Investigating Seismicity in Cape Town: Implications for Active Fault Lines in the Western Cape, South Africa

by Wade van Zyl (VZYWAD001)

A thesis project as part of the requirements of the MSc degree in Geological Sciences

Supervisor: Dr. Diego Quiros

Co-supervisor: Dr. Alastair Sloan

Declaration

I know that plagiarism is wrong. Plagiarism is using another's work and pretending that it is one's own. I have used the South African Journal of Geology (SAJG) convention for citation and referencing. Each contribution to, and quotation in, this project from the work (s) of other people has been attributed and has been cited and referenced. This project is my work. I have not allowed and will not allow anyone to copy my work to pass it off as his or her work.

Signature: Signed by candidate

Date: 2024/10/23

Abstract

Despite being in a stable continental region (SCR), South Africa has experienced significant seismic activity. Historical records cite a possible 6.5 magnitude earthquake in Cape Town in 1809. On September 29, 1969, a 6.3 magnitude earthquake struck the Ceres-Tulbagh region, less than 100 km from the Koeberg Nuclear Power Station (KNPS) in Cape Town. Previous studies have found a relationship between enhanced micro-seismicity over long periods and source zones of historical SCR earthquakes. This thesis seeks to identify heightened micro-seismic activity on regional fault structures to infer potential source zones for the 1809 event and future damaging earthquakes. To achieve this, eighteen three-component geophones were deployed across a 40 by 35-kilometre area near the KNPS. The geophones recorded data from August to October 2021 and were located near the Ceres-Tulbagh region, Cape Town, the proposed Milnerton fault, and the Colenso fault zone. Seismicity around these fault zones was analyzed using machine learning, visual inspection, and Short-Time Average to Long-time Average (STA/LTA) algorithms. Thirty-five events were found, categorized into two groups of elevated seismicity: one group was located offshore, outside the study area, while the other was situated between the proposed Milnerton fault and the Colenso fault system. Within the second group, the Colenso fault system shows elevated micro-seismicity, indicating that it is potentially active. Additional findings suggest that machine learning and visual examination of waveform data are more accurate than STA/LTA algorithms combined with manual assessment at detecting micro-seismic phases and consequently events.

Acknowledgments

I would like to express my deepest gratitude to my supervisory team, Dr. Diego Quiros and Dr. Alastair Sloan, for their invaluable guidance, advice, and support throughout this project.

Secondly, I extend my heartfelt thanks to my academic mentor, Prof. Jacob Jaftha, for his wisdom and guidance in helping me navigate the postgraduate journey.

A special thank you goes to Janine Carlse, Nuraan Kafaar, Chantal Swartz, Nathalie Barends, and Elretha Roos for their patience and assistance with registration processing and all related administrative tasks.

I am also grateful to the Accelerated Transformation of the Academic Programme (ATAP) scholarship, the Building Research Active Academic Staff (BRAAS) scholarship, and the Postgraduate Funding Office at the University of Cape Town for their financial support.

To my friends and family, thank you for your unwavering support.

Lastly, I would like to acknowledge and thank Schlumberger for providing the Vista 3-D seismic analysis software, as well as the teams behind Obspy, Generic Mapping Tools, SEISAN, and Seisbench.

Contents

List of Figures

List of Tables

1. Introduction

South Africa is an example of a Stable Continental Region (SCR) because of its placement far from recognized plate boundaries and its low observed strain rates of 1 nanostrain yr^{-1} (Brandt, 2011; Malservisi et al., 2013). Simplistic views of plate tectonics suggest that plate interiors should be rigid and undeforming with earthquakes only occurring near plate boundaries. However, it has long been recognized that large earthquakes do occur in plate interiors (Calais et al., 2016; Martín-González et al., 2023), but assessing seismic hazards in these areas is challenging due to infrequent seismicity, low strain rates and fault ruptures that are geologically difficult to locate (Brandt, 2011; Malservisi et al., 2013; Calais et al., 2016; Manzunzu et al., 2019; Martín-González et al., 2023).

Southern Africa consists of several Archean cratons, the closest one for this study is the Kaapvaal Craton, consisting of 3.2 Ga gneisses and narrow greenstone belts shown in Figure 1A (Begg et al., 2009). North of the Kaapvaal Craton lies the Limpopo Belt, formed around 2.6-2.7 Ga through the collision of the Kaapvaal Craton and the Zimbabwe Craton, resulting in the Kalahari Craton (Begg et al., 2009; Bumby et al., 2004). The southern boundary of the Kaapvaal Craton is marked by the Namaqua-Natal belt, a granulite terrane believed to have formed during the assembly of the supercontinent Rodinia around 1.0 to 1.2 Ga (Raith et al., 2003). The northern boundary between the Kaapvaal Craton and the Namaqua-Natal belt is referred to as the Kheis belt (Moen, 1999; Van Niekerk, 2009).

The two most significant tectonic provinces relevant to this study are the Saldania Belt and the Cape Fold Belt, both located South and West of the Kaapvaal craton. According to Kisters and Belcher, (2018) and Kisters et al. (2002), the Adamastor Ocean, which once connected South America, and the Kalahari Craton was subducted beneath the Kalahari craton resulting in the low-grade orogenic Saldania belt indicated in Figure 1A during the assembly of Gondwana (African Orogeny). The lithologies associated with the Saldania Belt are the polyphase deformed lower greenschist facies Malmesbury group (MG) (Kisters and Belcher, 2018; Kisters et al., 2002; Rozendaal et al., 1999). The Cape Granite Suite intruded the Saldania belt around 550-510 Ma (Kisters and Belcher, 2018; Kisters et al., 2002). These intrusions also intruded the northwest-to-southeast trending fault zone that is known as the Colenso Fault shown in Figure 2B and Figure 2C (Kisters et al., 2002). The Colenso Fault runs NW-SE from Stellenbosch to Saldanha Bay, with a surface-inferred trace length of roughly 150 km and a width of 7 km (Schoch, 1975; Theron et al., 1992; Gresse et al., 2006). Outcrop patterns associated with this fault are poorly constrained, except for some mylonitized and brecciated rocks along its strike indicating a history of ductile deformation overprinted by a period of brittle deformation (Schoch, 1975; Theron et al., 1992; Kisters et al., 2002). The fault and the Saldania Belt have a common syntectonic history; at roughly 540 Ma, the fault changed from being sinistral strike-slip to being dextral strike-slip, corresponding with the uplift of the Saldania Belt and the transition from a compressional to an extensional tectonic regime (Kisters et al., 2002).

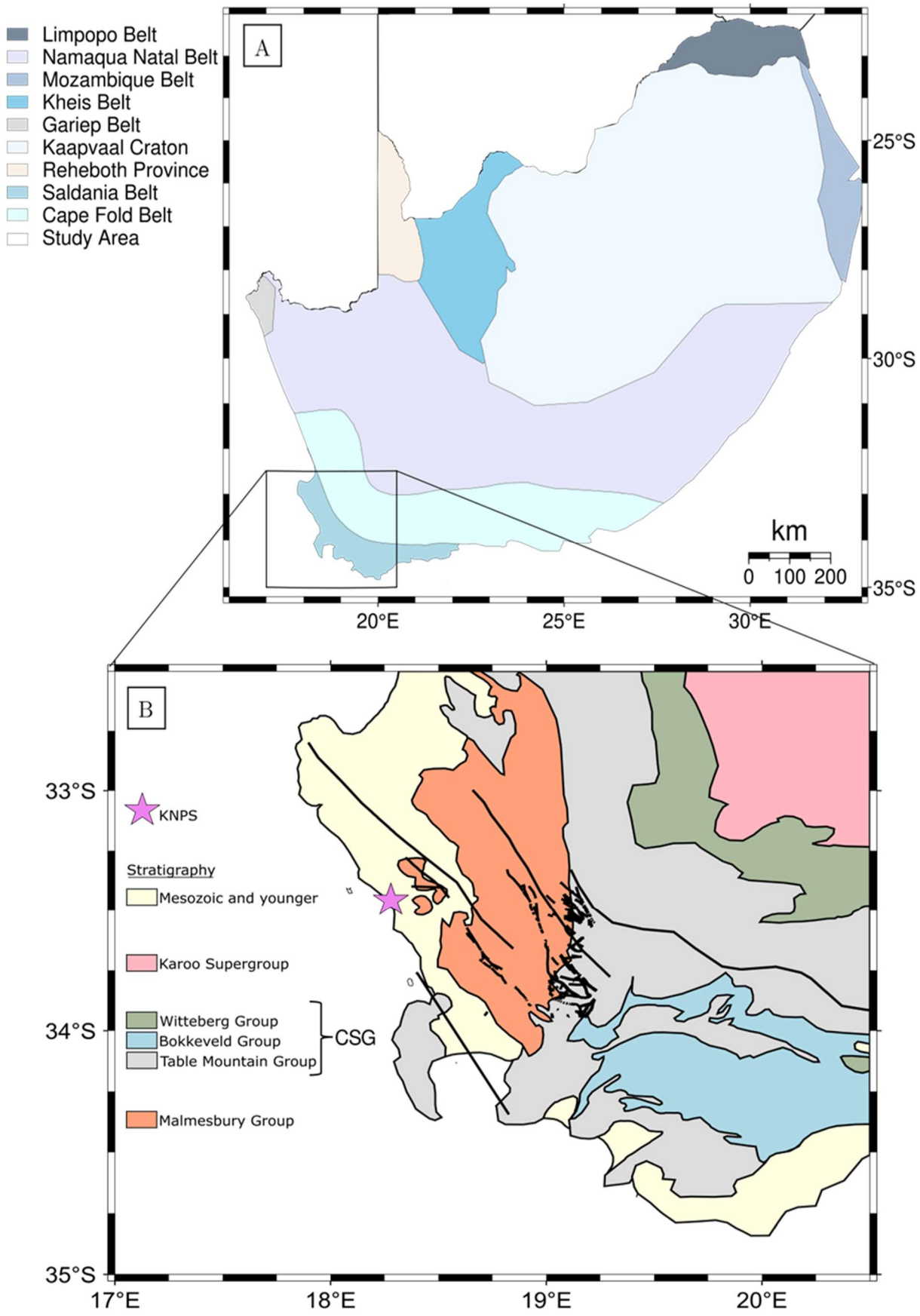


Figure 1: Figure 1A (adapted from Viola et al., 2012) shows the study area and all tectonic provinces of South Africa. The geology and stratigraphy of the research region are shown in Figure 1B, which was adapted from Flint et al. (2011), along with the location of the KNPS.

Overlying the Cape Granite Suite and Malmesbury group lies the approximately 6 to 10 km thick Cape Supergroup (CSG), as depicted in Figure 1B (Blewett et al., 2019). The Table Mountain Group (TMG) is at the base of the CSG, extending unconformably across the Malmesbury group beneath it (Blewett et al., 2019). The CSG is composed of three stratigraphic groups. The Bokkeveld group's cyclic shale and sandstone conformably cover the TMG, which is made up of basal conglomerates (Blewett et al., 2019; Thamm and Johnson., 2006). The top unit of the CSG is composed of mudrock deposits of the Witteberg Group (WG) (Blewett et al., 2019). On top of the WG is an unconformity between the CSG and Karoo Supergroup (KSG) due to a break in sedimentation (Blewett et al., 2019; Thamm and Johnson, 2006).

The CSG and MG were deformed during the Permian Orogeny, forming the 1300 km long, low-grade metamorphic Cape Fold Belt (CFB) shown in Figure 1A, due to subduction on Gondwana's southern margin (Smit et al., 2015; Lock, 1980). It is believed that the CFB formed both along pre-existing structures mostly produced by the African orogeny, as well as along structures that formed during the Permian Orogeny (Paton et al., 2006). The CFB is structurally divided into two main branches namely the NNW-trending western branch and the E-W axial orientated folded southern branch (Blewett et al., 2019; Smith et al., 2015). These two branches meet at what is referred to as the syntaxis (Johnston, 2000; Smith et al., 2015). The difference in orientations is inferred to be a product of the orientation of the forces that led to the observed structures, namely the western branch experienced strike-slip deformation whereas the southern branch experienced head-on north-directed compression (Johnston, 2000; Smith et al., 2015).

Although Southern Africa is considered a SCR, multiple seismic clusters have been identified within it. Figure 2A illustrates the recorded earthquakes of Southern Africa, along with their magnitudes, between 1620 and 2022, based on historical records and instrumental recordings. Figure 2A also indicates a number of seismic clusters that have been highlighted in the literature (Manzunzu et al. 2019; Brandt 2011) and are described in the following paragraphs.

The Witwatersrand Basin cluster in northeastern South Africa experiences seismicity related to both past and current mining activities (Brandt, 2011; Manzunzu et al., 2019). On August 5, 2014, the Orkney earthquake (M_L 5.5) struck, causing severe damage to surrounding homes and one fatality (Manzunzu et al., 2019; Miyamoto et al., 2022; Nkosi et al., 2022). This event is shown by a red star in Figure 2A. The earthquake's depth (4.7±1.2 km) and strike-slip focal mechanism suggest a tectonic origin (Manzunzu et al., 2019; Miyamoto et al., 2022; Nkosi et al., 2022).

The second cluster is the Ceres cluster, which is located northeast of Cape Town. The most significant earthquake struck this region on September 29, 1969 (Brandt, 2011; Kijko et al., 2021; Manzunzu et al., 2019; Smit et al., 2015). This 6.3-magnitude earthquake (green star in Figure 2A) occurred on a sinistral strike-slip fault, claiming 12 lives (Brandt, 2011; Kijko et al., 2021; Manzunzu et al., 2019; Smit et al., 2015). This earthquake damaged 70% of the buildings in the Ceres-Tulbagh region, displacing more than half of the population (Kijko et al., 2021).

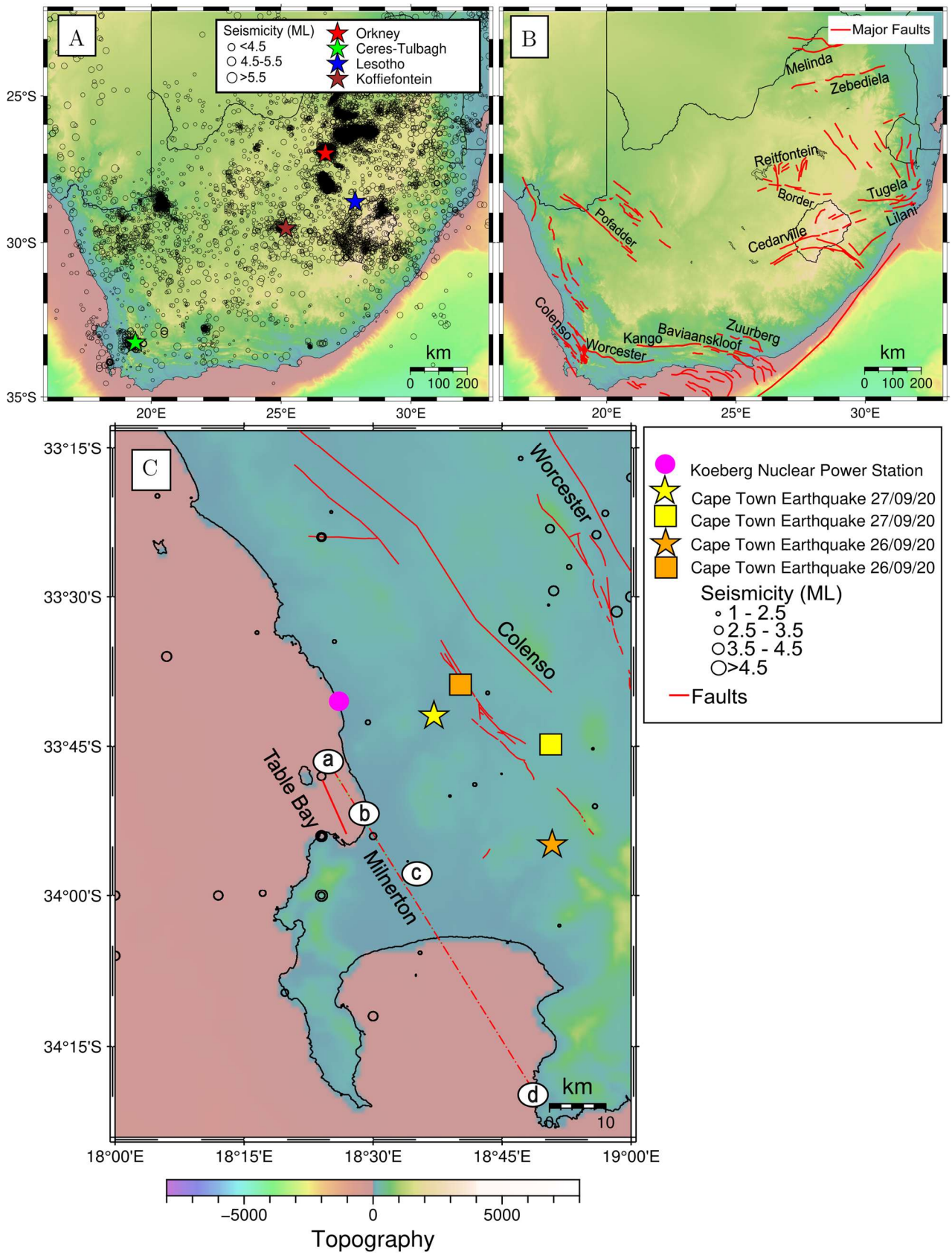


Figure 2: Figure 2A illustrates South Africa's overall seismicity as well as all major earthquakes from each seismic cluster (CGS, 2022). Figure 2B depicts all major fault systems in South Africa (Manzunzu et al., 2019), whereas Figure 2C depicts Cape Town's seismicity (CGS, 2022), and nearby fault systems such as the Colenso Fault (CGS, 2022), Milnerton Fault in dashed red (Hartnady, 2003), and Table Bay Fault (MacHutchon et al., 2020). The stars indicate the events recorded by CGS in 2020 and the squares indicate the events recorded by USGS.

The surface expression of the causative fault remains unidentified (Manzunzu et al., 2019; Smit et al., 2015), though seismic activity appears to align with the western termination of the Worcester-Kango- Baviaanskloof fault systems, provided in Figure 2B (Manzunzu et al., 2019).

The Koffiefontein cluster is cited as the most tectonically active region currently, with a 6.2 M_W earthquake on February 20, 1912, depicted by the brown star in Figure 2A (Bommer et al., 2015; Manzunzu et al., 2019; Strasser et al., 2015). This earthquake caused farmhouses to collapse and structural damage to townhouses, prompting the government to provide tents to residents of Koffiefontein and surrounding areas (Strasser et al., 2015). The Lesotho cluster's most significant occurrence is a 5.8 M_b earthquake that occurred on July 1, 1976, denoted with a blue star in Figure 2A (Manzunzu et al., 2019).

1.1. Seismicity in the vicinity of Cape Town

Earthquakes in the Cape Town region have been documented since 1602, with instrumental recordings starting in 1899 (CGS, 2022; Manzunzu et al., 2019). Two probable magnitudes ~6 earthquakes occurred before 1899, one on December 4, 1809, and another on June 2, 1811 (CGS, 2022; Hartnady, 2003).

The 1809 earthquake destroyed a farmhouse on Jan Biesjes Kraal farm near modern-day Milnerton and caused 1-inch-wide fractures extending for at least 1 mile (Von Buchenroder, 1830; Hartnady, 2003). Residents reported water spouting six feet high from numerous small craters on December 4th, indicating a vibration-induced liquefaction phenomenon (Von Buchenroder, 1830; Hartnady, 2003). This evidence suggests the epicenter may have been near Jan Biesjes Kraal (Hartnady, 2003), although it is possible this area may have been particularly susceptible to liquefaction from an earthquake within the region. The maximum Modified Mercalli (MM) intensity (I_{max}) for the Milnerton area was between VIII and IX (Hartnady, 2003).

Burchell (1822) descriptions likely indicate a time difference between the P and S phases of 3 to 4 seconds and an I_{max} upper limit of VII in the Milnerton area for the 1811 event (Hartnady, 2003). It is important to accept these values with caution because these descriptions are based on historical reports that may be inaccurate. Based on Burchell's location, this P-S time difference, along with an assumed focal depth, suggests a maximum epicentral distance of 21-32 km (Hartnady, 2003). The area around Milnerton and the historical site of Jan Biesjes Kraal, about 12 km northeast of Burchell's 1811 residence, aligns with a 3-4 second P-S interval, supporting it as near the epicenter for the 1809 event (Hartnady, 2003).

Recent seismic activity in the Cape Town cluster includes two earthquakes that occurred in 2020 (CGS, 2022; USGS, 2022). The first occurred on September 26, 2020, with a reported magnitude of 2.9, and the second occurred on September 27, 2020, with a reported magnitude of 2.7 and a projected hypocentral depth of 5 km ± 2 km for both occurrences (USGS, 2022). The epicenters shown by squares in Figure 2C are from the United States Geological Survey (USGS) database,

while the epicenters marked by stars are from the Council for Geoscience (CGS). The event on September 26th, 2020, is coloured orange, while the event on September 27th is coloured yellow. Before these two incidents, the Cape Town region was hit by a magnitude 6.3 event of Ceres-Tulbagh of 1969, followed by a magnitude 5.7 earthquake in 1970 with a very similar epicenter (CGS, 2022).

