the seismological society of America*, 94(4):1517–1527, 2004.
- [41] R.V. Allen. Automatic earthquake recognition and timing from single traces. *Bulletin of the Seismological Society of America*, 68(5):1521–1532, 1978.

- [42] Zachary E Ross, Men-Andrin Meier, and Egill Hauksson. P wave arrival picking and first-motion polarity determination with deep learning. *Journal of Geophysical Research: Solid Earth*, 123(6):5120–5129, 2018.
- [43] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [44] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.
- [45] Florian Marquardt. Machine learning and quantum devices. *SciPost Physics Lecture Notes*, page 029, 2021.
- [46] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010: 19th International Conference on Computational Statistics Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.
- [47] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [48] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.
- [49] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pages 253–256. IEEE, 2010.
- [50] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

- [52] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.
- [53] Manfred Baer and Urs Kradolfer. An automatic phase picker for local and teleseismic events. *Bulletin of the Seismological Society of America*, 77(4):1437–1445, 1987.
- [54] Bernd Weber, Jan Becker, Winfried Hanka, Anders Heinloo, M Hoffmann, Toni Kraft, Doreen Pahlke, Jörg Reinhardt, Joachim Saul, H Thoms, et al. Seiscomp3—automatic and interactive real time data processing. In *Geophys. Res. Abstr*, volume 9, page 219, 2007.
- [55] S. Gentili and A. Michelini. Automatic picking of p and s phases using a neural tree. *J. Seismol.*, 10:39–63, 2006.
- [56] Dmitry A Storchak, Domenico Di Giacomo, ER Engdahl, James Harris, István Bondár, William HK Lee, P Bormann, and Antonio Villaseñor. The isc-gem global instrumental earthquake catalogue (1900–2009): introduction. *Physics of the Earth and Planetary Interiors*, 239:48–63, 2015.
- [57] Amadej Trnkoczy, Peter Bormann, Winfried Hanka, L Gary Holcomb, Robert L Nigbor, Masanao Shinohara, Hajime Shiobara, and Kiyoshi Suyehiro. Site selection, preparation and installation of seismic stations. In *New manual of seismological observatory practice 2 (NMSOP-2)*, pages 1–139. Deutsches GeoForschungsZentrum GFZ, 2012.
- [58] Men-Andrin Meier, Zachary E Ross, Anshul Ramachandran, Ashwin Balakrishna, Suraj Nair, Peter Kundzicz, Zefeng Li, Jennifer Andrews, Egill Hauksson, and Yisong Yue. Reliable real-time seismic signal/noise discrimination with machine learning. *Journal of Geophysical Research: Solid Earth*, 124(1):788–800, 2019.
- [59] Dirk Gajewski and Ekkehart Tessmer. Reverse modelling for seismic event characterization. *Geophysical Journal International*, 163(1):276–284, 2005.
- [60] Honn Kao and Shao-Ju Shan. The source-scanning algorithm: Mapping the distribution of seismic sources in time and space. *Geophysical Journal International*, 157(2):589–594, 2004.

- [61] Francesco Grigoli, Simone Cesca, Maurizio Vassallo, and Torsten Dahm. Automated seismic event location by travel-time stacking: An application to mining induced seismicity. *Seismological Research Letters*, 84(4):666–677, 2013.
- [62] Francesco Grigoli, Simone Cesca, Lars Krieger, Marius Kriegerowski, Sergio Gammaldi, Josef Horalek, Enrico Priolo, and Torsten Dahm. Automated microseismic event location using master-event waveform stacking. *Scientific reports*, 6(1):25744, 2016.
- [63] Steven J Gibbons and Frode Ringdal. The detection of low magnitude seismic events using array-based waveform correlation. *Geophysical Journal International*, 165(1):149–166, 2006.
- [64] Robert J Skoumal, Michael R Brudzinski, Brian S Currie, and Jonathan Levy. Optimizing multi-station earthquake template matching through re-examination of the youngstown, ohio, sequence. *Earth and Planetary Science Letters*, 405:274–280, 2014.
- [65] S.M. Mousavi, W.L. Ellsworth, W. Zhu, and et al. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nat Commun*, 11(3952), 2020.
- [66] Zachary E Ross, Yisong Yue, Men-Andrin Meier, Egill Hauksson, and Thomas H Heaton. Phaselink: A deep learning approach to seismic phase association. *Journal of Geophysical Research: Solid Earth*, 124(1):856–869, 2019.
- [67] S. Mostafa Mousavi and Gregory C. Beroza. Machine learning in earthquake seismology. *Annual Review of Earth and Planetary Sciences*, 51(Volume 51, 2023):105–129, 2023.
- [68] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

- [69] Elisa Caredda, Marius Paul Isken, Simone Cesca, Maddalena Errico, Giampaolo Zerbinato, and Andrea Morelli. Improving microearthquake detection in the val d’agri region (southern italy) with deep learning. *Seismica*, 4(2), 2025.
- [70] G. Lavecchia, S. Bello, D. Cirillo, F. Pietrolungo, and F. Brozzetti. Quaternary-host faults database 2.0 (southern italy) [data set], 2023.
- [71] M. Zhang, W.L. Ellsworth, and G.C. Beroza. Rapid earthquake association and location. *Seismological Research Letters*, 90(6):2276–2284, 2019.
- [72] A. Lomax, J. Virieux, P. Volant, and C. Berge-Thierry. *Probabilistic Earthquake Location in 3D and Layered Models.*, volume 18, pages 101–134. Springer, Dordrecht, 2000.
- [73] A. Lomax, A. Michelini, and A. Curtis. *Earthquake Location, Direct, Global-Search Methods.*, volume 18. Springer, New York, NY., 2014.
- [74] Alexander Wickham-Piotrowski, Yvonne Font, Marc Regnier, Bertrand Delouis, Olivier Lengliné, Monica Segovia, and Quentin Bletery. Achieving a comprehensive microseismicity catalog through a deep-learning-based workflow: Applications in the central ecuadorian subduction zone. *Bulletin of the Seismological Society of America*, 114(2):823–841, 12 2023.
- [75] A. Michelini, S. Cianetti, S. Gaviano, C. Giunchi, D. Jozinović, and V. Lauciani. Instance – the italian seismic dataset for machine learning. *Earth System Science Data*, 13(12):5509–5544, 2021.
- [76] S. Cianetti, A. Lomax, A. Michelini, and et al. Comparison of deep-learning versus manual seismic arrival-time picks based on quality of high-precision earthquake locations and seismic tomography., 2025.
- [77] Albert Tarantola, Bernard Valette, et al. Inverse problems= quest for information. *Journal of geophysics*, 50(1):159–170, 1982.
- [78] A Tarantola. Inversion of travel times and seismic waveforms. *Seismic Tomography: With Applications in Global Seismology and Exploration Geophysics*, pages 135–157, 1987.

- [79] TJ Moser, T Van Eck, and G Nolet. Hypocenter determination in strongly heterogeneous earth models using the shortest path method. *Journal of Geophysical Research: Solid Earth*, 97(B5):6563–6572, 1992.
- [80] Donald Joseph R Meagher. *The octree encoding method for efficient solid modeling*. Rensselaer Polytechnic Institute, 1982.
- [81] W.H. Bakun and W.B. Joyner. The ml scale in central california. *Bulletin of the Seismological Society of America*, 75(5):1827–1843, 1984.
- [82] A. Morelli, M. Anselmi, T. Braun, S. Danesi, M. Errico, D. Famiani, A. Garcia, I. Molinari, M. Vassallo, L. Zaccarelli, and G. Zerbinato. A state-of-the-art analysis system designed to monitor microseismicity at a local scale. *Annals of Geophysics*, 67(6), 2025.
- [83] M. Isken, P. Niemz, J. Münchmeyer, P. Büyükakpınar, S. Heimann, S. Cesca, H. Vasyura-Bathke, and T. Dahm. Qseek: A data-driven framework for automated earthquake detection, localization and characterization. *Seismica*, 4(1), 2025.
- [84] Donald Meagher. Geometric modeling using octree encoding. *Computer graphics and image processing*, 19(2):129–147, 1982.
- [85] Anthony Lomax and Alexandros Savvaidis. High-precision earthquake location using source-specific station terms and inter-event waveform similarity. *Journal of Geophysical Research: Solid Earth*, 127(1):e2021JB023190, 2022. e2021JB023190 2021JB023190.
- [86] K. B. Richards-Dinger and P. M. Shearer. Earthquake locations in southern California obtained using source-specific station terms. *Journal of Geophysical Research: Solid Earth*, 105(B5):10939–10960, May 2000.
- [87] Francesco Grigoli, Simone Cesca, Lars Krieger, Marius Kriegerowski, Sergio Gammaldi, Josef Horalek, Enrico Priolo, and Torsten Dahm. Automated microseismic event location using master-event waveform stacking. *Scientific reports*, 6(1):25744, 2016.
- [88] G. Adinolfi, S. Cesca, M. Picozzi, S. Heimann, and A. Zollo. Detection of weak seismic sequences based on arrival time coherence and empiric network