Stein and Liu (2009) found that large earthquakes in slowly deforming areas appear to have very extended aftershock sequences. They suggested that many earthquakes observed in SCRs may be the aftershocks of major events from hundreds of years ago. This suggests a potential method for identifying the causative faults of historical SCR events. Identifying areas of increased seismicity and comparing them to surrounding fault systems (Figure 2C), may aid in determining where the 1809 and 1811 events likely occurred along with the fault systems responsible for each.

While the Milnerton Fault is speculative, ongoing research is exploring potential evidence for its existence. Figure 2C contains labels a to d, all which form arguments by Dames and Moore (1976, 1981) and Hartnady (2003) for its existence. Label (a) in Figure 2C represents the offshore NNW trending bathymetric anomaly, (b) represents outcrops with reported evidence of brittle deformation near Bloubergstrand, (c) references the geomorphology of the Cape Flats being similar to that of a large shear zone, and (d) an inferred fault zone based on Landsat imagery (Dames and Moore, 1976, 1981; Hartnady, 2003; Stamatakos et al., 2024). The Milnerton Fault is the straight line that forms when connecting all these points (Dames and Moore, 1976, 1981; Hartnady, 2003; Stamatakos et al., 2024).

The Koeberg Nuclear Power Stations (KNPS), commissioned by the South African government in 1984, is currently the continent's only nuclear power station (Bommer et al., 2015). The location of the KNPS is within the Cape Town cluster and lies within proximity of the Colenso Fault, Milnerton Fault and Table Bay fault (Figure 2C). Identifying which of these fault systems are currently active and most likely responsible for the 1809 and 1811 events will have direct implications for future seismic hazard assessments for the region surrounding the KNPS.

1.2. Aims of this Thesis

This thesis focuses on analysing micro-seismicity near the KNPS, including identifying events and comparing their epicenters with nearby fault systems, and finding regions of elevated seismicity to infer the potential locations of faults responsible for the 1809 and 1811 events. Additionally, it assesses the performance of traditional methods like the Short-Time Average to Long-Time Average (STA/LTA) and visual identification, as well as machine learning, in phase and earthquake identification. The evaluation examines how machine learning compares to traditional methods when applied to a dataset that is from a different region and crustal structure than that of the training dataset.

2. Deployment and Data

2.1. Field Deployment

Given the uncertainty surrounding the epicenters of the 1809 and 1811 earthquakes, and the possibility that they occurred along the Milnerton Fault (Hartnady, 2003), a temporary seismic network was established in the Cape Town area. This network aimed to investigate local seismicity, with particular focus on the region near the KPNS which was undergoing a license renewal application at the time (Figure 3A). The areal extent of the network encompassed the Colenso Fault and the epicenters of the 2020 Cape Town earthquakes reported by both the USGS and the CGS. Additionally, it was located 10 km east from the offshore trace of the proposed Milnerton fault.

The network was primarily situated within the greenschist Malmesbury Group (Figure 3B) and consisted of 18 stations across a 40 by 35 km area north of Cape Town (Figure 4A). Inter-station spacing ranged from 2.73 km to 33.42 km with an overall average of 8.94 km for the whole network. Table 1 provides the coordinates of each station. Each station consisted of a three-component SENSOR Nederland PE-6/B3 geophone linked to an Omnirec DATA-CUBE^3 digitizer sampling at 200 Hz. The geophones used in this deployment had specific parameters associated with them including a natural frequency (f_0) of 4.5 Hz, a coil resistance (R_c) of 375 Ω, a datalogger input impedance (Z_{amp}) of 100 000 Ω, an intrinsic sensitivity (G_0) of 28.8 V/ (m/s), open circuit damping (b_0) of 0.56 and an internal moving mass (m) of 0.0111 kg.

Data was recorded between July 27th and October 25th, 2021 (Figure 4B). Stations ACF, ACZ and AD2 were active for most of the duration of the deployment with stations AD6, ACT and ACX exhibiting gaps in their waveform recording (Figure 4B). The geophones were oriented to true north. Each geophone was inserted into a 30-50 cm hole, with the spikes buried at the bottom to ensure secure coupling with the ground. The data cube and external battery were enclosed in a durable plastic bag, sealed with waterproof tape for protection. After installation, each hole was backfilled, leaving only the GPS antenna exposed. Stations ACT, ACZ, AD6, ACS, AD0, and ACM showed muddy ground conditions upon retrieval. The dataloggers and battery connectors showed signs of corrosion from water infiltration, most likely caused by Cape Town's abnormally high rainfall in August and September.

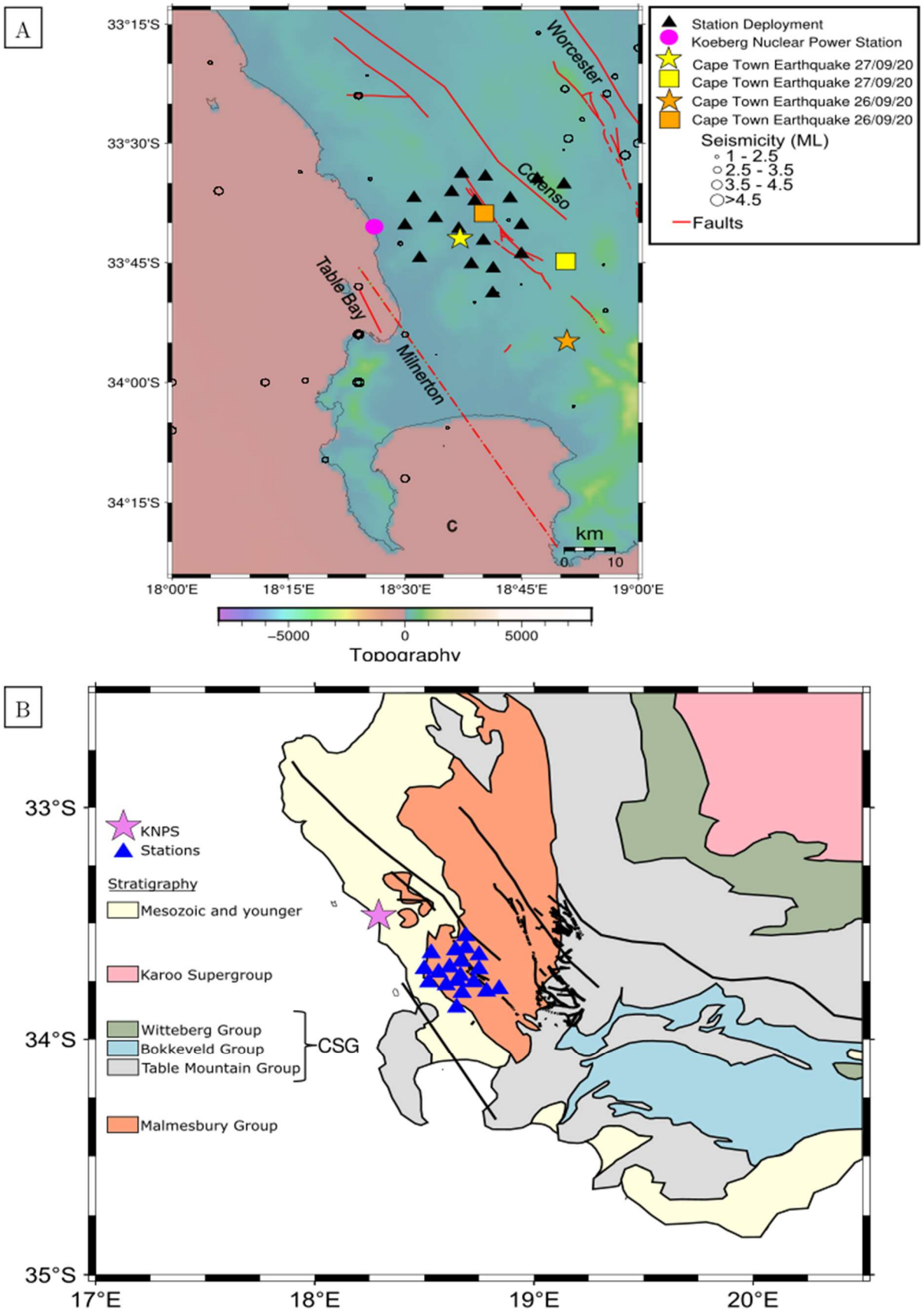


Figure 3: Figure 3A illustrates the overall seismicity of Cape Town (CGS, 2022) and the nearby fault systems, including the Colenso Fault (CGS, 2022), the Milnerton Fault (dashed red) (Hartnady, 2003), and the Table Bay Fault (MacHutchon et al., 2020). The stars represent events recorded by the CGS in 2020, while the squares indicate events recorded by the USGS. Figure 3B depicts the local geology and stratigraphy of the study area (adapted from Flint et al., 2011), along with the location of the KNPS.

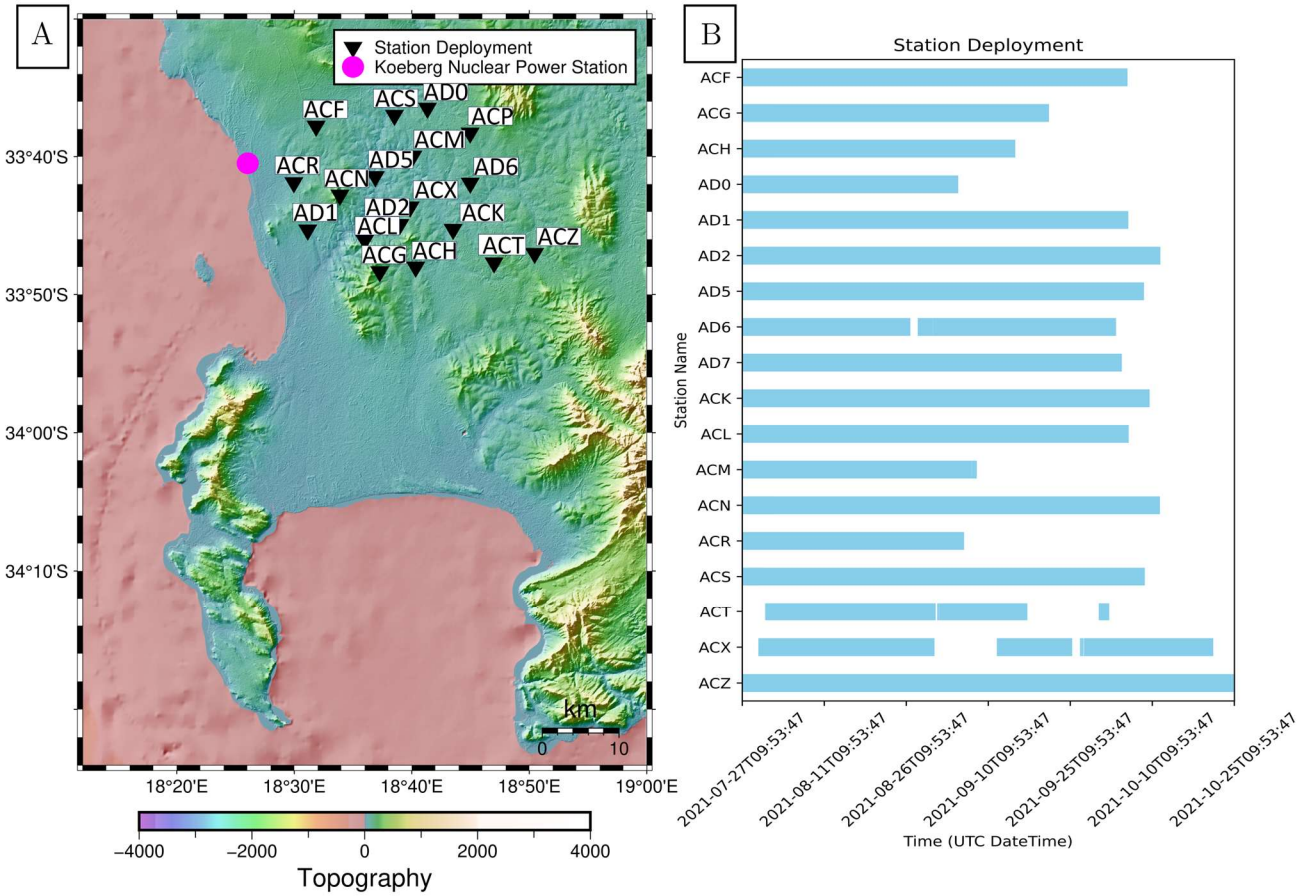


Figure 4: Figure 4A shows the location of each seismic station along with the location of the KNPS. Figure 4B indicates the recording duration of each station.

Table 1: Locations of the seismic stations.

Station	Longitude	Latitude	Elevation (m)
AD7	18.69	-33.56	131
AD0	18.69	-33.60	107
ACP	18.75	-33.64	145
AD6	18.75	-33.70	89
ACM	18.67	-33.67	72
ACF	18.53	-33.63	118
ACR	18.50	-33.70	64
AD1	18.52	-33.76	61
ACL	18.60	-33.77	163
ACX	18.66	-33.73	79
AD5	18.61	-33.70	42
ACS	18.64	-33.62	72
ACK	18.73	-33.75	100
ACZ	18.84	-33.78	157
ACT	18.78	-33.76	159
ACN	18.56	-33.71	253
AD2	18.65	-33.75	99
ACH	18.67	-33.80	107
ACG	18.62	-33.81	230

Figure 5 illustrates the instrument response, with the sensor's lower corner frequency around 4.5 Hz. The upper corner frequency is near 100 Hz, governed by the Nyquist frequency of the 200 Hz sampling rate used in this project.

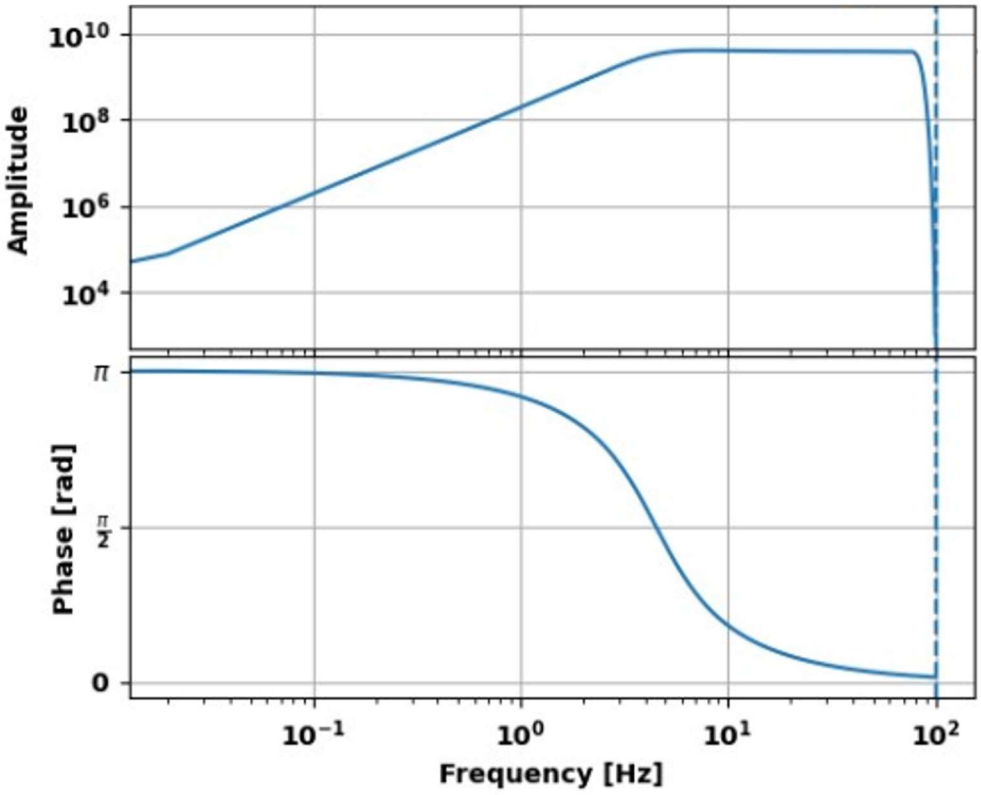


Figure 5: Bode diagram depicting the instrument's velocity response.

2.2. Seismic Data

The recorded data was converted from DiGOS' proprietary cube format to the mini-Standard for the Exchange of Earthquake Data (MSEED) format using cube2mseed, a program provided by DiGOS. It was then converted to Society of Exploration Geophysicists-Y (SEG-Y) format using ObsPy (Beyreuther et al., 2010). Figure 6A shows the raw data used to rapidly scan the SEG-Y data volumes visually for earthquakes, with an example of background noise shown in Figure 6B. The average signal-to-noise ratio across all stations for the earthquake record in Figure 6A is roughly 400, computed from the rms amplitudes of a 2-second window containing the direct P-wave and a 20-second window before the P-wave arrival.

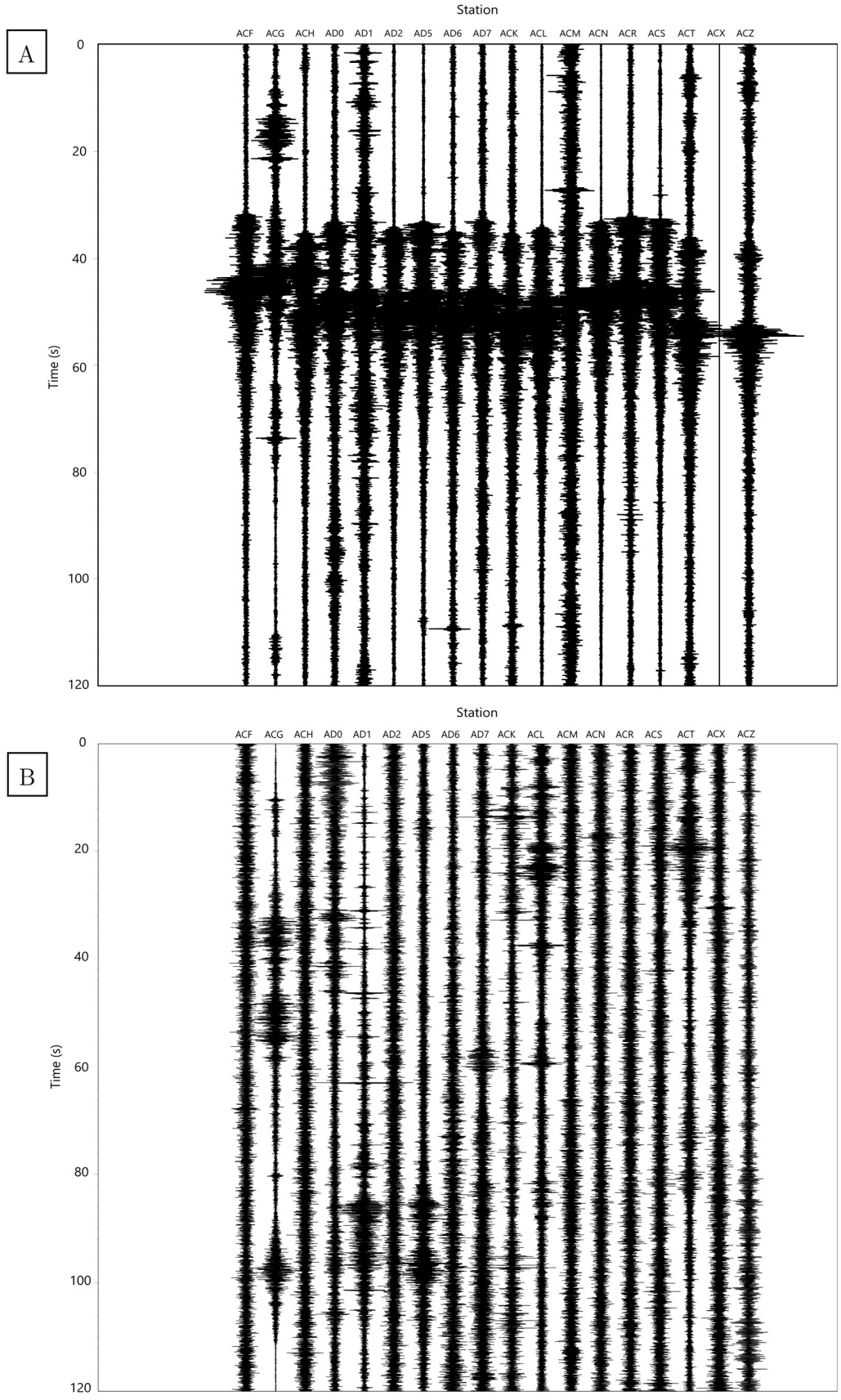


Figure 6: Figure 6A shows the vertical components of each station with an example of an earthquake viewed using Vista. Figure 6B provides an example of noise for the same vertical components.