- detectability: an application at a near fault observatory. *Geophysical Journal International*, 218(3):2054–2065, 07 2019.
- [89] D. E. McNamara and R. I. Boaz. Seismic noise analysis system using power spectral density probability density functions: A stand-alone software package. *USGS Open-File Report*, 1438, 2005.
- [90] H.K Gupta, B. K. Rastogi, and H. Narain. Common features of the reservoir-associated seismic activities. *Bulletin of the Seismological Society of America*, 62(2):481–492, 1972.
- [91] Andrea Berbellini, Lucia Zaccarelli, Licia Faenza, Alexander Garcia, Luigi Improta, Pasquale De Gori, and Andrea Morelli. Effect of groundwater on noise-based monitoring of crustal velocity changes near a produced water injection well in val d’agri (italy). *Frontiers in Earth Science*, 9, 2021.
- [92] Luciano Telesca, Serena Panebianco, Vincenzo Serlenga, and Tony Alfredo Stabile. Fractal, spectral, and topological analysis of the reservoir-induced seismicity of pertusillo area (southern italy). *Fractal and Fractional*, 9(4), 2025.
- [93] S.M. Mousavi and G.C. Beroza. Deep-learning seismology. *Science*, 377(1), 2022.
- [94] S. M. Mousavi and G. C. Beroza. Machine learning in earthquake seismology. *Annual Review of Earth and Planetary Sciences*, 51(1):105–129, 2023.
- [95] Beyreuther, R. M., Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann. Obspy: A python toolbox for seismology. *Seismological Research Letters*, 81(3):530–533, 2010.
- [96] S. Heimann, M. Kriegerowski, Isken, S. M., Cesca, S. Daout, F. Grigoli, C. Juretzek, Megies, Nooshiri T., A. N., Steinberg, H. Sudhaus, H. Vasyura-Bathke, T. Willey, and T. Dahm. Pyrocko - an open-source seismology toolbox and library, 2017.
- [97] J. D. Hunter. "matplotlib: A 2d graphics environment". *Computing in Science & Engineering*, 9(3):90–95, 2007.

- [98] Elisa Caredda, Marius Paul Isken, Simone Cesca, Maddalena Errico, Giampaolo Zerbinato, and Andrea Morelli. Earthquake catalogs for: Improving detection of micro-earthquakes in the val d’agri region (southern italy) using deep learning algorithms, September 2025.
- [99] Beno Gutenberg and Charles F Richter. Frequency of earthquakes in california. *Bulletin of the Seismological society of America*, 34(4):185–188, 1944.
- [100] Jochen Woessner and Stefan Wiemer. Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty. *Bulletin of the Seismological Society of America*, 95(2):684–698, 2005.
- [101] Stefan Wiemer and Max Wyss. Minimum magnitude of completeness in earthquake catalogs: Examples from alaska, the western united states, and japan. *Bulletin of the seismological society of America*, 90(4):859–869, 2000.
- [102] K. Aki. Maximum likelihood estimate of b in the formula $\log n = a - bm$ and its confidence limits. *Bull. Earthq. Res. Inst.*, (43):237–239, 1965.
- [103] Richard J Rossi. *Mathematical statistics: an introduction to likelihood based inference*. John Wiley & Sons, 2018.
- [104] Yosihiko Ogata and Koichi Katsura. Analysis of temporal and spatial heterogeneity of magnitude frequency distribution inferred from earthquake catalogues. *Geophysical Journal International*, 113(3):727–738, 1993.
- [105] Yosihiko Ogata and Koichi Katsura. Analysis of temporal and spatial heterogeneity of magnitude frequency distribution inferred from earthquake catalogues. *Geophysical Journal International*, 113(3):727–738, 1993.
- [106] Robert J Geller and Charles S Mueller. Four similar earthquakes in central california. *Geophysical Research Letters*, 7(10):821–824, 1980.
- [107] Hansruedi Maurer and Nicholas Deichmann. Microearthquake cluster detection based on waveform similarities, with an application to the western swiss alps. *Geophysical Journal International*, 123(2):588–600, 1995.