3. Standard Seismological Methods

3.1. Visual inspection and Manual Picking

Earthquakes in the upper crust are generally the result of shear failure, and as such they generate several different seismic waves at the time of nucleation (Shearer, 2000). The pair of body waves known as P and S, for primary and secondary, are commonly used to identify and locate these events (Shearer, 2000). In standard practice, seismologists visually inspect the data looking for large amplitude arrivals correlated across several seismic stations, such as the ones shown in Figure 7A (Shearer, 2000). Due to the difference in seismic velocities between P and S waves, which are controlled by several elastic moduli, it is possible to get an estimate of epicentral distance (distance from event to the station) by measuring the arrival time difference of the two waves (Shearer, 2000). This method is time consuming, and becomes intractable for very large datasets, such as when there are many stations and or long recording durations. Manual identification of P and S wave forms were done using Vista, while SEISAN was used for picking of phases (Figure 7B) and HYPOCENTER was used for earthquake location.

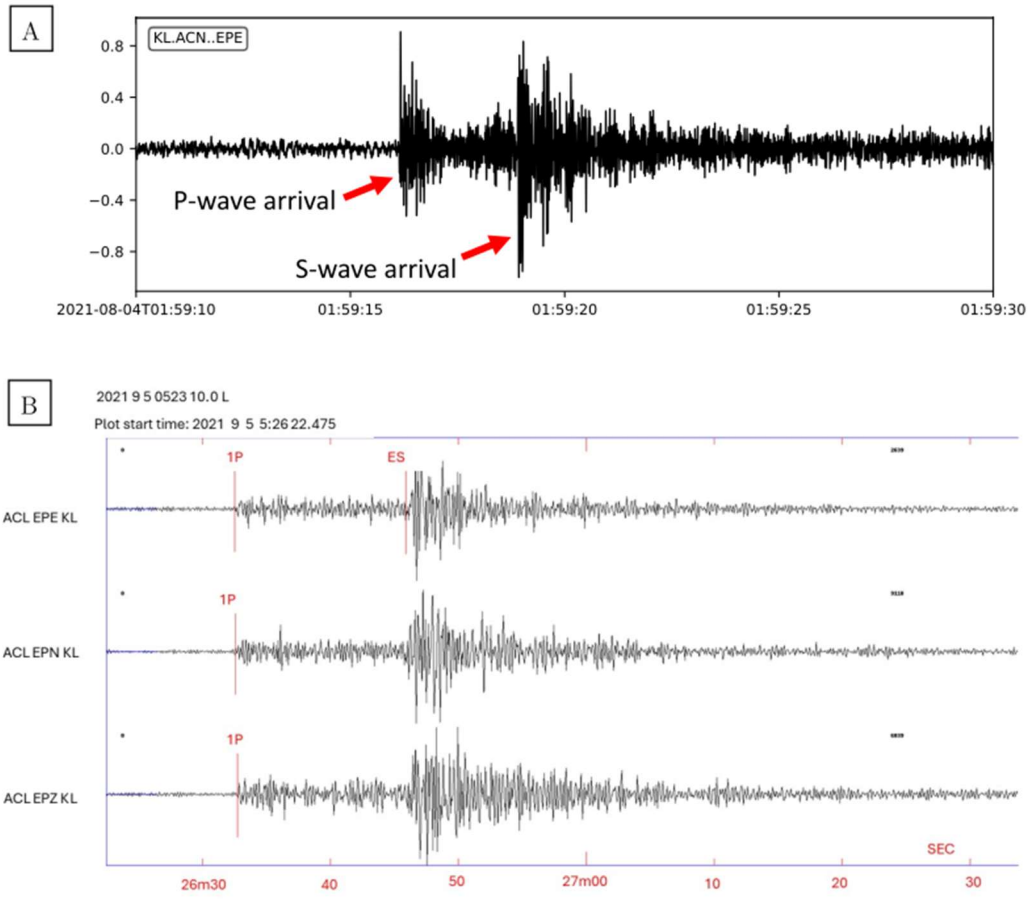


Figure 7: Figure 7A illustrates the two body waves (P and S) commonly used to identify an earthquake. Figure 7B shows an example of how SEISAN was used for phase-picking of P-waves (1P) and S-waves (ES).

3.2. Short-Time Average to Long-Time Average (STA/LTA)

An alternative to visually identifying phase arrivals is the Short-Time Average to Long-Time Average (STA/LTA) algorithm (Gaol et al., 2021; Ross et al., 2019; Trnkoczy, 1999; Vaezi and Van Der Baan, 2015; Woollam et al., 2019). STA/LTA approaches rely on amplitude threshold trigger algorithms when the average amplitude of the Short-Time Average (STA) divided by the average amplitude of the Long-Time Average (LTA) exceeds a predefined user threshold value (Gaol et al., 2021; Ross et al., 2019; Trnkoczy, 1999; Vaezi and Van Der Baan, 2015). When the STA/LTA reaches this threshold at any one station, a channel trigger is recorded (Figure 8). However, for an event to be identifiable, it must be recorded by most of the stations in the network, generating what can be called a "network trigger" (Ahmed et al., 2021; Gaol et al., 2021; Trnkoczy, 1999; Vaezi and Van Der Baan, 2015). After a network event trigger is activated, another user-defined parameter, the STA/LTA de-triggering threshold, is activated (Trnkoczy, 1999). The de-triggering threshold indicates a reduction in the STA/LTA ratio, signaling the end of a network trigger (Trnkoczy, 1999). This thesis utilized the Classic STA/LTA method implemented in Obspy by Beyreuther et al. (2010). Mathematically, the Classic STA/LTA and all its components can be written as:

$$STA_{(i)} = \frac{1}{N_{STA}} \left(\sum_{j=i-N_{STA}}^{i} A_j\right) \tag{3.1}$$

$$LTA_{(i)} = \frac{1}{N_{LTA}} \left(\sum_{j=i-N_{LTA}}^{i} A_j\right) \tag{3.2}$$

$$(STA/LTA)_{(i)} = \frac{STA_{(i)}}{LTA_{(i)}} \tag{3.3}$$

Where STA is the short-term average, LTA is the long-term average, and RSL is the short-term average divided by the long-term average (Beyreuther et al., 2010; Gaol et al., 2021). N_{STA} is the STA window length (in samples), N_{LTA} is the LTA window length (in samples), and A is the absolute signal amplitude (Beyreuther et al., 2010; Gaol et al., 2021).

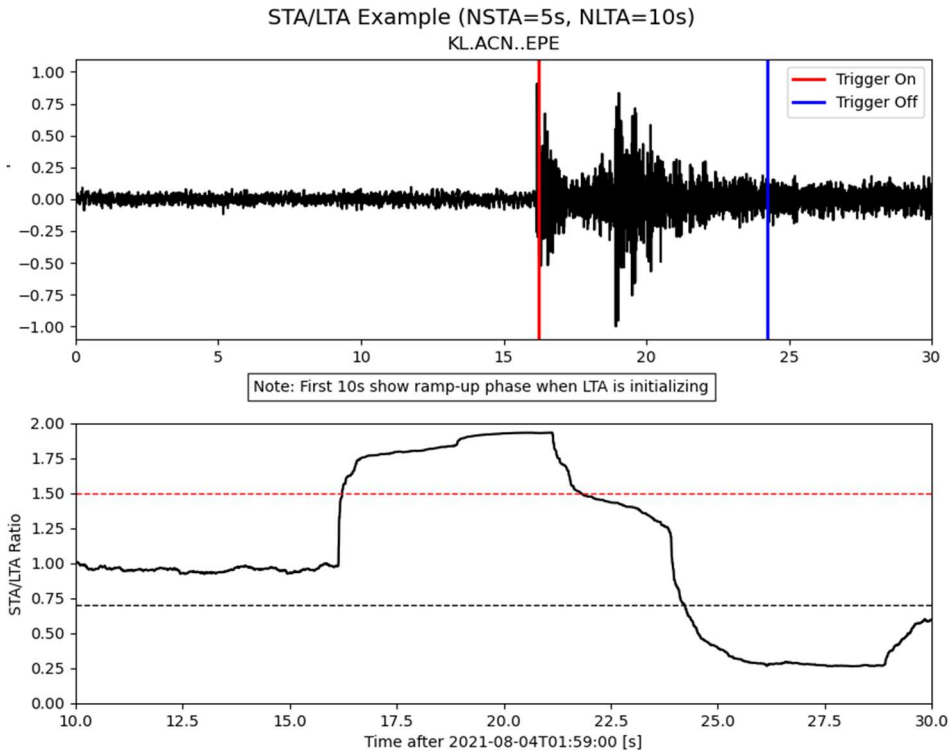


Figure 8: STA/LTA detection example applied to local waveform data from Klipheuwel Network (Station ACN.). Parameters used: NSTA=5s, NLTA=10s, trigger-on threshold=1.5, trigger-off threshold=0.7. Note: First 10s show ramp-up phase when LTA is initializing. These thresholds and parameters were selected for methodological demonstration purposes only and are not intended for universal application.

After considering a range of optimum parameter settings for identifying local micro-seismic activity in noisy data, as indicated in Trnkoczy (1999), the parameters listed in Table 2 were selected for this project. Leeway, given in Table 2, is the period between the initial channel trigger and the number of stations needed for a network trigger.

Table 2*:* Table containing all the STA/LTA parameters used in this thesis for event identification.

Month	N_{STA}	N_{LTA}	STA/LTA on	STA/LTA off	Leeway (in seconds)	Network Trigger
July	100	10 000	9	2	5	5 stations
August	100	10 000	9	2	5	6 stations
September	100	10 000	9	2	5	5 stations
October	100	10 000	9	2	5	4 stations

3.3. Hypocenter Locations

Earthquake locations were determined using HYPOCENTER, a package provided by SEISAN. Earthquake location is a non-linear problem that relates the observed arrival time at station i to the source location k through the following equation (Lienert, Berg and Frazer 1986; Ottemöller et al. 2015):

$$\left(t_{k,i}\right)^{obs} = T(\boldsymbol{x_i}, \boldsymbol{x_k}) + \tau_k \tag{3.4}$$

Where $\boldsymbol{T(x_i, x_k)}$ is the travel time from the hypocenter $\boldsymbol{x_k}$ to station location $\boldsymbol{x_i}$ and incorporates both the velocity model and distances between the receiver and source. Depending on the type of velocity model, travel time can take various forms (Lienert, Berg and Frazer 1986; Ottemöller et al. 2015). The vector formulation of the problem for an n-dimensional arrival times vector $\boldsymbol{t}$ can be written as:

$$\boldsymbol{t^{obs} = f\,(m)} \tag{3.5}$$

Where f maps the model parameters $\boldsymbol{m}$ to the observations.

HYPOCENTER uses a linear-iterative model to find an approximation of $\boldsymbol{m}$, it starts with a starting model vector $\boldsymbol{m_0} = < x_0, y_0, z_0, t_0 >$ where $\boldsymbol{f(m_0)}$ will have different values from $\boldsymbol{t^{obs}}$ and adjusting $\boldsymbol{m_0}$ will have to be done such that predicted arrival times become closer to observed arrival times. The update of $\boldsymbol{m_0}$ is $\boldsymbol{m_1 = m_0 + \Delta m}$ where $\boldsymbol{\Delta m}$ brings the starting model closer to the observed arrival times. Substituting $\boldsymbol{m_1}$ into (3.5) results in the following:

$$\boldsymbol{t^{obs} = f(m_1) = f(m_0 + \Delta m)} \tag{3.6}$$

Further approximation of $\boldsymbol{t^{obs}}$ is achieved using the first order Taylor expansion of $\boldsymbol{f}$ given by:

$$\boldsymbol{t^{obs} = f(m_1) = f(m_0 + \Delta m) \approx f(m_0) + \frac{\partial f(m_0)}{\partial m} \Delta m} \tag{3.7}$$

Substituting the zero-term of the Taylor expansion with $\boldsymbol{f(m_0) = t_0^{obs}}$ and taking it across the approximately equal sign results in defines the *travel-time residual*, expressed mathematically as:

$$\boldsymbol{\Delta t_0 = t^{obs} - t_0^{calculated} = \frac{\partial f(m_0)}{\partial m} \Delta m} \tag{3.8}$$

Noting that $\boldsymbol{m = < x, y, z, t >}$ and that $\boldsymbol{\Delta m = < \Delta x, \Delta y, \Delta z, \Delta t >^T}$, an equivalent form of writing this vector-matrix equation of time-travel residuals is given below, applicable for the first iteration and for n-observed travel-time operations:

$$[\Delta t_1, \dots, \Delta t_n]^T = \begin{pmatrix} \frac{\partial f_1(m_0)}{\partial x}\Delta x & \frac{\partial f_1(m_0)}{\partial y}\Delta y & \frac{\partial f_1(m_0)}{\partial z}\Delta z & \frac{\partial f_1(m_0)}{\partial t}\Delta t \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n(m_0)}{\partial x}\Delta x & \frac{\partial f_n(m_0)}{\partial y}\Delta y & \frac{\partial f_n(m_0)}{\partial z}\Delta z & \frac{\partial f_n(m_0)}{\partial t}\Delta t \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1(m_0)}{\partial x} & \frac{\partial f_1(m_0)}{\partial y} & \frac{\partial f_1(m_0)}{\partial z} & \frac{\partial f_1(m_0)}{\partial t} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n(m_0)}{\partial x} & \frac{\partial f_n(m_0)}{\partial y} & \frac{\partial f_n(m_0)}{\partial z} & \frac{\partial f_n(m_0)}{\partial t} \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \\ \Delta t \end{pmatrix} \quad (3.9)$$

Replacing the matrix containing the partial derivatives with G allows us to rewrite the vector-matrix equations as:

$$\Delta t_0 = G\Delta m \quad (3.10)$$

Note that $\mathbf{G}$ and Δt_0 are known, Δm can be solved by multiplying on the left-hand side by the generalized inverse in the least square sense of $\mathbf{G}$ which is $(G^T G)^{-1} G^T$ denoted by G^{-g} . This provides the solution for Δm as:

$$\Delta m = G^{-g}\, \Delta t_0 \quad (3.11)$$

This gives us the value of the new model parameters closer to the observed arrival times (i.e., the model parameters for the second results from $m_1 = m_0 + \Delta m$). The arrival times from the model parameters can therefore be subsequently calculated as:

$$t_1^{calculated} = f(m_1) \quad (3.12)$$

Which is closer to the observed arrival times. If the misfit calculation obtained from using $\mathbf{m_1}$ is accepted then $\mathbf{m_1}$ becomes the hypocenter and origin time solution, else iteration continues until an acceptable misfit is achieved.

3.4. Velocity Model

An inaccurate velocity model induces uncertainty in hypocentral locations (Gesret et al., 2014; Kanaujia et al., 2015; Midzi et al., 2010). Given the project's focus on hypocenter locations and fault systems, employing the most accurate velocity model possible is critical for reducing uncertainty in observed seismicity near major fault systems and the KNPS.

One-dimensional velocity models have typically been used to calculate hypocenter locations at all scales because they are more efficient than three-dimensional velocity models (Kanaujia et al., 2015; Midzi et al., 2010). The study area for this project is mostly located in the Cape Fold Belt, a tectonic province where Smit et al. (2015) inverted a 1D velocity model (Figure 9), for a segment near this project's study area. They used the VELEST software program to evaluate 59 local earthquakes from their dataset.

In Smit et al. (2015), the change in P-wave velocity of 0.4 km/s at 6 km depth is correlated with a change in subsurface lithologies. Smit et al. (2015) acknowledges that this change could have occurred anywhere between 2 and 6 km but suggests it's most likely at 6 km due to specified input model parameters. Implementation of the Smit et al. (2015) model was used in this project without including the Cape Supergroup since the network lies mostly within the Malmesbury Group (MG) (Figure 3B). Furthermore, the shift from 5.4 to 5.8 km/s at 6 km may indicate a transition in lithology from the MG to the Namaqua Natal Metamorphic Belt beneath it (Smit et al., 2015)

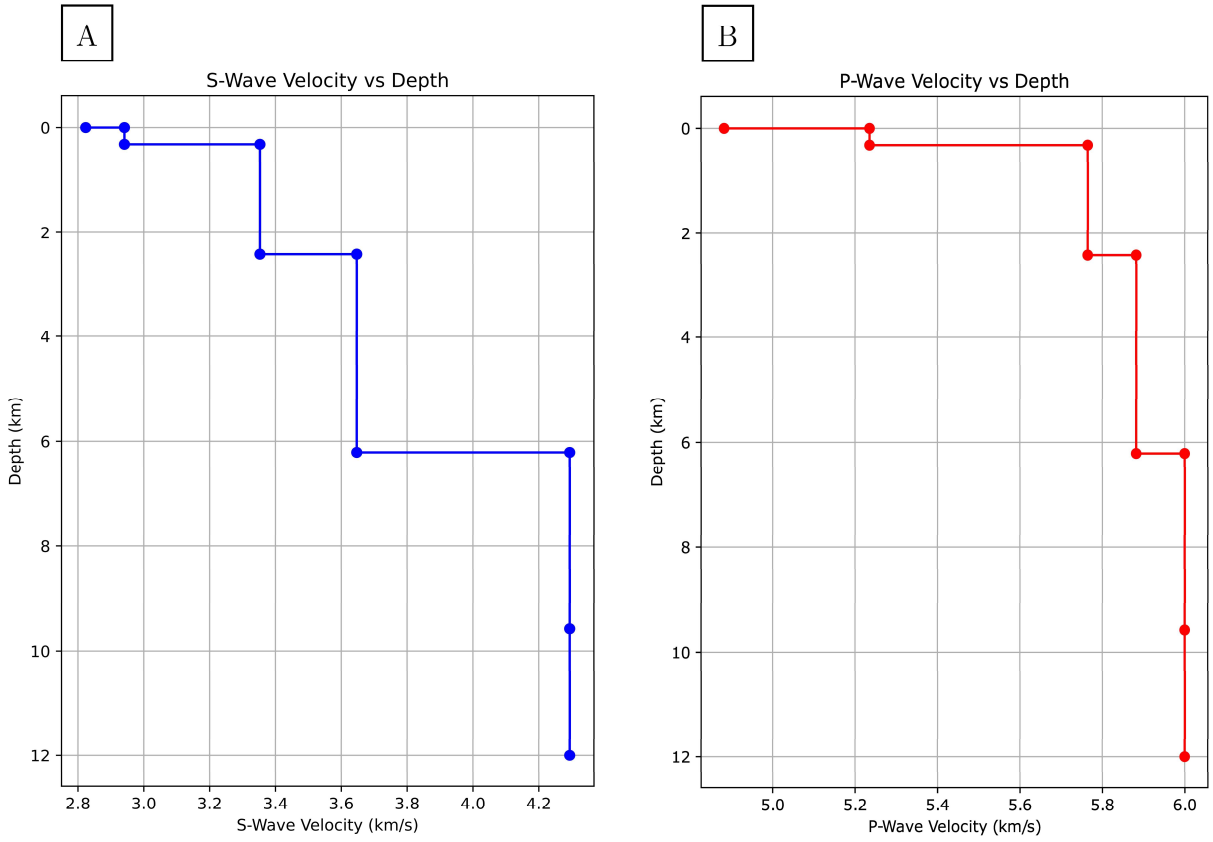


Figure 9: Velocity model provided Smith et al., (2015) for the Ceres-region in the Cape Fold Belt. Figure 9A shows the change in S-wave speed with depth and Figure 9B shows the P-wave speed with depth.

4. Machine Learning Methods

Machine learning (ML), a rapidly evolving branch of artificial intelligence, is increasingly applied to identify patterns, predict outcomes, and categorize both structured and unstructured data (Raschka & Mirjalili, 2019). ML is classified into three types: supervised machine learning, unsupervised machine learning, and reinforcement machine learning (Liermann and Stegmann, 2019; Raschka and Mirjalili, 2019). Supervised and unsupervised machine learning have been widely used in recent years for seismic phase identification, earthquake detection, and phase association and have proven to be more effective in many case studies when compared to traditional methods (Bergen et al., 2019; Mousavi et al., 2020; Münchmeyer et al., 2022; Ross et al., 2019; Woollam et al., 2022; Zhu and Beroza, 2018).

In this thesis, ML algorithms from the Seisbench library were utilized, incorporating a range of training datasets and machine learning models (Münchmeyer, 2022; Woollam et al., 2022). The Seisbench library is composed of several modules namely the Data, Generate and Models modules (Münchmeyer, 2022; Woollam et al., 2022). The Data module contains all the benchmark datasets on which the machine learning algorithms were trained (Münchmeyer 2022; Woollam et al. 2022). Figure 10A shows the station geometry and epicenters from the Southern California Earthquake Data Center (SCEDC) dataset as an example (Hauksson et al., 2020; Münchmeyer, 2022; Woollam et al., 2022). The Generate module is composed of preprocessing function that can be used to train the Seisbench algorithms (Münchmeyer 2022; Woollam et al. 2022). Six deep learning-based detection and phase selection models, along with a waveform denoising model, are included in the Model's framework (Münchmeyer 2022; Woollam et al. 2022). Figure 10B represents the PhaseNet deep-learning model, which is a widely used phase arrival model (Bergen et al., 2019; Kong et al., 2019; Münchmeyer et al., 2022; Ross et al., 2019; Schultz et al., 2020).