- [108] Masaru Tsujiura. Characteristic frequencies for earthquake families and their tectonic implications: evidence from earthquake swarms in the kanto district, japan. *pure and applied geophysics*, 121(4):573–600, 1983.
- [109] Peter M Shearer, Jeanne L Hardebeck, Luciana Astiz, and Keith B Richards-Dinger. Analysis of similar event clusters in aftershocks of the 1994 northridge, california, earthquake. *Journal of Geophysical Research: Solid Earth*, 108(B1), 2003.
- [110] Toshihiro Igarashi, Toru Matsuzawa, and Akira Hasegawa. Repeating earthquakes and interplate aseismic slip in the northeastern japan subduction zone. *Journal of Geophysical Research: Solid Earth*, 108(B5), 2003.
- [111] Yosio Nakamura. A1 moonquakes-source distribution and mechanism. In *In: Lunar and Planetary Science Conference, 9th, Houston, Tex., March 13-17, 1978, Proceedings. Volume 3.(A79-39253 16-91) New York, Pergamon Press, Inc., 1978, p. 3589-3607.*, volume 9, pages 3589–3607, 1978.
- [112] William Menke. Using waveform similarity to constrain earthquake locations. *Bulletin of the Seismological Society of America*, 89(4):1143–1146, 1999.
- [113] Felix Waldhauser and William L Ellsworth. A double-difference earthquake location algorithm: Method and application to the northern hayward fault, california. *Bulletin of the seismological society of America*, 90(6):1353–1368, 2000.
- [114] Peter Shearer, Egill Hauksson, and Guoqing Lin. Southern california hypocenter relocation with waveform cross-correlation, part 2: Results using source-specific station terms and cluster analysis. *Bulletin of the Seismological Society of America*, 95(3):904–915, 2005.
- [115] Václav Vavryčuk. Moment tensor decompositions revisited. *Journal of Seismology*, 19(1):231–252, 2015.
- [116] ML u Jost and RB Herrmann. A student’s guide to and review of moment tensors. *Seismological Research Letters*, 60(2):37–57, 1989.
- [117] Bruce R Julian, Angus D Miller, and GR Foulger. Non-double-couple earthquakes 1. theory. *Reviews of Geophysics*, 36(4):525–549, 1998.

- [118] Angus D Miller, GR Foulger, and Bruce R Julian. Non-double-couple earthquakes 2. observations. *Reviews of Geophysics*, 36(4):551–568, 1998.
- [119] G M Petersen, P Niemz, S Cesca, V Mouslopoulou, and G M Bocchini. Clusty, the waveform-based network similarity clustering toolbox: concept and application to image complex faulting offshore zakynthos (greece). *Geophysical Journal International*, 224(3):2044–2059, 11 2020.
- [120] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.
- [121] Christopher C Paige and Michael A Saunders. Lsq: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71, 1982.
- [122] Sebastian Heimann, Marius Isken, Daniela Kühn, Henriette Sudhaus, Andreas Steinberg, Simon Daout, Simone Cesca, H Bathke, and Torsten Dahm. Grond: A probabilistic earthquake source inversion framework. 2018.
- [123] Donald B Rubin. The bayesian bootstrap. *The annals of statistics*, pages 130–134, 1981.
- [124] M Parotidis, E Rothert, and SA Shapiro. Pore-pressure diffusion: A possible triggering mechanism for the earthquake swarms 2000 in vogtland/nw-bohemia, central europe. *Geophysical Research Letters*, 30(20), 2003.
- [125] Stephen A Miller, Cristiano Collettini, Lauro Chiaraluce, Massimo Cocco, Massimiliano Barchi, and Boris JP Kaus. Aftershocks driven by a high-pressure co2 source at depth. *Nature*, 427(6976):724–727, 2004.
- [126] Sebastian Hainzl and Yosihiko Ogata. Detecting fluid signals in seismicity data through statistical earthquake modeling. *Journal of Geophysical Research: Solid Earth*, 110(B5), 2005.
- [127] XL Lei, S-l Ma, XZ Wen, J-r Su, and F Du. Integrated analysis of stress and regional seismicity by surface loading—a case study of zipingpu reservoir. *Seismology and Geology*, 30(4):1046–1064, 2008.