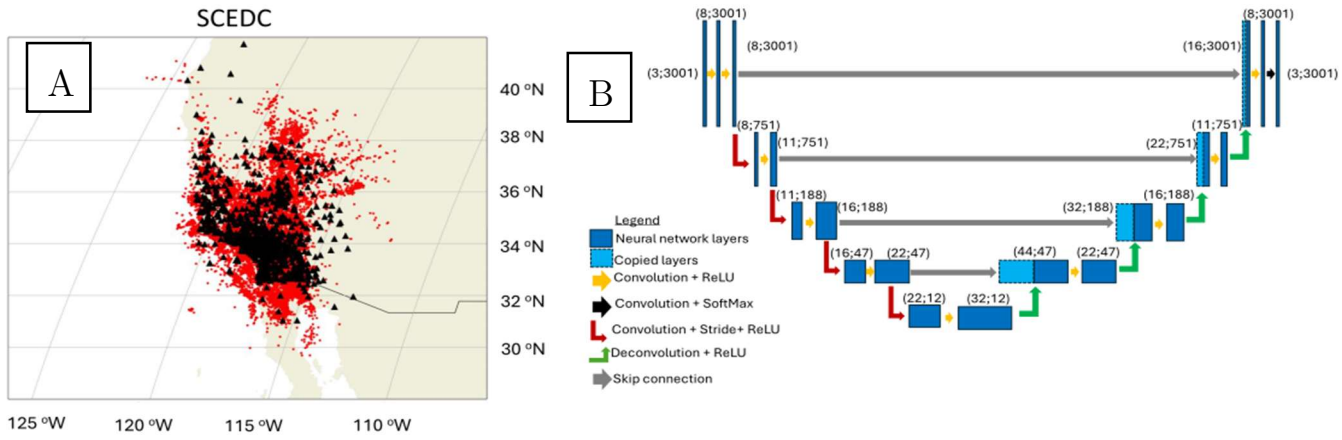


Figure 10: Figure 10A shows the location of the SCEDC dataset as an example of a training dataset (Woollam et al., 2022). Figure 10B shows the network architecture of the PhaseNet model (modified from Zhu and Beroza, 2018).

The next two subsections lay the groundwork for the unsupervised machine learning techniques used in this project. Subsection 4.1.1 discusses the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, explaining how it clusters closely spaced P- and S-picks in space and time while filtering out noise. Subsection 4.1.2 provides the foundation for the Gaussian Mixture Model Association (GaMMA) algorithm, introduced as a seismic phase associator to an event.

4.1. Unsupervised Machine Learning

The use of machine learning algorithms to identify patterns or subgroups based on physical proximity is known as unsupervised learning (Ferraro and Giordani, 2019; Kong et al., 2018; Liermann and Stegmann, 2019; Raschka and Mirjalili, 2019). In this thesis, the focus will be on the subfield of unsupervised machine learning termed clustering.

To illustrate the concept of clustering Figure 11 shows a number of data points. Given the spatial data in Figure 11A, unsupervised machine learning can be applied to identify four unique subgroups as shown in Figure 11B. In this case, the cluster centre determines how closely the data points lie within a specific constraint, which determines the clustering. Noise points are defined as extreme points, or points that are not strongly connected to any other point. There are two subcategories of clustering approaches, soft clustering and hard clustering (Liermann and Stegmann, 2019; Miyamoto et al., 2008). Figure 11B illustrates hard clustering, with each data point assigned to a single cluster and no overlap between them. Soft clustering allows data points to be assigned to multiple clusters with varying degrees of membership, reflecting how well each data point belongs to each cluster (Liermann and Stegmann, 2019; Miyamoto et al., 2008). The only relevant soft clustering strategy in this thesis is model-based soft clustering. The seismological equivalent of data points are P- and S-wave picks, which are grouped based on their close temporal and spatial proximity. Similarly, a cluster can be attributed to a seismic event that generates a signal detectable by most stations in the network. P- and S-wave picks associated with a specific seismic event can be modeled using a model-based clustering method. A soft partition of phase picks and the posterior likelihood of a pick being part of the event are the outcomes of model-based clustering (Ferraro and Giordani, 2019; Zhu et al., 2022). This is the framework through which seismic phase association will be conducted in subsection 4.1.2.

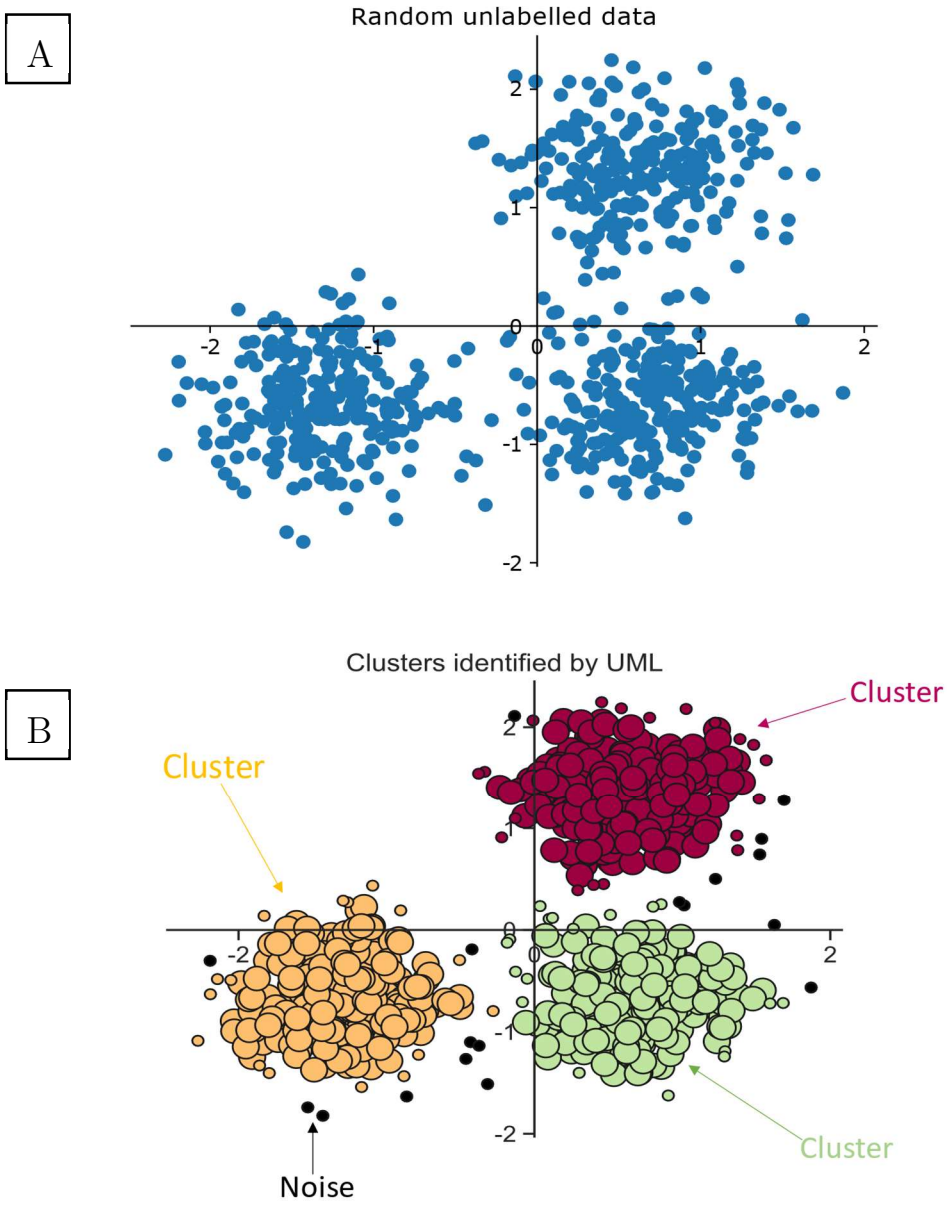


Figure 11: An example of random data without the use of unsupervised machine learning is shown in Figure 11A. The same data are shown in Figure 11B, where DBSCAN, an unsupervised machine learning technique, has been applied. Large circles are used to denote core points, whereas smaller circles are used to highlight border points. Small black circles that are solid are used to signify noise.

4.1.1. Density-Based Spatial Clustering of Application with Noise (DBSCAN)

When the number of data points within a specified radius (epsilon) exceeds a user-defined minimum sample threshold, an algorithm should be able to identify that these data points form a cluster. This is the purpose of Density-Based Spatial Clustering of Application with Noise (DBSCAN), a hard clustering machine learning algorithm (Rehman et al., 2014; Schubert et al., 2017). In this context, a high concentration of data points defines a core point (Rehman et al., 2014; Schubert et al., 2017), conversely, a region with low data point density within the same radius is classified as containing "noise" (Rehman et al., 2014; Schubert et al., 2017). Border points serve as an intermediate point between noise and core points. They are located within an epsilon (ε) radius of a core point but are unable to form a core point due to not having the required number of samples in their epsilon neighborhood (Schubert et al., 2017). To expand identified clusters, border points are essential (Schubert et al., 2017). A new cluster is created by joining the two original clusters if two core points with a given ε intersect and share border points (Schubert et al., 2017). If two core points with a given ε radius intersect and share common border points, a new cluster will form that is the union of the two initial clusters (Schubert et al., 2017).

The DBSCAN algorithm is shown in Figure 12A, with an arbitrarily defined epsilon radius, ε, and a minimum of 3 samples needed to form a core point. Red squares represent core points, which satisfy the need for at least three neighbouring points inside ε. The solid black dots representing border points are located within ε of a core point. In addition, Figure 12A shows the non-empty intersection of two distinct clusters, and Figure 12B illustrates the union of those two clusters. The border point within the intersection is what creates the green cluster that results, as seen in Figure 12B. Figure 12 illustrates the main goal of DBSCAN, which is to eliminate noise from unstructured data and locate clusters according to pre-defined input parameters.

DBSCAN has been successfully used in seismology to cluster seismic picks in both time and space (Grigoli et al., 2018), and to cluster slowness vectors to identify seismic arrivals (Ward et al., 2021). Its performance in these applications demonstrates its reliability in performing such tasks. In this project DBSCAN performs clustering based on identified seismic phase arrivals in the station network. It does this by converting each pick's northing and easting into time equivalents (in seconds) using an assumed average P-wave velocity. This allows the DBSCAN algorithm to use the epsilon radius in seconds to identify picks closely spaced in space and time, as shown in Figure 13. Six spaced clusters in time and space were identified following an event that triggered seismic activity detectable at neighboring stations within the DBSCAN cluster. Crosses mark noise, which DBSCAN filters out.

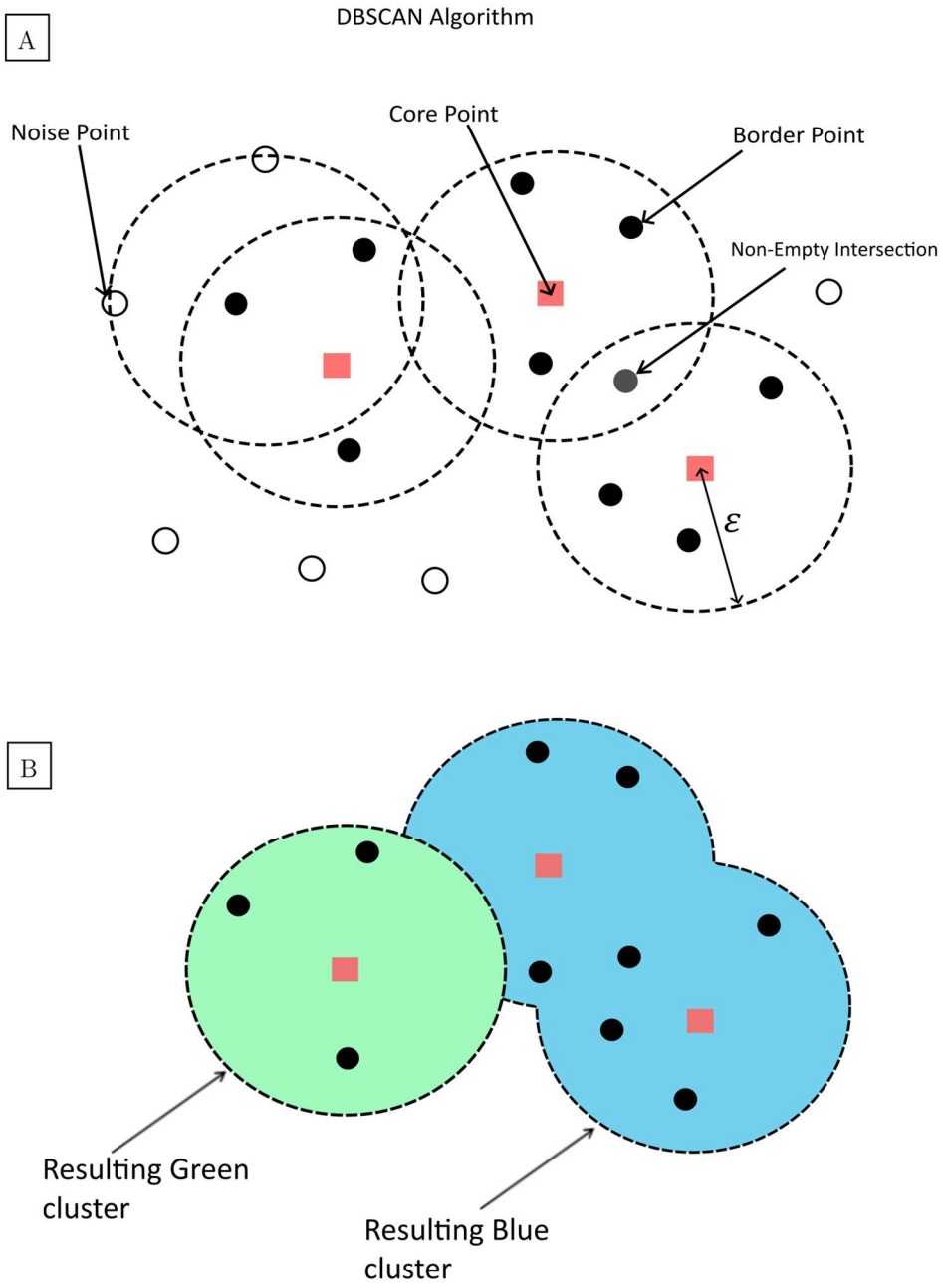


Figure 12: Figure 12A illustrates DBSCAN-identified points such as core points, boundary points, and noise points (modified from Chauhan, 2022). Core points are identified by the presence of more than three data points inside their epsilon radius, indicating their role in cluster formation. Boundary points do not have the required minimum of three neighboring points inside their epsilon, but they do reside within the epsilon radius of a core point. Outliers are defined as noise points that remain outside the epsilon radius of any core point. The non-empty intersection in Figure 12A results in a bigger cluster shown in Figure 12B.

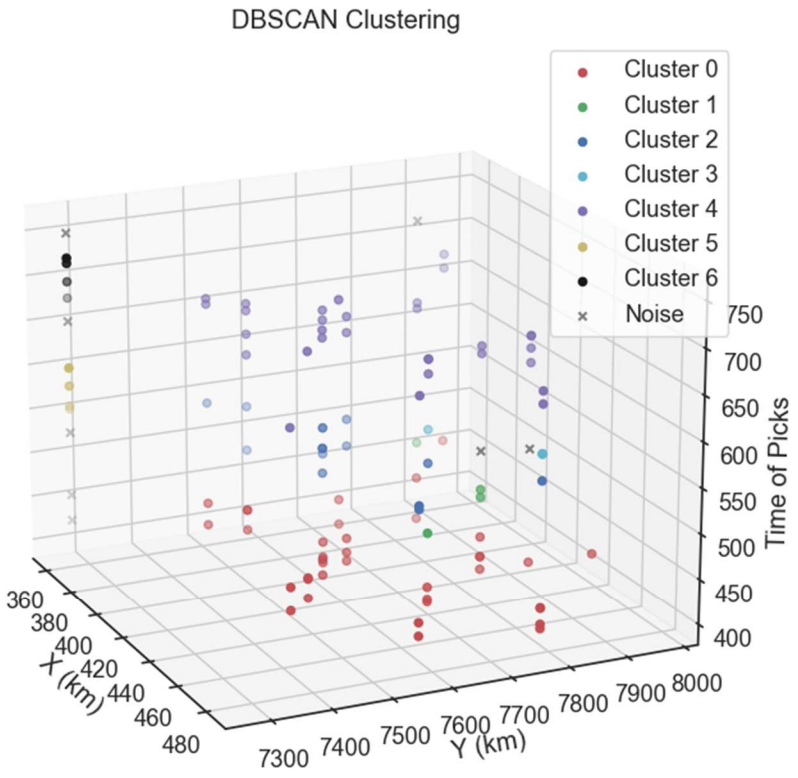


Figure 13: An example of DBSCAN illustrating its use as a pick partition based on station coordinates in Northings and Eastings. A DBSCAN cluster represents a dense region of points separated from lower-density areas. Waveform data sourced from the Chilean Seismic Network (Barrientos and National Seismological Center (CSN) Team, 2018) were used to identify DBSCAN clusters in space with parameters: an epsilon radius of 25 and a minimum of 3 samples

4.1.2. Gaussian Mixture Model Association (GaMMA)

Seismic phase association refers to the mapping of observed groups of phases across a station network to a common hypocenter (Zhu et al., 2022; Ross et al., 2023; Münchmeyer, 2024). Traditionally, this has been accomplished by combining grid-search and back projection approaches to determine a common origin (Ross et al., 2023). However, with the advent of machine learning (ML) in seismology, ML-based phase association algorithms have surpassed traditional approaches (Zhu and Beroza, 2018; Mousavi et al., 2020; Münchmeyer, 2024). In this thesis, phase association was performed using Gaussian Mixture Model Association (GaMMA) (Zhu et al., 2022), a probabilistic ML associator.

Phase association is approached using a probabilistic framework, to determine the likelihood that a set of recorded phases can be assigned to many earthquakes. An optimization algorithm was

then used to match each phase pick to the most likely earthquake (Zhu et al., 2022; Ross and Zhu, 2023; Münchmeyer, 2024). In ML, a similar approach involves determining the likelihood that a set of data may be represented by several Gaussians (Bishop, 2006; Zhu et al., 2022). To gain insight into the ideas behind probabilistic associators used in ML, the following descriptions begin with a single Gaussian (representing one source location), followed by a discussion of numerous Gaussians (representing multiple source locations) modelled via a Gaussian mixture model (GMM).

The probability that a phase arrival at station x_i is generated by an event with hypocenter x_m can be modelled by the univariate Gaussian equation given by:

$$p(t_i|\mu,\sigma^2) = \gamma(\mu,\sigma^2) \tag{4.1}$$

Where t_i is the observed arrival time, μ the theoretical arrival time $t(x_i, x_m)$ between source location and station location, and σ^2 the variance (Zhu et al., 2022; Ross and Zhu, 2023). The theoretical arrival time follows a hyperbolic moveout and is given by the equation $t(x_i, x_m) = \frac{|x_i - x_m|}{v} + t_o$ where v is a user-defined velocity for P and S-wave speeds depending on the pick identified (Zhu et al., 2022; Münchmeyer, 2024). From here on, the terms "Gaussian" and "earthquake" will be used interchangeably.

This approach can be used to solve the forward problem indicated in the introduction: estimating the likelihood of a phase arrival at station x_i is generated by k different earthquakes. This is accomplished by altering equation (4.1) as follows (Zhu et al., 2022; Ross and Zhu, 2023):

$$p(t_i|\mu_k,{\sigma_k}^2) = \ w_i \sum_{k=1}^{K} \varphi_k \gamma(t_i,|\mu_k,{\sigma_k}^2) \tag{4.2}$$

Where t_i is the observed arrival time, μ_k the theorical arrival time between the kth earthquake at source location x_k and station location x_i (also the mean of the kth Gaussian), and $\sigma_k$the variance of the kth earthquake (Zhu et al., 2022). The w_i and φ_k terms are the phase quality score of the identified pick and mixture coefficient of the kth earthquake, respectively. Two restrictions apply to equation (4.2), the first is that $\sum_{k=1}^{K} \varphi_k = 1$ and the second that $0 \leq w_i \leq 1$.

Equation (4.2) is referred to as a Gaussian Mixture Model (GMM) composed of K different Gaussians. Figure 14 shows an example of what a GMM would look like in 1 dimension with 3 different mixture components (i.e., the mixture model is composed of 3 different Gaussians). The goal is to determine the best Gaussian in the GMM for each phase arrival, which necessitates using a maximization strategy to optimize the association of each arrival with a specific Gaussian (Bishop, 2006; Zhu et al., 2022).

Let $\theta = \{\varphi_l, \mu_l, \sigma_l : 0 < l \ll K\}$ be the set containing all unknown parameters of each Gaussian in the GMM (Bishop, 2006). Assuming m time arrivals are independent and identically distributed (i.i.d.), the log-likelihood of the GMM's probability density function can be introduced (Bishop, 2006; Zhu et al., 2022):

$$L(\theta) \ = \ \sum_{i=1}^{m} ln(\gamma(t_i|\theta)) = \sum_{i=1}^{m} ln(\sum_{k=1}^{K} \varphi_k \gamma(t_i,|\mu_k,{\sigma_k}^2)) \tag{4.3}$$

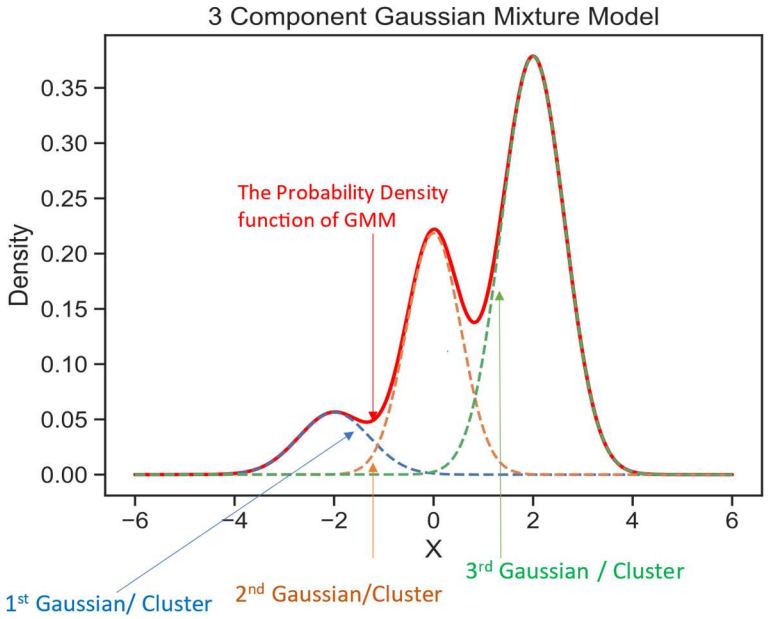


Figure 14: This figure portrays a GMM in one dimension, which is made up of three unique mixture components (Gaussians). A dashed line represents each Gaussian component, highlighting their distinct contributions to the total GMM. The weighted sum of these various components yields the combined density curve (solid red line).

Maximizing equation (4.3) concerning θ poses several challenges. First, the solution pertains to the entire GMM, rather than each individual Gaussian component (Bishop, 2006). Second, the solution does not have a closed-form expression (Bishop, 2006).To calculate the ideal parameters of each Gaussian to explain the data or to approximate the likelihood of the data, an iterative process is required (Bishop, 2006). The iterative process that is commonly used in the context of maximizing GMMs is the *Expectation Maximization (EM)* algorithm (Bishop, 2006; Zhu et al.,2022).

Figure 15 illustrates an example of a four-component GMM, which identified two associated earthquakes using the EM algorithm. Additionally, Equation 4.1 represents an individual univariate Gaussian for a specific data point, assuming an orange cross-hypocenter. When focusing solely on this data point, the combined hypocenters (blue, green, red, and orange mixture components) are indicated, with the arrow next to Equation 4.1 highlighting the relationship captured by Equation 4.2. The maximization of Equation 4.2 via the EM algorithm is what designates the "associated" earthquake, represented by the blue mixture component for example. Furthermore, note the arrival times of the picks and the mean, which reflects the hyperbolic moveout from the earthquake's epicenter, modeling the blue mixture component quite accurately.

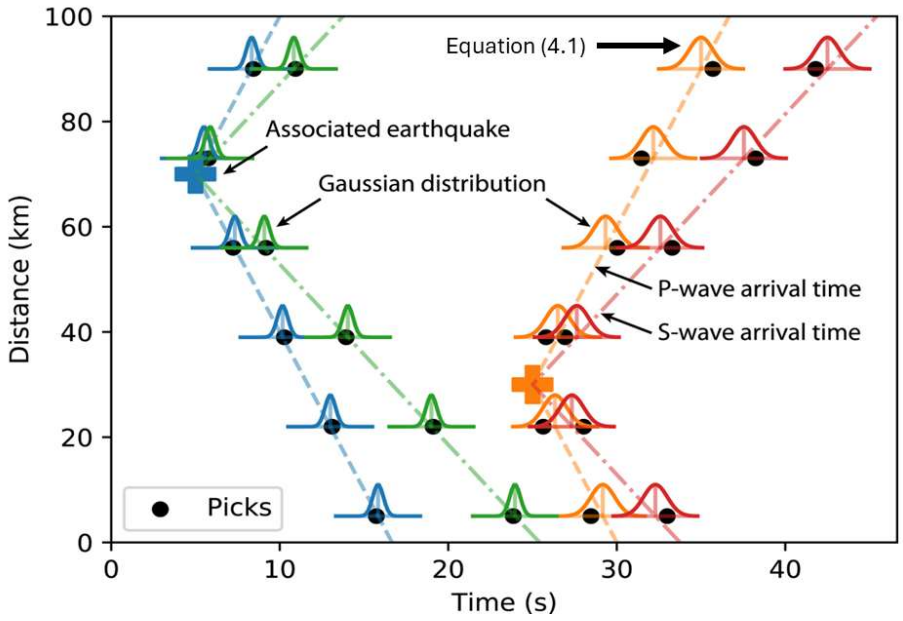


Figure 15: Example of a 4-mixture component GMM used for phase association (modified from Zhu et al., 2022).

One drawback of the EM algorithm is that the number of underlying clusters must be determined prior to its application (Bishop, 2006; Zhu et al.,2022). It is possible to overestimate the number of clusters and use an algorithm that reduces the impact of overestimation by setting the unnecessary Gaussians' mixing coefficients near zero (φ_k in equation 4.2) (Zhu et al., 2022). To solve this, priors are added to the posterior probabilities using the Bayes rule into the EM-algorithm, resulting in the creation of what is referred to as Bayesian Gaussian Mixture Models (BGMM). Zhu et al. (2022) used a BGMM algorithm for earthquake association, by introducing three conjugate prior distributions to address the latent variables of an overestimated clustered GMM, specifically a Dirichlet prior for the mixture coefficient, which determines the scaling factor of a specific earthquake. Earthquakes that do not help explain the findings will have a scaling factor close to zero (Zhu et al, 2022). Zhu et al. (2022) also employed a Gaussian prior distribution for mean conditioned on precision, followed by a Wishart prior distribution for precision. Exact update rules are provided in (Bishop, 2006; Zhu et al.,2022), with an additional minimization step incorporated into BGMM after finding the ideal Gaussian (Zhu et al., 2022). This step determines the origin time and earthquake location forming the Gaussian Mixture Model Association (GaMMA) algorithm that was used in this thesis and provided by (Zhu et al., 2022).

4.2. Supervised Machine Learning (SML)

Supervised machine learning (SML) includes regression and classification, with classification more commonly used in seismology (Waldeland et al., 2018; Woollam et al., 2019, 2022; Zhu and Beroza, 2018). Classification is used to establish decision boundaries between classes, such as P-wave and S-wave arrivals. Decision boundaries, as shown in Figure 16A, are learned from training data, enabling neural networks to classify new information. A basic neuron representation (Figure 16B) receives input vectors derived from training examples. Weights connect input nodes to the neuron, forming the network (Raschka and Mirjalili, 2019). Each line has a value associated with it and is referred to as the weight of the line that connects the input node to the neuron (Raschka and Mirjalili, 2019). The weights play an important role since they become the "learnable" parameters that describe the input vector during the training process (Woollam et al., 2019; Raschka and Mirjalili 2019). The neuron itself is composed of a weighted input (z) and an activation function f (Raschka and Mirjalili, 2019), where the neuron's output depends on the activation function.

In the simplest terms, training a network begins with the output from a network initialized with random weights, consistent with the architecture outlined in Figure 16B. The output from the neuron serves as the predicted output from the network based on the weighted sum of the input and the criteria outlined by the activation (i.e., if the weighted sum is greater than some than the y=x line in Figure 16A then it will be assigned to Class A) (Raschka and Mirjalili, 2019). To calculate the error in the predicted output, the difference between the predicted output and the true output is determined, and this difference is used to update the weights during training (Raschka and Mirjalili, 2019). This process continues through each training iteration (or epoch) until the error approaches zero (Raschka and Mirjalili, 2019). At this point, the weights represent a model of the training data, enabling the classification of new data that was not included in the training set (Raschka and Mirjalili, 2019).

The type of neural networks that are mostly used in seismology and image segmentation are Convolutional Neural Networks (CNNs) (Waldeland et al., 2018; Woollam et al., 2019, 2022; Zhu and Beroza, 2018). Most CNNs have a similar architecture that is composed of an input image or time series, convolutional layers, pooling layers, and transpose convolutional layers (i.e., deconvolutional layers/up sampling) as shown in Figure 16C (Dertat, 2018, Woollam et al, 2019). The most fundamental part of a CNN is the convolutional layer where feature maps get generated by an elementwise product between filters and the input image followed by the summation of all entries (Ahmed et al., 2021; Dertat, 2018; Woollam et al., 2019). Feature maps are important since they are generated by repeated convolutional operations to input data, resulting in a set of learnable filters that automatically engineer the appropriate features for classification (Ahmed et al., 2021; Woollam et al., 2019). During training, the objective is to find optimal weights that minimize error. For CNNs, these weights serve as entries to kernels and deconvolutional kernels for generating and up-sampling feature maps (Woollam et al., 2019).

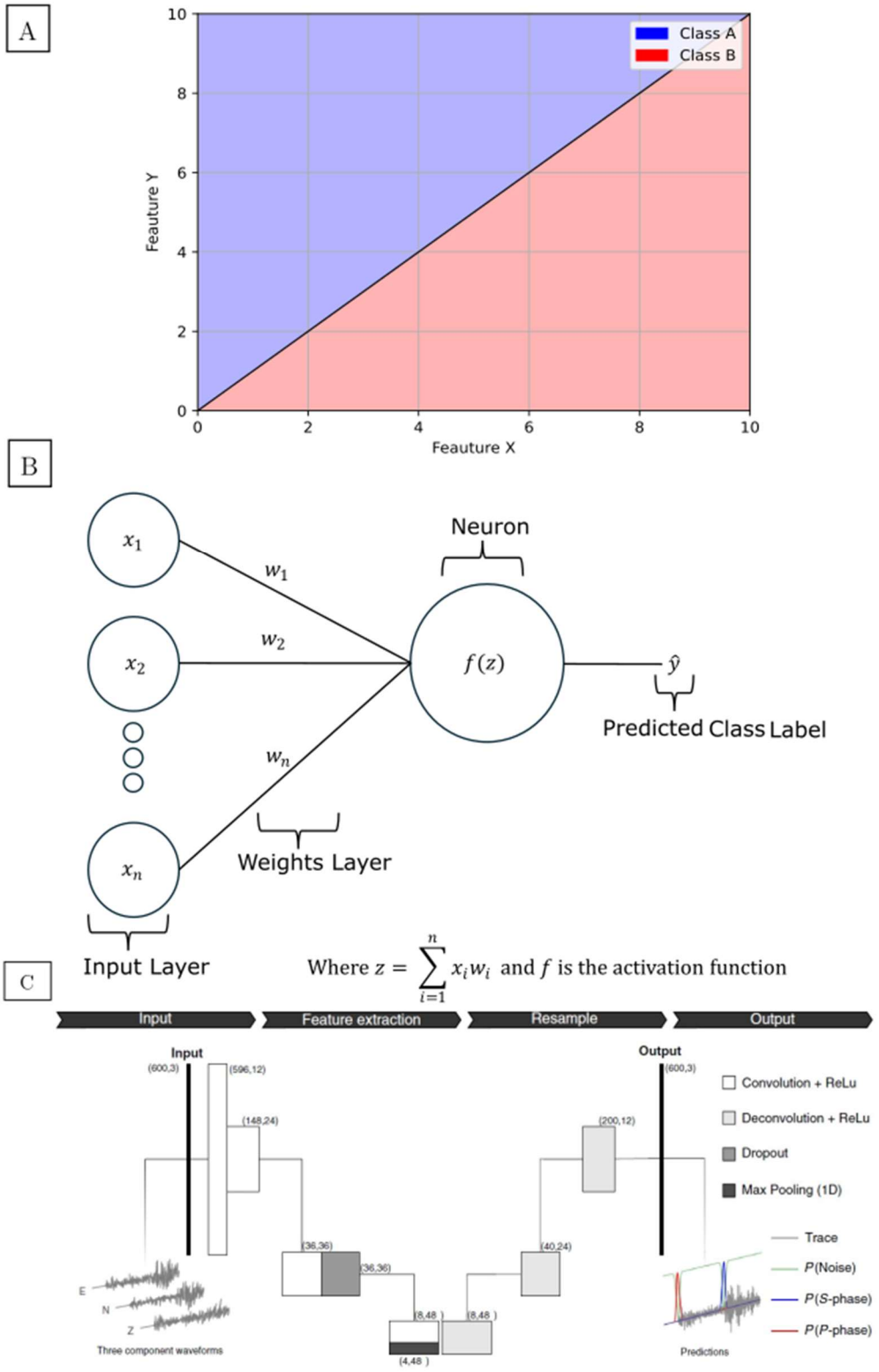


Figure 16: Figure 16A illustrates a decision boundary between two classes, based on two features that are shared by both classes. Figure 16B indicates the simple outlay of a single neuron with multiple inputs and single output. Figure 16C shows an example of a typical CNN architecture, illustrating the process from three-component waveform data to three probability classes (Woollam et al., 2019).

For this project, PhaseNet, a CNN model for identifying phase arrival times in waveforms, was used (Zhu and Beroza, 2018). The following subsection is dedicated to explaining the training process for this specific model. Note that no training was conducted in this project, only model applications were performed using the available datasets, this is due to the short experiment duration (2.8 months) and the low seismicity rates expected for the region.

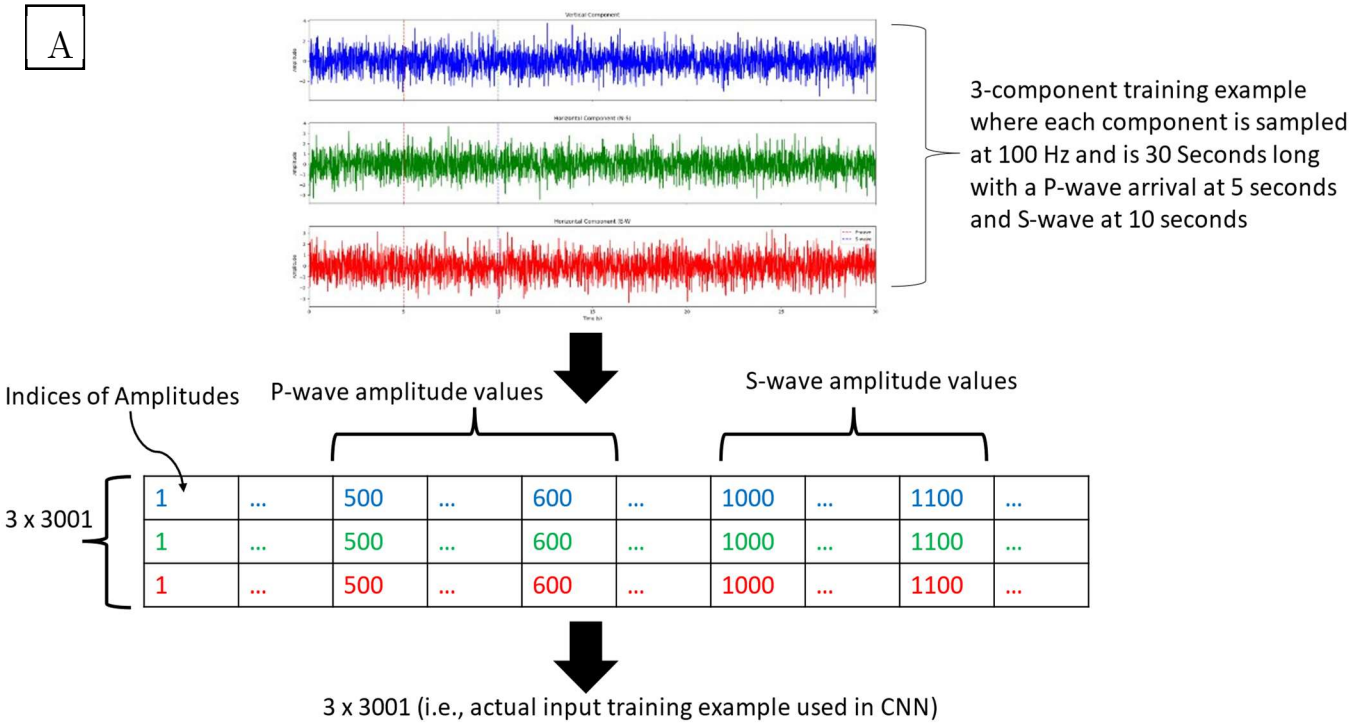
4.2.1. PhaseNet

PhaseNet was trained on 30s window length data sampled at 100 Hz (Woollam et al., 2022; Zhu and Beroza, 2018). Figure 17A illustrates what the input matrix of a training example looked like for PhaseNet. In this example, the P-wave arrives at 5 seconds and the S-wave at 10 seconds. Each 30-second channel is transformed into a 3001-amplitude vector where the indices of values between 500 and 600 represent the P-wave and the indices between 1000 and 1100 represent the S-wave (i.e., since it was sampled at 100 Hz). The network architecture of PhaseNet follows a U-Net architecture that was created by Ronneberger et al. (2015) for biomedical image processing (Zhu and Beroza, 2018). It was trained with the intention of localizing a time series composed of earthquakes into three probability classes: P-wave class, S-wave class, and Noise (Zhu and Beroza, 2018). Figure 17B shows the overall architecture of the neural network with emphasis on the Encoder (i.e., decreasing branch, down sampling) and Decoder (i.e., ascending branch, up-sampling) along with the skip connection between them.

The encoder is for extracting and shrinking useful information from raw seismic waveform to a few neurons so that each neuron in the last layer of this branch will contain a broadly receptive view of the input waveforms (Zhu and Beroza, 2018). The decoder expands on this last layer and proceeds to classify them into three classes. The skip connection at each step concatenates the left-most output to the right-most to improve convergence during training (Zhu and Beroza, 2018).

The last layer, as shown in Figure 17B, has dimensions of 8 by 3001, meaning there are 8 feature maps whose dimensions correspond to the input training examples. PhaseNet aims to convert the input vector into a probability distribution, but in the last layer, the information still pertains only to the amplitude values of the input training example. Zhu and Beroza (2018) describe these values as unscaled values that serve as input for the SoftMax normalized exponential to produce a probability distribution. Mathematically, this is written as:

$$q_i(x) = \frac{ez_i(x)}{\sum_{k=1}^{3} e^{z_k(x)}} \text{ where } i = 1,2,3 \text{ represents noise, P and S classes (4.4)}$$



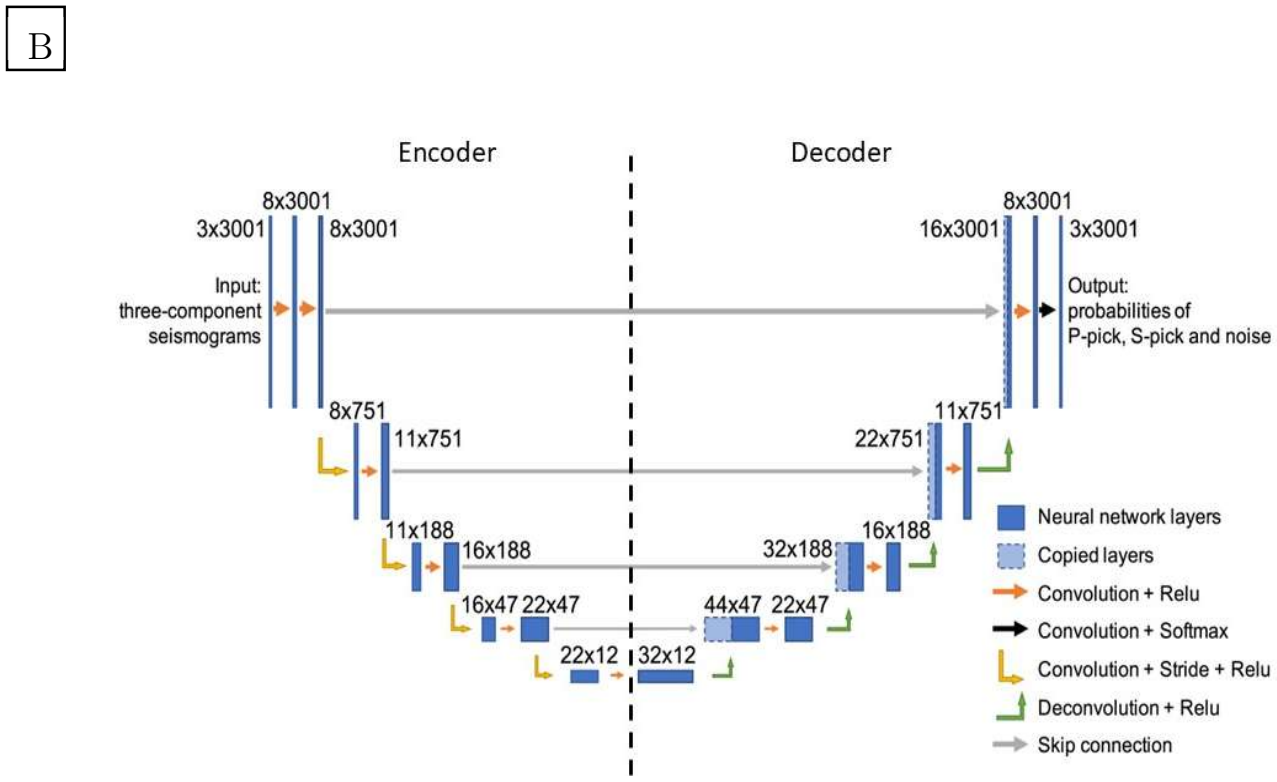


Figure 17: Figure 17A shows what the input matrix looks like. This includes going from a 3-component time series data of 30 seconds sampled at 100 Hz to a 3 by 3001 matrix that is used in PhaseNet's training as discussed in the text. Figure 17B shows the network architecture of PhaseNet (modified from Zhu and Beroza, 2018).