- [128] Kai Deng, Shiyong Zhou, Rui Wang, Russell Robinson, Cuiping Zhao, and Wanzheng Cheng. Evidence that the 2008 m w 7.9 wenchuan earthquake could not have been induced by the zipingpu reservoir. *Bulletin of the Seismological Society of America*, 100(5B):2805–2814, 2010.
- [129] Rongjiang Wang. A simple orthonormalization method for stable and efficient computation of green’s functions. *Bulletin of the Seismological Society of America*, 89(3):733–741, 1999.
- [130] Donald Wells and Kevin Coppersmith. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin of the seismological Society of America*, 84(4):974–1002, 1994.
- [131] Robert W Graves and Arben Pitarka. Broadband ground-motion simulation using a hybrid approach. *Bulletin of the Seismological Society of America*, 100(5A):2095–2123, 2010.
- [132] KB Broberg. How fast can a crack go? *Materials Science*, 32(1):80–86, 1996.
- [133] David J Wald, Vincent Quitoriano, Thomas H Heaton, and Hiroo Kanamori. Relationships between peak ground acceleration, peak ground velocity, and modified mercalli intensity in california. *Earthquake spectra*, 15(3):557–564, 1999.
- [134] Carlo Meletti, Romano Camassi, and Viviana Castelli. A reappraisal of the seismicity of sardinia, italy. *Seismological Society of America*, 92(2A):1148–1158, 2021.
- [135] Andrea Rovida, Mario Locati, Romano Camassi, Barbara Lolli, Paolo Gasperini, Andrea Antonucci, et al. Catalogo parametrico dei terremoti italiani (cpti15). *INGV versione 4.0*, 2022.
- [136] Fabrizio Magrini, Giovanni Diaferia, Islam Fadel, Fabio Cammarano, Mark Van Der Meijde, and Lapo Boschi. 3-d shear wave velocity model of the lithosphere below the sardinia–corsica continental block based on rayleigh-wave phase velocities. *Geophysical Journal International*, 220(3):2119–2130, 2020.

- [137] Fabio Cammarano, Henrique Berger Roisenberg, Alessio Conclave, Islam Fadel, and Mark van der Meijde. On the crustal composition of the sardinia-corsica continental block inferred from receiver functions. *Solid Earth*, 16(2):135–154, 2025.
- [138] Andrea Rovida, Mario Locati, Romano Camassi, Barbara Lolli, and Paolo Gasperini. The italian earthquake catalogue cpti15. *Bulletin of Earthquake Engineering*, 18(7):2953–2984, 2020.
- [139] Ruijia Wang, Brandon Schmandt, Monique Holt, and Keith Koper. Advancing local distance discrimination of explosions and earthquakes with joint p/s and ml-mc classification. *Geophysical Research Letters*, 48(23):e2021GL095721, 2021.
- [140] James E Quick, Silvano Sinigoi, Luisa Negrini, Gabriella Demarchi, and Adriano Mayer. Synmagmatic deformation in the underplated igneous complex of the ivrea-verbano zone. *Geology*, 20(7):613–616, 1992.
- [141] SM Schmid, Andre Zingg, and M Handy. The kinematics of movements along the insubric line and the emplacement of the ivrea zone. *Tectonophysics*, 135(1-3):47–66, 1987.
- [142] Mattia Pistone, Luca Ziberna, György Hetényi, Matteo Scarponi, Alberto Zanetti, and Othmar Müntener. Joint geophysical-petrological modeling on the ivrea geophysical body beneath valsesia, italy: Constraints on the continental lower crust. *Geochemistry, Geophysics, Geosystems*, 21(12):e2020GC009397, 2020.
- [143] Judith M Confal, Simone Salimbeni, Adriano Cavaliere, Stefania Danesi, Irene Molinari, Maddalena Errico, Ludovic Baron, Steffen Uhlmann, György Hetényi, and Silvia Pondrelli. Divenet: a local seismographic network monitoring the lower continental crust drilling activities for the icdp-dive project. *Annals of Geophysics*, 68(5):DM579–DM579, 2025.
- [144] Jon R Peterson. Observations and modeling of seismic background noise. Technical report, US Geological Survey, 1993.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Prof. Andrea Morelli, for his kind and valuable guidance and support throughout the course of this PhD, and for his constant availability and thoughtful advice. His scientific rigor has greatly influenced my way of thinking as a researcher.

I am sincerely thankful to Simone Cesca and Marius P. Isken for their collaboration and mentorship during my research stay at the GFZ Helmholtz Centre for Geosciences in Potsdam. Working with them was an enriching experience that significantly contributed to the development of this work and broadened my scientific perspective. Their insight, patience, and enthusiasm have been a constant source of inspiration.

I express my sincere gratitude to the CMS analysts and board for providing the data and facilitating this research. In particular, I would like to thank Maddalena Errico for her valuable insights and discussions throughout various stages of this work, and Paolo Zerbinato for his technical support.

Special thanks go to Spina Cianetti for her initial guidance on software usage. I also extend my appreciation to all network operators for maintaining the seismic instrumentation utilized in this study, particularly the commercial operators ENI and Total E&P for their continued interest in our research, support, and collaboration.

I would like to thank Henrique Berger Roisenberg and Prof. Fabio Cammarano for involving me in the study of the Sardinia dataset, and Silvia Pondrelli and Simone Salimbeni for the collaboration on the DIVE network data. I hope that these collaborations will continue and further develop in the future.