Where z(x) represents the unscaled entries in each of the 3001 feature vectors. Figure 18 shows what this would be like when computed for the last layer and the general order of operations from moving from the unscaled values to a probability distribution for each data point.

The theme for this subsection is supervised machine learning, requiring an introduction to the labeled ground truth vector and the loss function used to evaluate the error between predicted probabilities (equation 4.4) and actual ground truth values.

Ground truth values will be set to 1 at each phase arrival. For instance, in Figure 17A, the ground truth will have a probability of 1 at $q_1(500)$ to indicate the P-wave arrival at 5 seconds within the 30-second window. Similarly, a probability of 1 at $q_2(10000)$ will mark the S-wave arrival at 10 seconds. Each labeled arrival time will form a Gaussian distribution, with means coinciding with the 1s and a standard deviation of 0.1 seconds (Zhu and Beroza, 2018). This ensures that the highest probability is at the arrival time, with reduced uncertainty at nearby points (Zhu and Beroza, 2018). Noise, as described by Zhu and Beroza (2018), is calculated based on the labeled truths at q_1 and q_2, and is trained using P(Noise) = 1 - Probability(P-wave-arrival) - Probability(S-wave-arrival).

The loss function that Zhu and Beroza (2018) used is the cross-entropy loss function and is mathematically formulated as:

$$H(p,q) = -\sum_{i=1}^{3}\sum_{x} p_i(x)\ln\left(q_i(x)\right) \tag{4.5}$$

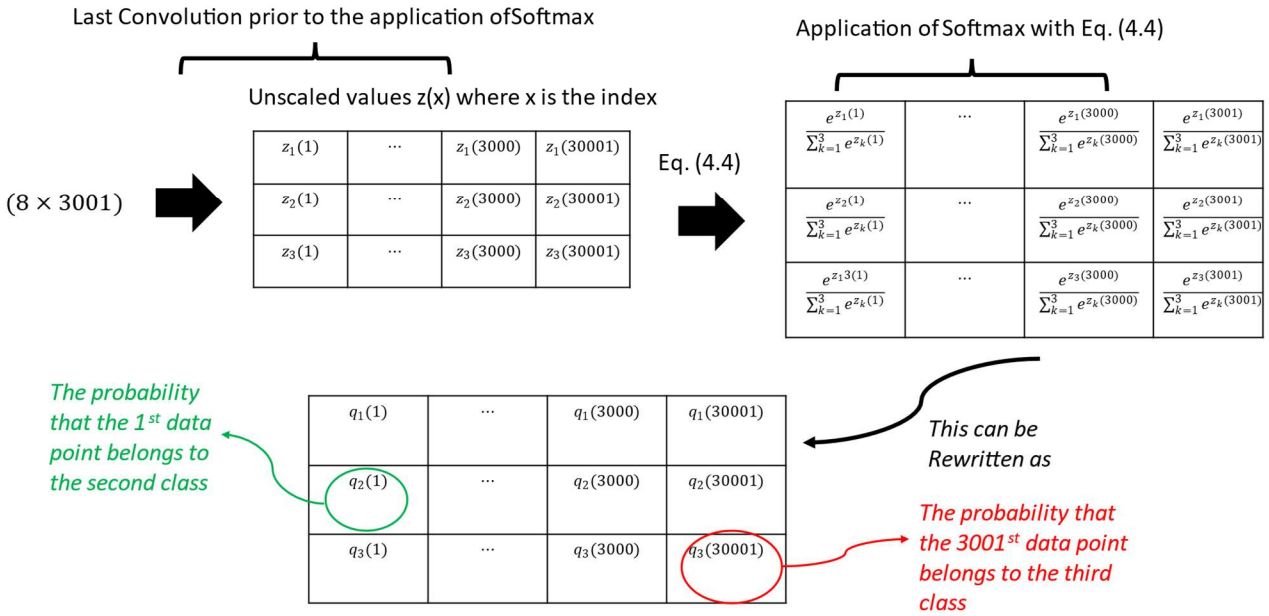


Figure 18: Order of operations going from 8 by 3001 feature maps to a 3-class probability distribution

Where $q_i(x)$ is the predicted probability distribution and $p_i(x)$ the true probability distribution. This loss function is used to measure the given between these two given probability distributions.

4.3. Applications of Machine Learning in this project

This section provides a brief overview of the machine learning workflow implemented in this thesis, covering the application of pre-trained models, clustering and phase association, and a summary of specific parameters.

First, the supervised machine learning methods used in the thesis are discussed, focusing on the application of pre-trained models on the waveform data. The workflow commenced with loading the waveform data and the associated station inventory, followed by segmenting the data into monthly intervals. Subsequently, the pre-trained PhaseNet model (Zhu and Beroza, 2018), as discussed in subsection 4.2.1, was applied using the trained weights from the Southern California Earthquake Data Center (SCEDC) dataset (Hauksson et al., 2020). Before applying the model, the waveform data underwent pre-processing steps such as detrending and resampling to 100 Hz to match the sampling rate of the training set (Woollam et al., 2022; Zhu and Beroza, 2018). The model application produced outputs including picks with arrival times, station positions, probability scores, and phase types. Following this, pick times were normalized relative to the earliest pick, and station coordinates were converted into Northings (km) and Eastings (km).

Then second step in the workflow deals with unsupervised ML and process of clustering and phase association. The former is implemented with DBSCAN, as it clusters data based on a defined radius and the minimum number of data points required for a cluster, as described in subsection 4.1.1. It accomplishes this by translating each pick's northing and easting coordinates to time equivalents (in seconds) using the assumed p-wave speed of 6 km/s, with a Vp/Vs ratio of 1.73 (Mooney, 2007). This method allows DBSCAN to use the epsilon radius in seconds to detect picks that are tightly separated both geographically and temporally, as seen in Figure 13.

Each DBSCAN cluster undergoes initialization of an overestimated number of BGMMs (subsection 4.1.2) using the GaMMA algorithm. The overestimated number of clusters is proportional to the number of picks in the DBSCAN cluster divided by the number of stations in the network which is referred to as the over-sampling ratio as discussed in Zhu et al. (2022). The result is an output catalogue of events.

Finally, the PhaseNet machine learning algorithm was evaluated using multiple DBSCAN input parameters, benchmark datasets, and P- and S-wave detection thresholds. The ideal parameters for minimizing false detections were determined using waveform data from three earthquakes on September 1, 2021 (Table 3). An additional filter was added based on the characteristics of these 3 earthquakes. Each earthquake was consistently identified with at least six P-wave picks, while false positives had fewer P-wave picks.

Table 3: All ML parameters used in this thesis.

Model	Benchmark Dataset	DBSCAN (Min. samples)	DBSCAN (epsilon)	P and S detection threshold	Additional Filters
PhaseNet	SCEDC	5	3 seconds	40%	6 P-wave detections.

To validate the machine learning methods, the dataset was further evaluated using two regularly utilized approaches: manual waveform identification and the Short-Time Average to Long-Time Average (STA/LTA) trigger and discussed in sections 3.1 and 3.2, respectively.

5.Results

5.1. Visual Inspection and Manual Picking Results

As a comparison to the machine learning results, visual inspection of the data resulted in the identification of 41 events. Out of these 13 events were identified as blasts from quarries located in the greater Cape Town area, and the remaining 28 events were identified as earthquakes based on their timings and waveform characteristics. Earthquakes were recognized using the unique P- and S-waves outlined in subsection 3.1. The low signal-to-noise ratio of the data and the use of high frequency geophones made first motion picking challenging, resulting in the omission of magnitude calculations for this project. Picking S-wave arrivals was likewise problematic; thus, most earthquake locations were determined using P-wave arrivals from at least eight stations where possible.

Quarry blasts were identified by their origin times during the day and depth (Table 4), proximity to known quarry locations (Figure 19A), and lower S-wave amplitudes (Figure 19B). The blasts shown in Figure 19A, which have no known associated quarry locations, exhibit waveforms like those in Figure 19B, but may also be misidentified earthquakes.

Table 5 lists every earthquake that has been manually identified, and Figure 20 provides a map of the epicenters of each earthquake. It is also likely that some of these earthquakes in the network could be quarry blasts. However, both blasts and earthquakes are considered true positives when identified by a method, and therefore do not affect the evaluation of each method's effectiveness (subsections 5.4 and 6.1).

Table 4: Manually identified 13 quarry blasts with the time of the blasts, latitude and depth.

Seisan Origin Time	Latitude	Longitude	Depth (km)
2021/08/10 11:59:51	-33.82	18.578	0.3
2021/08/16 13:39:35	-33.992	18.744	0.1
2021/08/22 00:52:27	-33.811	18.58	0.2
2021/08/24 10:23:33	-33.816	18.579	0.1
2021/08/26 10:58:40	-33.826	18.556	0.2
2021/08/31 15:30:26	-33.811	18.556	0.3
2021/09/04 09:54:36	-33.771	18.57	0
2021/09/09 12:23:59	-33.422	18.768	0.5
2021/09/14 11:43:37	-33.814	18.586	0.2
2021/09/15 15:53:54	-33.778	18.578	0
2021/09/20 16:08:59	-33.764	18.571	0
2021/09/29 11:19:18	-33.907	18.703	0
2021/09/30 15:30:37	-33.803	18.56	0.1

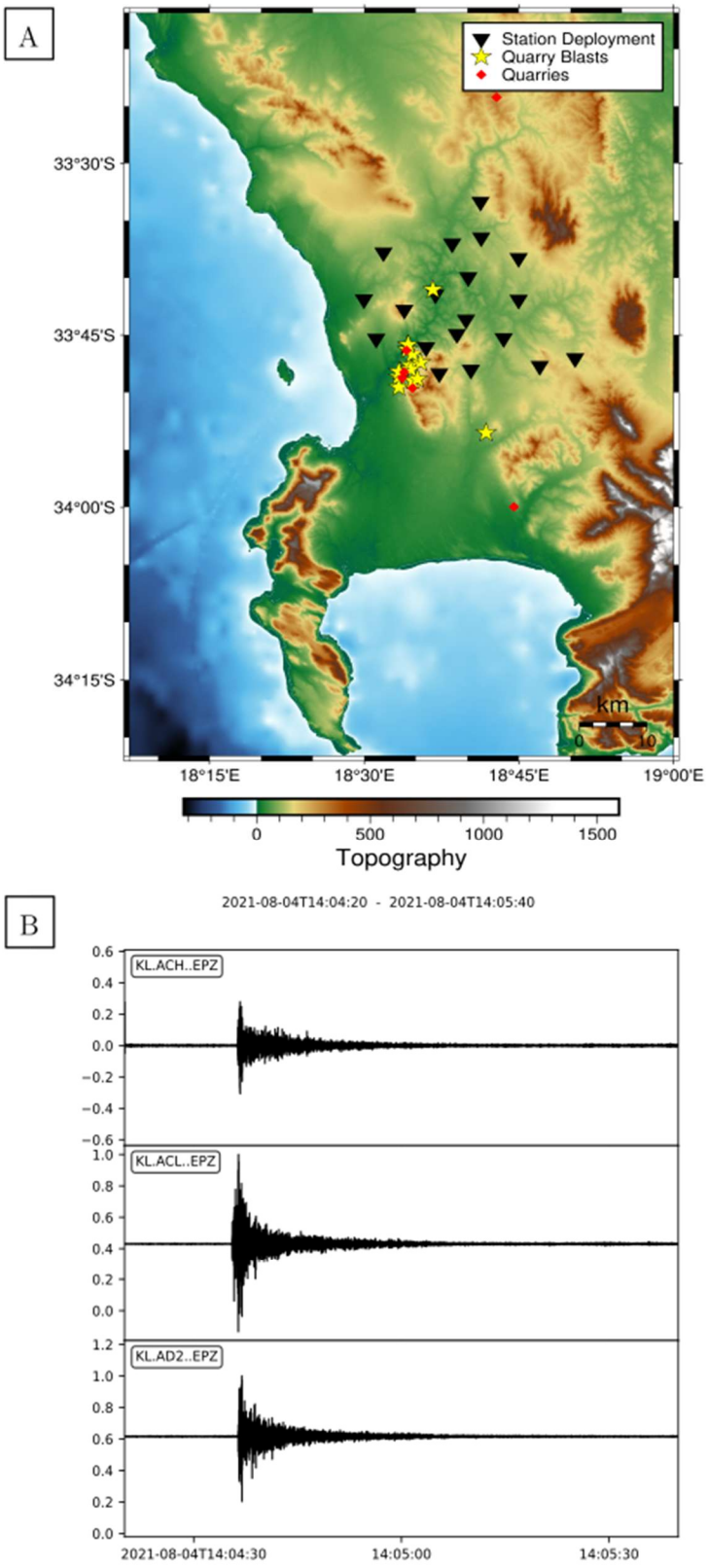


Figure 19: Figure 19A shows the locations of the manually identified quarry blasts (yellow stars) along with known quarry locations (red diamonds) relative to the station network. Figure 19B presents an example of the decreased S-wave amplitudes used to visually identify quarry blasts in the waveform data.

Table 5: Manually identified 28 earthquakes with the origin time, latitude and depth.

Seisan Origin Time	Latitude	Longitude	Depth (km)
2021/08/02 12:34:49	-33.319	18.472	0.7
2021/08/04 01:59:12	-33.664	18.346	6.9
2021/08/09 08:45:16	-32.327	17.213	0.2
2021/08/13 09:49:21	-33.755	18.673	2.1
2021/08/17 00:30:14	-33.6	17.424	6.2
2021/08/18 11:44:01	-33.612	18.564	1.8
2021/08/19 05:12:49	-33.055	17.24	0
2021/08/19 11:19:18	-33.699	18.725	0.3
2021/08/25 10:08:40	-33.686	18.62	0
2021/08/26 00:16:19	-33.679	18.797	0
2021/08/27 01:28:42	-33.465	18.2	5.5
2021/09/01 11:04:29	-32.836	17.413	10.6
2021/09/01 14:50:52	-33.751	18.646	2.9
2021/09/01 16:10:03	-34.352	17.53	11.3
2021/09/02 11:12:17	-33.739	18.739	4.9
2021/09/05 05:26:13	-33.714	18.585	0.7
2021/09/07 12:10:39	-32.991	18.366	0.1
2021/09/09 15:26:30	-33.821	18.571	2.2
2021/09/10 12:57:02	-33.569	18.702	0.9
2021/09/17 10:52:00	-33.679	18.65	2
2021/09/27 02:55:34	-33.474	17.398	12.4
2021/09/27 03:31:06	-33.192	17.328	12
2021/09/27 10:47:08	-33.671	18.723	0.3
2021/09/28 09:16:07	-33.545	18.704	0.6
2021/09/30 12:25:32	-33.473	18.157	0.3
2021/09/30 12:49:42	-33.602	18.655	0
2021/10/02 21:17:41	-33.695	18.85	11.8
2021/10/02 23:17:29	-33.725	18.476	0

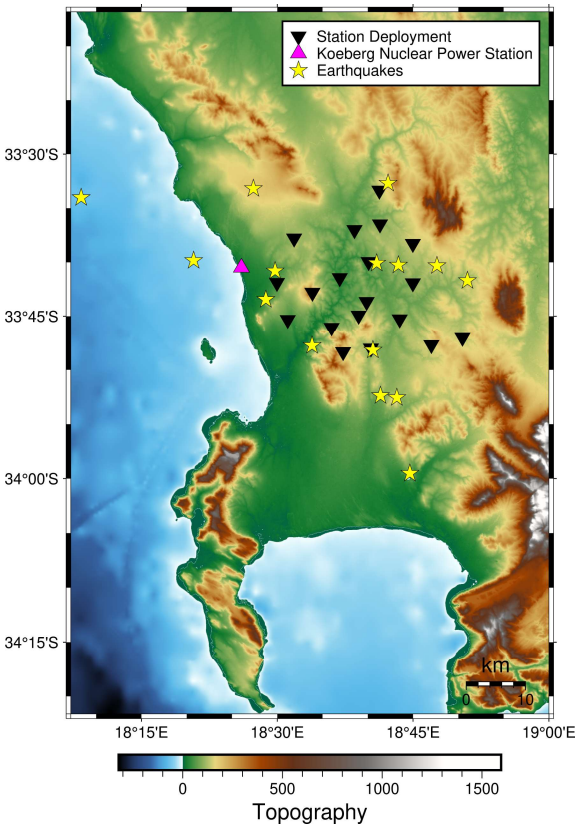


Figure 20: Epicenters of the visually identified earthquakes.

5.2. Short-Time Average to Long-Time Average Results

A total of 1,395 network triggers were detected by the Short-Time Average to Long-Time Average (STA/LTA) algorithm. Of these, only 41 were likely true positives, consisting of 15 quarry blasts and 26 earthquakes. The distinction between blasts and earthquakes was made as discussed in subsection 5.1, and therefore, the possibility of misidentifying an earthquake as a quarry blast, or vice versa, is also present in the STA/LTA results. Table 6 contains the catalog of blasts along with the stations that triggered a network alert. Similarly, Table 7 lists the identified earthquakes and the stations responsible for the network triggers. The hypocenter locations of these earthquakes are provided in subsequent subsections in the complete earthquake catalog.

Not all 1,395 network triggers underwent manual review to confirm if an event occurred at the corresponding time. Only events with comparable times to those identified in the manual review and machine learning algorithms (subsection 5.3) were examined. Therefore, it is possible that the STA/LTA algorithm could have identified additional potential events, but this would require a more extended review of waveforms for each network trigger.

Table 6: STA/LTA identified quarry blasts with the time of the blasts and the stations which detected the blast to result in a network trigger.

Network Trigger Time	STA/LTA Stations Triggered
2021/07/28 16:01:20	ACF, ACL, ACM, ACR, AD0, AD1, AD6, AD7
2021/07/29 14:31:50	ACF, ACL, ACM, ACR, AD0, AD1, AD6, AD7
2021/08/04 14:04:40	ACF, ACG, ACH, ACK, ACL, ACM, ACN, ACR, ACS, ACX, ACZ, AD0, AD1, AD2, AD5, AD6, AD7
2021/08/10 11:59:56	ACF, ACG, ACH, ACK, ACL, ACN, ACR, ACS, ACT, ACX, ACZ, AD0, AD2, AD5, AD6, AD7
2021/08/16 13:39:41	ACK, ACL, ACN, ACS, ACT, ACX, ACZ, AD2, AD5, AD6
2021/08/24 10:23:38	ACH, ACL, ACN, ACX, ACZ, AD0, AD1, AD2, AD5
2021/08/26 10:58:43	ACG, ACH, ACL, ACN, ACR, ACT, ACX, ACZ, AD1, AD2, AD5
2021/08/31 15:30:30	ACF, ACG, ACL, ACN, ACR, ACS, ACT, ACZ, AD0, AD1, AD2, AD5, AD7
2021/09/04 09:54:39	ACG, ACH, ACK, ACL, ACN, ACR, ACS, ACT, ACZ, AD0, AD1, AD2, AD5
2021/09/09 12:24:08	ACF, ACG, ACL, ACN, ACR, AD1, AD5
2021/09/14 11:43:41	ACF, ACG, ACH, ACK, ACL, ACN, ACR, ACS, ACT, ACX, ACZ, AD1, AD2, AD5, AD6, AD7
2021/09/15 15:53:58	ACG, ACK, ACL, ACN, ACT, ACX, ACZ, AD2, AD5
2021/09/20 16:09:05	ACG, ACN, ACR, ACS, ACX, ACZ, AD1, AD2, AD5, AD6
2021/09/29 11:19:24	ACF, ACL, ACN, AD2, AD5
2021/09/30 15:30:41	ACF, ACK, ACL, ACN, ACR, ACS, ACZ, AD1, AD2, AD5, AD6, AD7

Table 7: STA/LTA identified earthquakes and the stations which detected the event to result in a network trigger.

Network Trigger Time	STA/LTA Stations Triggered
2021/07/30 10:33:32	ACF, ACL, ACM, ACR, AD0, AD1, AD6, AD7
2021/08/02 12:34:54	ACF, ACL, ACM, ACR, ACS, ACX, AD7
2021/08/04 01:59:18	ACF, ACH, ACL, ACN, ACR, ACS, AD1, AD2, AD5
2021/08/04 15:28:51	ACF, ACG, ACH, ACK, ACL, ACM, ACN, ACR, ACS, ACT, ACX, ACZ, AD1, AD2, AD5, AD6, AD7
2021/08/06 11:39:58	ACF, ACH, ACL, ACM, ACN, ACR, ACS, ACT, ACX, ACZ, AD2, AD5
2021/08/09 08:45:39	ACN, ACR, ACX, ACZ, AD0, AD2, AD6
2021/08/13 09:52:22	ACH, ACN, ACR, ACS, ACT, AD5, AD6
2021/08/17 00:30:39	ACK, ACL, ACN, ACS, ACT, ACX, ACZ, AD2, AD5, AD6
2021/08/18 11:44:22	ACG, ACL, ACN, ACR, ACX, AD0
2021/08/19 05:13:05	ACG, ACL, ACN, ACR, ACS, AD2, AD5

2021/08/22 03:45:39	ACG, ACH, ACL, ACN, ACS, ACT, ACX, ACZ, AD0, AD2, AD5, AD7
2021/08/31 16:20:19	ACF, ACL, ACN, ACR, ACZ, AD1, AD2, AD7
2021/09/01 16:10:08	ACF, ACG, ACL, ACX, AD6
2021/09/02 11:12:18	ACF, ACH, ACK, ACL, ACN, ACR, ACT, ACZ, AD0, AD2, AD5
2021/09/05 05:26:31	ACF, ACG, ACL, ACN, ACR, AD1, AD2, AD5, AD7
2021/09/07 12:10:58	ACF, ACG, ACH, ACL, ACN, ACR, AD7
2021/09/09 15:26:34	ACG, ACH, ACK, ACL, ACN, ACR, AD1, AD6
2021/09/10 12:57:12	ACF, ACG, ACK, ACL, ACN, ACR, ACS, ACZ, AD1, AD2, AD5, AD7
2021/09/17 10:52:55	ACF, ACK, ACL, ACN, ACR, ACS, AD2, AD5
2021/09/27 02:55:55	ACF, ACL, ACR, ACS, AD2, AD5, AD7
2021/09/27 03:31:30	ACF, ACL, ACN, ACR, ACS, ACX, AD1, AD2, AD5, AD7
2021/09/28 09:16:28	ACL, ACN, ACS, ACX, AD2, AD5, AD7
2021/09/30 12:26:51	ACF, ACN, ACR, AD1, AD7
2021/09/30 12:49:56	ACL, ACN, ACR, AD2, AD7
2021/10/02 21:17:47	ACL, ACN, AD2, AD5, AD7
2021/10/02 23:17:31	ACL, ACN, AD1, AD5

5.3. Machine Learning Results

The overall total number of detections identified by the ML algorithm is 39. After visually reviewing all 39 events, no clear false detections were found among them. The ML algorithm identified 15 quarry blasts indicated in the blast catalog given in Table 8. The output map from Seisbench shown in Figure 21, includes station locations, quarry blast epicenters, and the locations of quarries that were active during the deployment.

Table 8: Catalog of the 15 quarry blasts identified by ML algorithm. Nine parameters are associated with each catalog output from the ML algorithm: Sigma Time (root mean square error between observed and predicted arrivals), Gamma Score (sum of probabilities used in event association), Total Picks , P-picks (count of p-picks out of total picks), S-picks (count of s-picks out of total picks), and hypocenter coordinates (x for easting, y for northing, z for depth of the event).

Origin Time	Sigma Time	Gamma Score	Total Picks	P-picks	S-picks	x(km)	y(km)	z(km)
2021-07-28T16:01:16.816	0.09	9	9	8	1	-282.31	6229.06	0.31
2021-07-29T14:31:43.547	0.11	8	8	8	0	-281.25	6222.5	0.12
2021-08-04T14:04:34.636	0.16	19	19	17	2	-283.09	6226.95	0.16
2021-08-10T11:59:50.559	0.12	15	15	14	1	-280.54	6225.62	0.27
2021-08-16T13:39:35.392	0.21	11.84	12	11	1	-264.82	6207.59	0.67
2021-08-24T10:23:32.315	0.3	9	9	6	3	-283.82	6223.04	0
2021-08-26T10:58:41.229	0.12	15	15	12	3	-282.85	6226.31	0.3
2021-08-31T15:30:26.206	0.11	15	15	13	2	-282.47	6226.07	0.25
2021-09-04T09:54:36.030	0.17	15.96	16	13	3	-282.81	6231.58	0.8
2021-09-09T12:24:03.365	0.25	7	7	7	0	-268.63	6271	1.25
2021-09-14T11:43:37.482	0.11	19	19	15	4	-279.95	6226.5	0.17
2021-09-15T15:53:54.443	0.12	15	15	12	3	-282.57	6230.45	0.28
2021-09-20T16:08:59.371	0.12	12.95	13	9	4	-282.88	6231.36	0.19
2021-09-29T11:19:19.796	0.1	8	8	7	1	-263.55	6209.12	0.3
2021-09-30T15:30:37.467	0.11	13	13	13	0	-282.85	6226.47	0.25

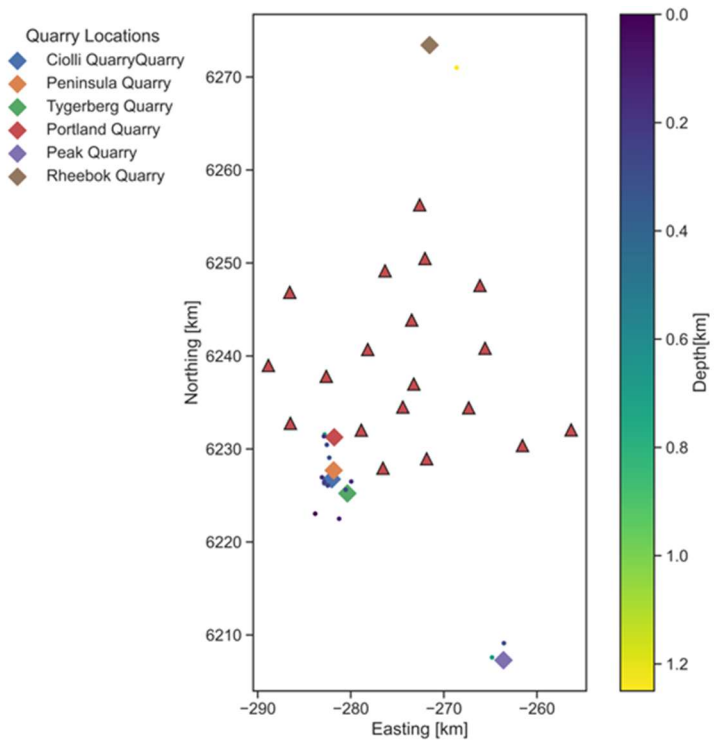


Figure 21: Map output of the epicenters from the ML algorithm along with added quarry locations.

The ML algorithm identified 24 earthquakes whose output catalog is shown in Table 9. Figure 22A is an output map from Seisbench showing all the events and their relation to the station deployment. Figure 22B shows the locations of all ML-identified events, but with their x (km) and y (km) (i.e., Eastings and Northings) transformed back into latitude and longitude to view in relation to the KNPS's location.

Table 9: Catalog of the 24 earthquakes identified by the machine algorithm.

Origin Time	Sigma Time	Gamma Score	Total Picks	P-picks	S-picks	x(km)	y(km)	z(km)
2021-07-30T10:33:26.183	0.07	6	6	6	0	-274.1	6228.46	0.01
2021-07-31T22:45:29.039	0.56	6	6	6	0	-259.25	6251.67	0.06
2021-08-02T12:34:50.211	0.19	9	9	9	0	-294.12	6271	0
2021-08-04T01:59:12.641	0.11	19.99	20	8	12	-301	6240.8	10.37
2021-08-04T15:28:45.731	0.14	17	17	16	1	-283.15	6225.7	0.18
2021-08-06T11:39:54.183	0.25	8	8	8	0	-276.97	6266.51	0
2021-08-09T08:45:35.954	0.81	7	7	6	1	-272.41	6244.6	9.8
2021-08-13T09:52:16.514	0.69	6	6	6	0	-279.57	6271	0.1
2021-08-19T05:13:04.460	0.48	8	8	7	1	-286.4	6243.01	0.28

2021-08-22T03:45:36.603	0.16	14	14	14	0	-254	6262.58	0.04
2021-08-27T01:28:39.695	0.21	18	18	18	0	-301	6250.14	8.92
2021-08-31T16:20:13.740	0.2	6	6	6	0	-285.49	6271	0.18
2021-09-01T11:04:27.941	0.19	12	12	12	0	-301	6259.44	7.97
2021-09-01T14:50:50.620	0.19	6	6	6	0	-270.45	6223.66	0.12
2021-09-01T16:10:04.791	0.11	16	16	16	0	-301	6217.85	16
2021-09-02T11:12:14.739	0.09	12	12	10	2	-263.99	6207.36	0.44
2021-09-05T05:26:26.984	0.19	8	8	8	0	-301	6252.97	16
2021-09-07T12:10:52.137	0.21	7	7	7	0	-269.52	6271	11.19
2021-09-09T15:26:29.892	0.13	9	9	7	2	-283.18	6225.16	0.27
2021-09-10T12:57:06.118	0.2	10	10	10	0	-262.9	6271	0.81
2021-09-17T10:52:50.468	0.22	7	7	7	0	-270.8	6271	0
2021-09-21T02:26:47.602	0.34	7	7	7	0	-262.79	6203.48	16
2021-09-28T09:16:26.558	0.3	8	8	6	2	-276.56	6269.01	1.81
2021-10-02T21:17:39.981	0.2	7	7	7	0	-254	6261.18	0.74

The events at the edges of Figures 22A and 22B are algorithm artifacts known as boundary effects. These events have hypocenters that fall outside the defined Northing and Easting bounds in our algorithm, resulting in them to be placed at the maximum bound values, forming a square shape. In Figure 22A, these boundary-affected events are indicated by red rectangles on the output map from the Seisbench algorithm and in Figure 22B these events are shown by red stars. These events are included in the catalog because their origin times are correct. The analysis in subsections 5.4 and 6.1 focuses solely on origin times when evaluating the usefulness of each method, without comparing the accuracy of machine learning-generated event locations to those done with HYPOCENTER.

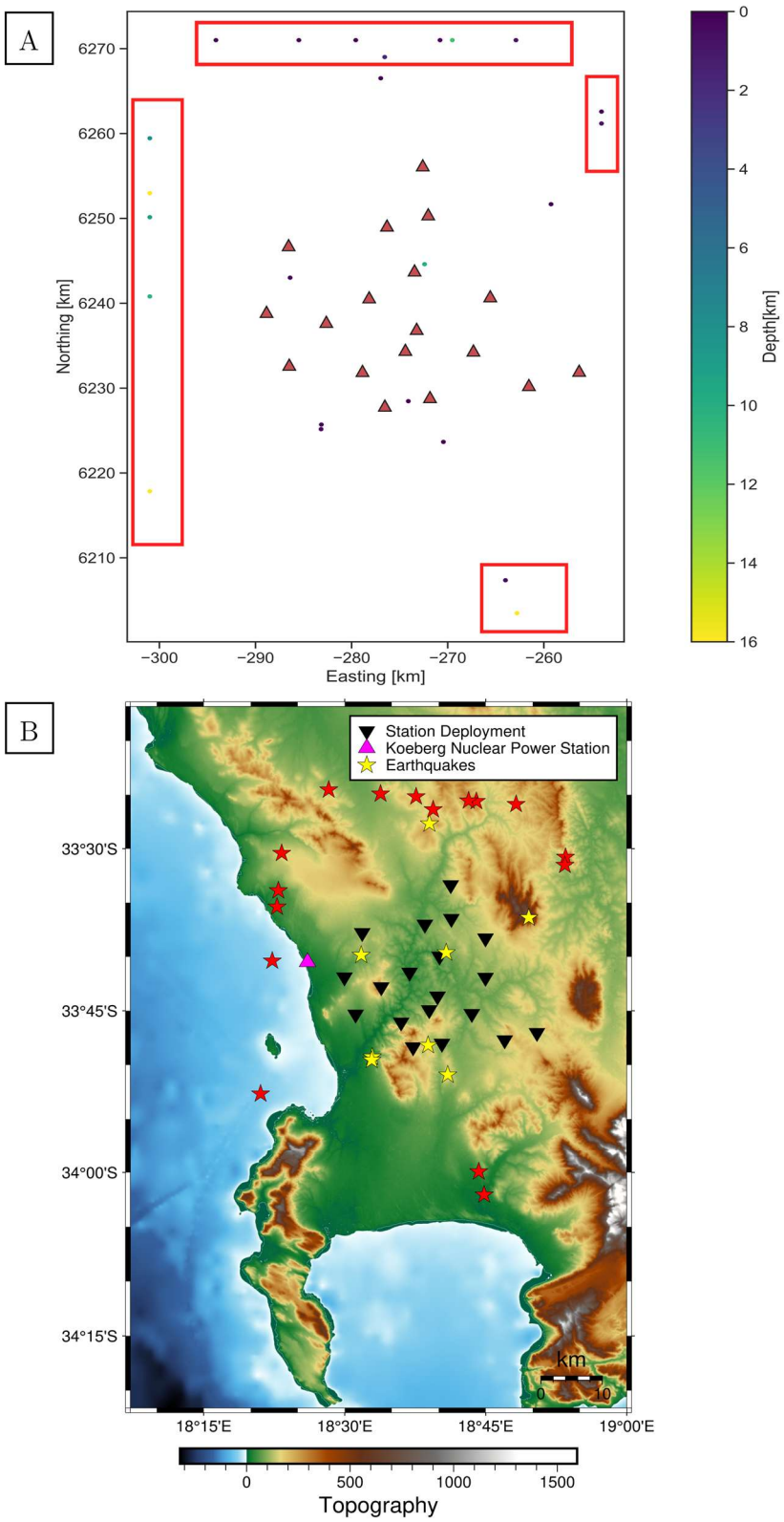


Figure 22: Figure 22A shows an output map from Seisbench with all the epicenters for each earthquake given in Northing (y(km)) and Eastings (x(km)). Figure 22B indicates the same information, but in latitude and longitude with the location of the KNPS included.

5.4. Detection Statistics: Visual Identification, STA/LTA and ML

The complete catalog includes a total of 51 events, comprising 35 recorded earthquakes and 16 quarry blasts. Table 10 presents the origin times and identification methods for each event, with missed events from each method highlighted in red.

Machine learning has accurately recognized 39 occurrences (true positives) with no false detections. This method missed 12 events (false negatives) as shown in Table 10 and Figure 23. The false negative rate (miss rate) is the likelihood that an event will be missed by a specific method, computed by dividing the total number of false negatives by the sum of false negatives and true positives. The machine learning method has a false negative rate of 24%, indicating a true positive rate of 76%, which means that 76% of all true positives in the catalog were detected using this method. Precision is the final metric to consider when determining the effectiveness and accuracy of each method. Precision is defined as the ratio of true positives to the sum of true and false positives. High precision indicates a greater proportion of true positives across all detected events than false positives. The machine learning algorithm's precision is 100%.

The total number of true positives identified by the STA/LTA algorithm is 41, accompanied by 1,354 potential events that were not manually confirmed. Although most of these unconfirmed events are likely false positives, it is not possible to definitively rule out the possibility that some smaller or less distinct events might be present among them. The false negative rate is 20%, suggesting that the method may miss approximately one in five true positive events in the dataset. The precision of 3% indicates that most events identified by this method require careful human evaluation to distinguish genuine signals from noise. This outcome is consistent with the STA/LTA approach, which is primarily designed to flag potential events for further human-driven scrutiny and differentiation.

Manual identification resulted in the identification of 41 out of the total of 51 events in the dataset. This gives this method a true positive number of 41 and a false negative number of 10 as indicated in Table 10 and Figure 23. This method did not identify any false detections resulting in 0 false positives. The false positive rate for this method is 20 % with the precision at 100%.

Table 10: Origin time calculated with SEISAN versus, Seisbench origin time and the Network Trigger time given by the STA/LTA algorithm.

SEISAN Origin Time	Seisbench Origin Time	Network Trigger Time	Manual Detection
2021-07-28 16:01:16	2021-07-28T16:01:16.816	2021-07-28 16:01:20	Event missed
2021-07-29 14:31:44	2021-07-29T14:31:43.547	2021-07-29 14:31:50	Event missed
2021-07-30 10:33:19	2021-07-30T10:33:26.183	2021-07-30 10:33:32	Event missed
2021-07-31 22:45:28	2021-07-31T22:45:29.039	Event missed by STA/LTA	Event missed
2021-08-02 12:34:49	2021-08-02T12:34:50.211	2021-08-02 12:34:54	2021-08-02 12:34:49
2021-08-04 01:59:12	2021-08-04T01:59:12.641	2021-08-04 01:59:18	2021-08-04 01:59:12
2021-08-04 14:04:34	2021-08-04T14:04:34.636	2021-08-04 14:04:40	Event missed
2021-08-04 15:28:45	2021-08-04T15:28:45.731	2021-08-04 15:28:51	Event missed
2021-08-06 11:39:55	2021-08-06T11:39:54.183	2021-08-06 11:39:58	Event missed
2021-08-09 08:45:16	2021-08-09T08:45:35.954	2021-08-09 08:45:39	2021-08-09 08:45:16
2021-08-10 11:59:50	2021-08-10T11:59:50.559	2021-08-10 11:59:56	2021-08-10 11:59:50
2021-08-13 09:49:21	2021-08-13T09:52:16.514	2021-08-13 09:52:22	2021-08-13 09:49:21
2021-08-16 13:39:35	2021-08-16T13:39:35.392	2021-08-16 13:39:41	2021-08-16 13:39:35
2021-08-17 00:30:14	Event missed by PhaseNet	2021-08-17 00:30:39	2021-08-17 00:30:14
2021-08-18 11:44:01	Event missed by PhaseNet	2021-08-18 11:44:22	2021-08-18 11:44:01
2021-08-19 05:12:48	2021-08-19T05:13:04.460	2021-08-19 05:13:05	2021-08-19 05:12:48
2021-08-19 11:19:18	Event missed by PhaseNet	Event missed by STA/LTA	2021-08-19 11:19:18
2021-08-22 00:52:26	Event missed by PhaseNet	Event missed by STA/LTA	2021-08-22 00:52:26
2021-08-22 03:45:27	2021-08-22T03:45:36.603	2021-08-22 03:45:39	Event missed
2021-08-24 10:23:33	2021-08-24T10:23:32.315	2021-08-24 10:23:38	2021-08-24 10:23:33
2021-08-25 10:08:40	Event missed by PhaseNet	Event missed by STA/LTA	2021-08-25 10:08:40
2021-08-26 00:16:18	Event missed by PhaseNet	Event missed by STA/LTA	2021-08-26 00:16:18
2021-08-26 10:58:40	2021-08-26T10:58:41.229	2021-08-26 10:58:43	2021-08-26 10:58:40
2021-08-27 01:28:42	2021-08-27T01:28:39.695	Event missed by STA/LTA	2021-08-27 01:28:42
2021-08-31 15:30:26	2021-08-31T15:30:26.206	2021-08-31 15:30:30	2021-08-31 15:30:26
2021-08-31 16:20:18	2021-08-31T16:20:13.740	2021-08-31 16:20:19	Event missed
2021-09-01 11:04:29	2021-09-01T11:04:27.941	Event missed by STA/LTA	2021-09-01 11:04:29
2021-09-01 14:50:51	2021-09-01T14:50:50.620	Event missed by STA/LTA	2021-09-01 14:50:51
2021-09-01 16:10:02	2021-09-01T16:10:04.791	2021-09-01 16:10:08	2021-09-01 16:10:02
2021-09-02 11:12:17	2021-09-02T11:12:14.739	2021-09-02 11:12:18	2021-09-02 11:12:17

2021-09-04 09:54:36	2021-09-04T09:54:36.030	2021-09-04 09:54:39	2021-09-04 09:54:36
2021-09-05 05:26:13	2021-09-05T05:26:26.984	2021-09-05 05:26:31	2021-09-05 05:26:13
2021-09-07 12:10:39	2021-09-07T12:10:52.137	2021-09-07 12:10:58	2021-09-07 12:10:39
2021-09-09 12:23:58	2021-09-09T12:24:03.365	2021-09-09 12:24:08	2021-09-09 12:23:58
2021-09-09 15:26:30	2021-09-09T15:26:29.892	2021-09-09 15:26:34	2021-09-09 15:26:30
2021-09-10 12:57:02	2021-09-10T12:57:06.118	2021-09-10 12:57:12	2021-09-10 12:57:02
2021-09-14 11:43:37	2021-09-14T11:43:37.482	2021-09-14 11:43:41	2021-09-14 11:43:37
2021-09-15 15:53:54	2021-09-15T15:53:54.443	2021-09-15 15:53:58	2021-09-15 15:53:54
2021-09-17 10:52:00	2021-09-17T10:52:50.468	2021-09-17 10:52:55	2021-09-17 10:52:00
2021-09-20 16:08:59	2021-09-20T16:08:59.371	2021-09-20 16:09:05	2021-09-20 16:08:59
2021-09-21 02:25:47	2021-09-21T02:26:47.602	Event missed by STA/LTA	Event missed
2021-09-27 02:55:34	Event missed by PhaseNet	2021-09-27 02:55:55	2021-09-27 02:55:34
2021-09-27 03:31:06	Event missed by PhaseNet	2021-09-27 03:31:30	2021-09-27 03:31:06
2021-09-27 10:47:08	Event missed by PhaseNet	Event missed by STA/LTA	2021-09-27 10:47:08
2021-09-28 09:16:06	2021-09-28T09:16:26.558	2021-09-28 09:16:28	2021-09-28 09:16:06
2021-09-29 11:19:18	2021-09-29T11:19:19.796	2021-09-29 11:19:24	2021-09-29 11:19:18
2021-09-30 12:25:32	Event missed by PhaseNet	2021-09-30 12:26:51	2021-09-30 12:25:32
2021-09-30 12:49:42	Event missed by PhaseNet	2021-09-30 12:49:56	2021-09-30 12:49:42
2021-09-30 15:30:37	2021-09-30T15:30:37.467	2021-09-30 15:30:41	2021-09-30 15:30:37
2021-10-02 21:17:40	2021-10-02T21:17:39.981	2021-10-02 21:17:47	2021-10-02 21:17:40
2021-10-02 23:17:29	Event missed by PhaseNet	2021-10-02 23:17:31	2021-10-02 23:17:29

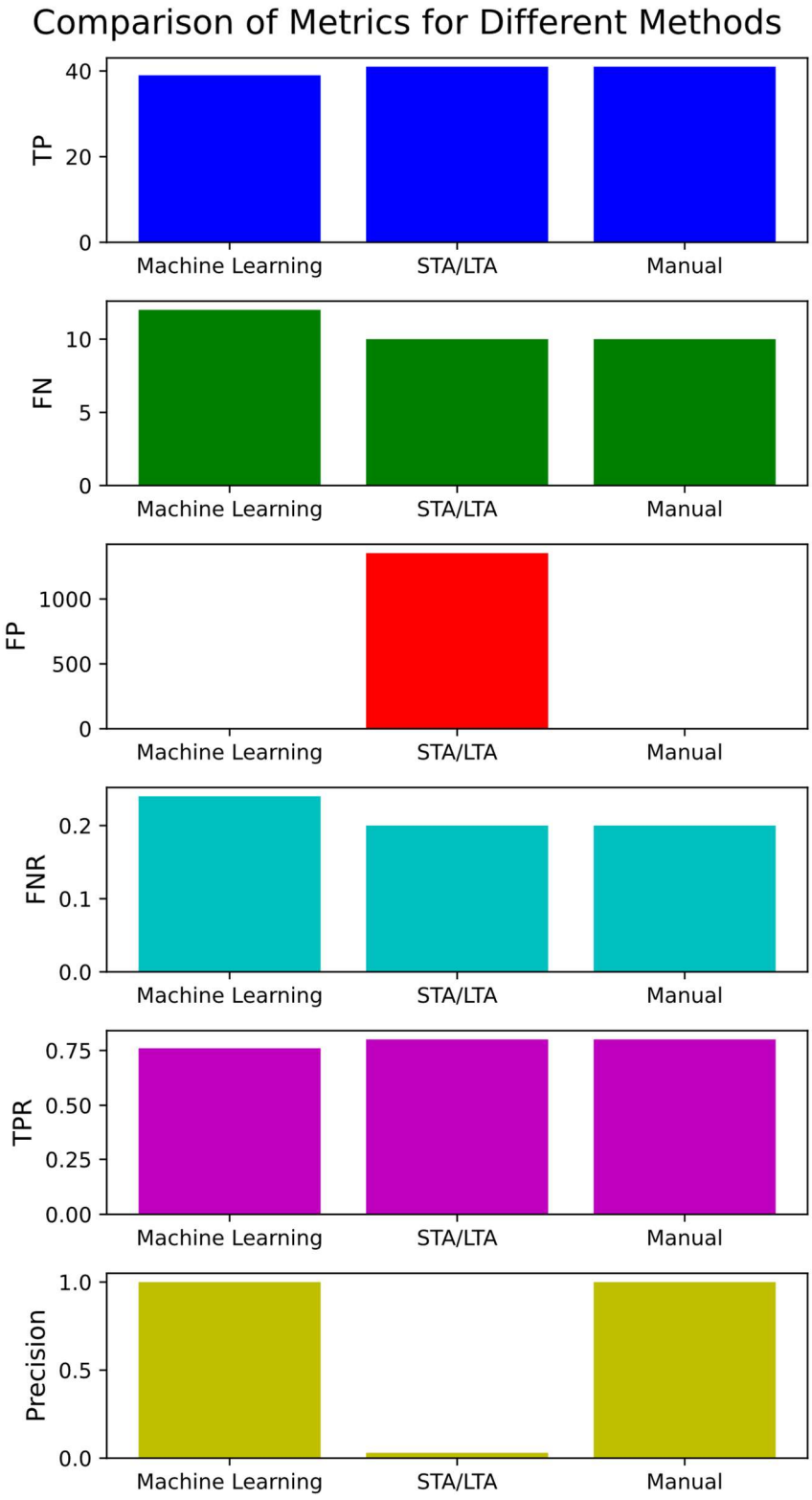


Figure 23: Detection statistics for various methods of phase identification. True Positive (TP) events are those that have been correctly detected. The number of false detections of a specific method is indicated by False Positives (FP). False Negative (FN) represents the number of missed detections by a specific method. The True Positive Rate (TPR) is calculated by dividing the TP by the sum of the TP and FN. The False Negative Rate (FNR) is calculated by dividing

5.5. Epicenters of Final Blast Catalog

The horizontal uncertainty ellipses depicted in Figures 24 and 25 (subsection 5.6) represent the 90% confidence regions, and are an output from HYPOCENTER. The morphology and orientation of uncertainty ellipses provide critical insights into the geometric constraints governing location determination.

For earthquake events within the station network (Figure 25B), the ellipses manifest approximately circular geometries with horizontal uncertainties typically have 2-5 km long minor axis and <10 km long major axis, reflecting optimal azimuthal coverage and robust P- and S-wave arrival constraints within the network.

In contrast, offshore events exhibit strongly anisotropic uncertainty ellipses characterized by two predominant elliptical patterns: (1) radially elongated ellipses (error ellipse numbers 1, 2, and 6 in Figure 25A) pointing toward or away from the network, occurring when backazimuth is well-constrained by at least 6 P-arrivals but distance remains poorly resolved due to limited S-wave observations; and (2) obliquely oriented narrow ellipses (error ellipse numbers 3, 4, and 5 in Figure 25A) that point neither directly toward the array nor tangentially, resulted from fewer than average picks constrained using picks from 4 stations. Overall, picking these events has proven to be extremely difficult to accurately constrain their location.

Due to these substantial uncertainties, offshore events should be interpreted with considerable caution. While some potential interpretations of offshore seismicity patterns are provided in Section 6, these should be considered preliminary hypotheses rather than definitive conclusions, given the inherent location uncertainties described above.

Notably, an additional source of uncertainty for offshore events is not accounted for in the uncertainty ellipses—specifically, uncertainty in the velocity model. Assuming velocities that are too slow (e.g., because the model does not incorporate faster layers below 12 km) would imply that the true epicenters are farther from the array, and vice versa. The 12 km cutoff depth of the velocity model also posed challenges.

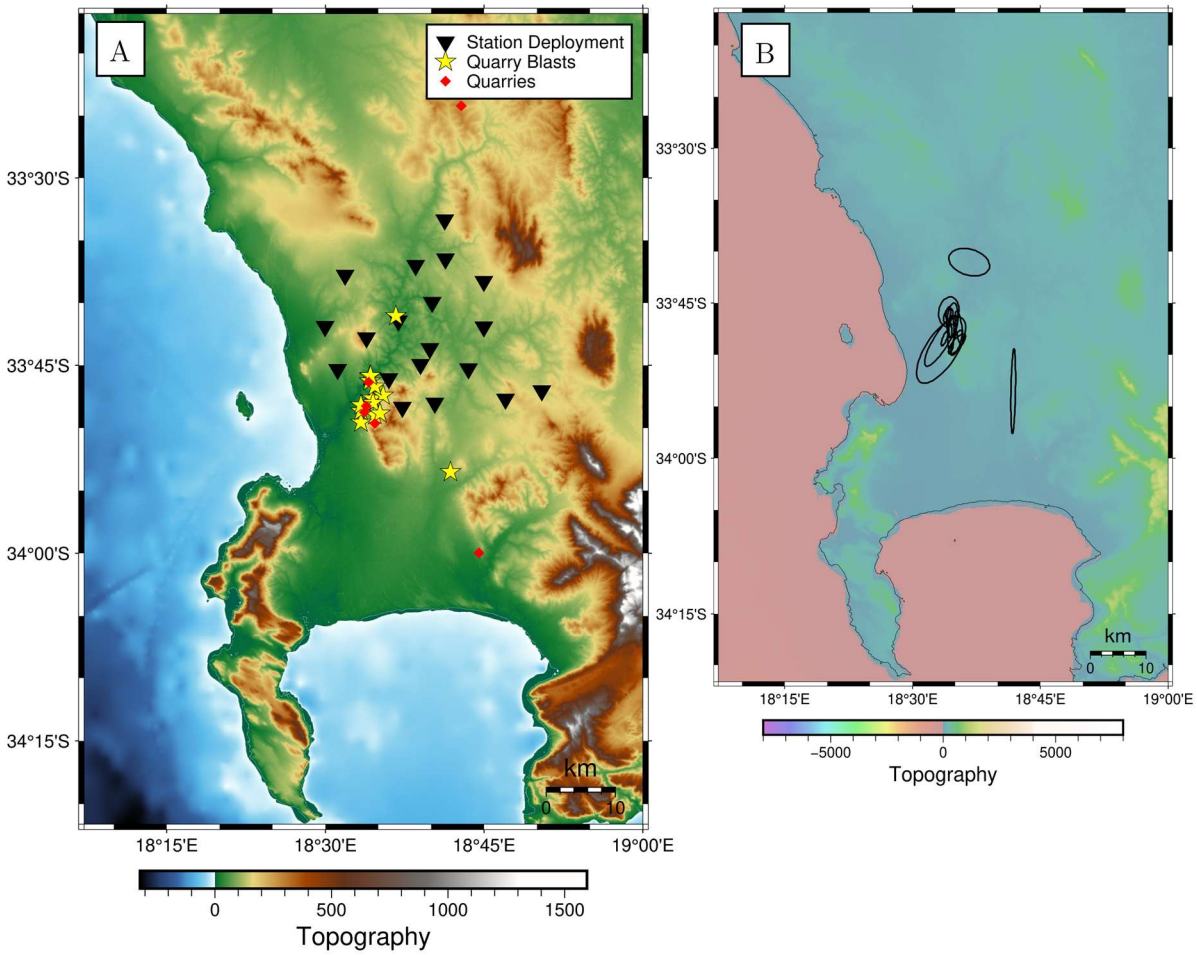


Figure 24: Figure 24A shows the locations of all 16 quarry blasts that occurred during the deployments with Figure 24B showing the error associated with calculated locations. The error increases away from the center of the array-deployment.

5.6. Epicenters of Final Earthquake Catalog

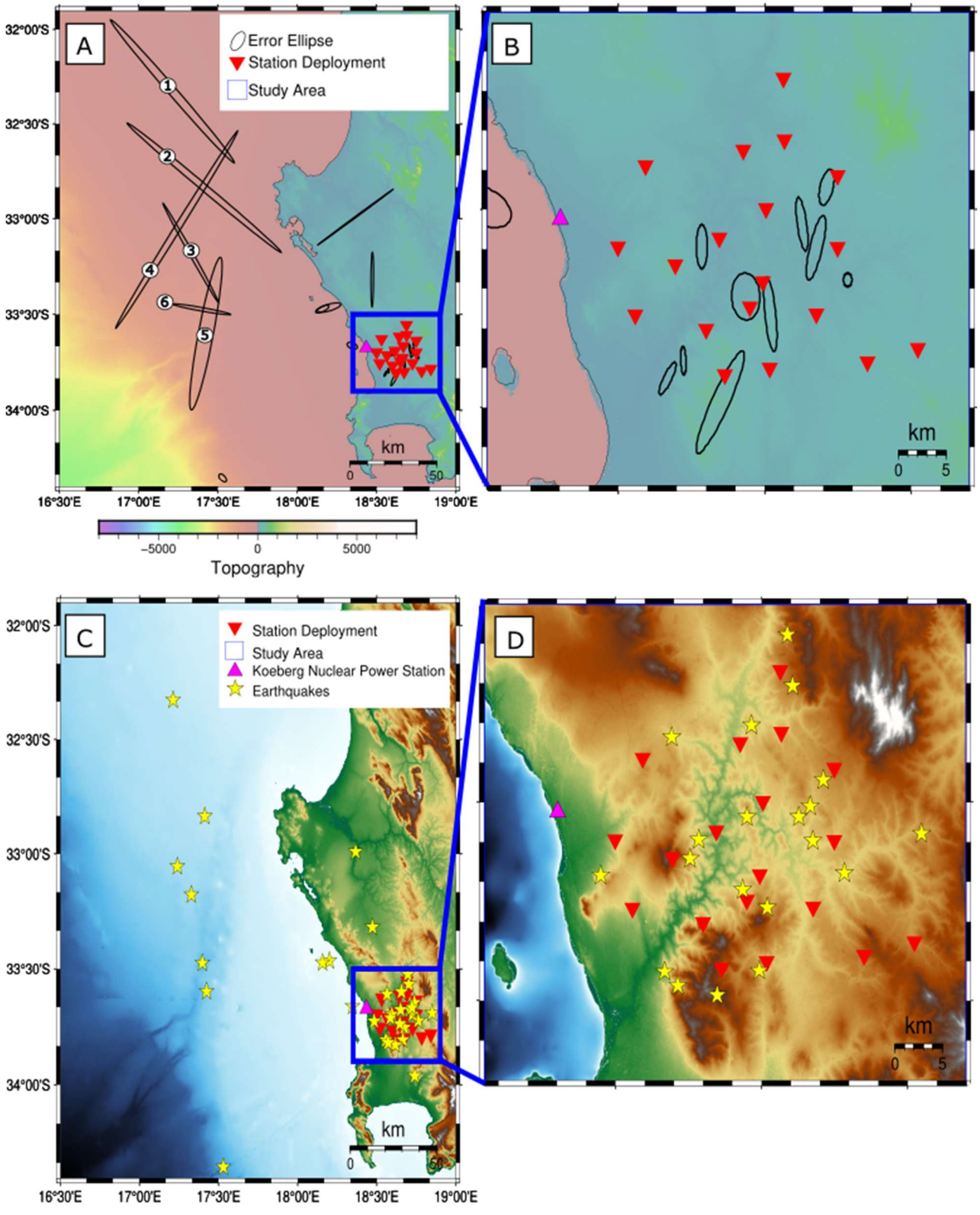


Figure 25: Figure 25A illustrates the uncertainty of all 35 identified earthquakes, with offshore events numbered to emphasize the significant uncertainties, as discussed in the text. Figure 25B provides a closer view focusing exclusively on the station network. Figure 25C shows the epicenters along with the Koeberg Nuclear Power Station (KNPS), while Figure25D offers a zoomed-in view concentrating on the station network.

6. Discussion

The results indicate that within the 2.8 months of recordings, the region experienced 35 micro-earthquakes, none of which were reported as felt. In the context of seismically active versus quiet areas, Figure 26A and 26B indicate heightened seismicity within the greater Cape Town area and the KNPS, specifically within the seismic station network. It is important to mention that the detection limit within the network is smaller (i.e., small magnitude events are detectable within and near the network) than for regions further away from the network centroid. The increased seismic activity within the station array is primarily located between the Colenso Fault and the proposed Milnerton Fault (Figure 26B). Notably, the epicenters are more densely concentrated to the north of the array, with seven micro-seismic events occurring along the Colenso Fault, providing evidence for it being potentially active. These events align with two previously recorded epicenters from 2020: one on the 26th of September with a magnitude of 2.9 in the north and another on the 27th of September with a magnitude of 2.7 in the south, both documented by USGS and CGS (as indicated by the yellow encirclements in Figure 26B). It is important to note that these are not four separate events in Figure 26B; the two in the north represent the same event, and the two in the south are also the same, but each was located by different agencies (i.e., the CGS and USGS), highlighting the discrepancies in identifying reliable epicenters for this region.

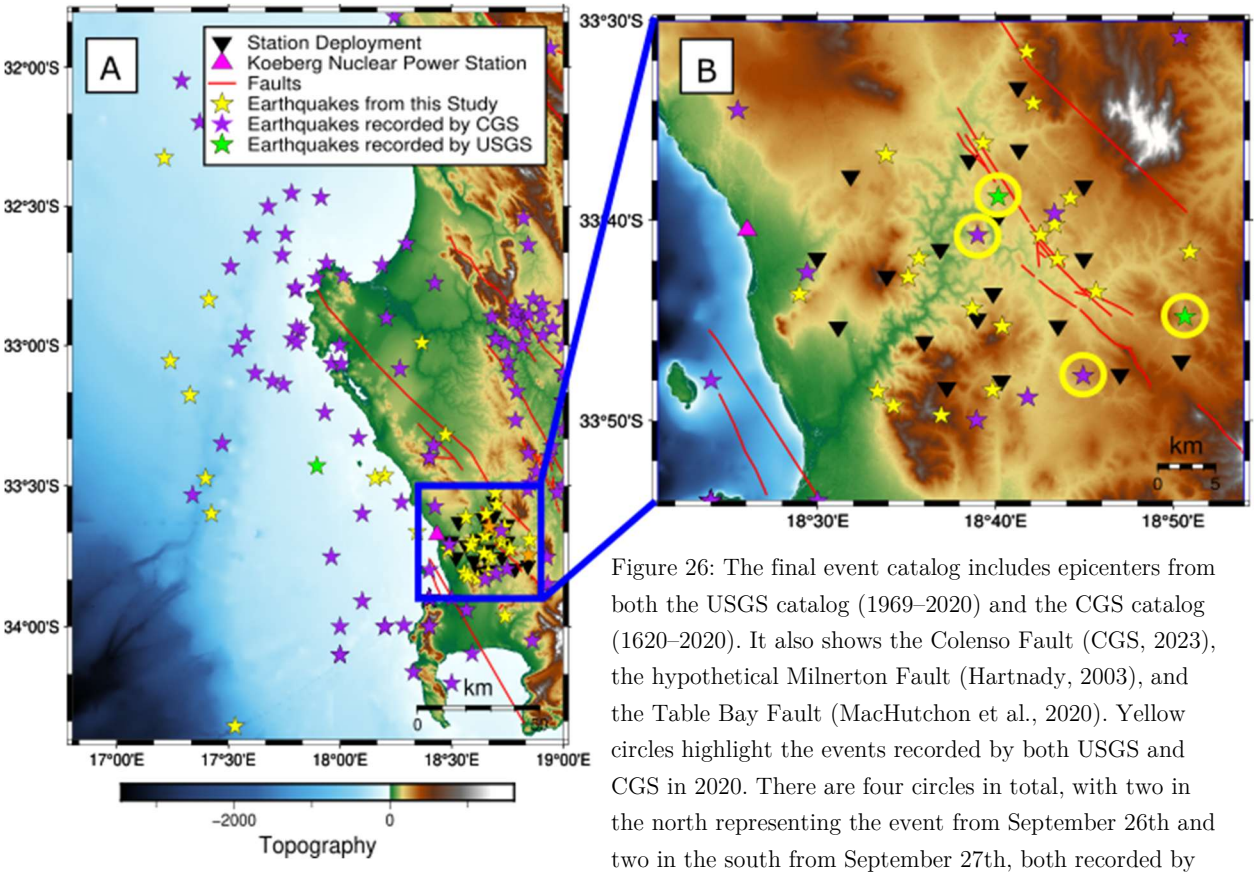


Figure 26: The final event catalog includes epicenters from both the USGS catalog (1969–2020) and the CGS catalog (1620–2020). It also shows the Colenso Fault (CGS, 2023), the hypothetical Milnerton Fault (Hartnady, 2003), and the Table Bay Fault (MacHutchon et al., 2020). Yellow circles highlight the events recorded by both USGS and CGS in 2020. There are four circles in total, with two in the north representing the event from September 26th and two in the south from September 27th, both recorded by USGS and CGS.

In the case of offshore events detected in this dataset, the associated uncertainty is considerable (Figure 25A); nonetheless, they appear to agree with previous events reported by the CGS (2020) (Figures 26A). An interesting observation is that some of the offshore events (Figure 26A) seem to align with the Cape Canyon (Wigley and Compton, 2005), a submarine canyon carved into the continental shelf. However, due to the large uncertainty of the epicenters (Figure 25A), no definitive conclusion can be drawn. Nevertheless, this alignment provides a potential avenue for future research.

Additionally, no direct seismic events from our results correlate with the Table Bay Fault (Figures 2C and 26), a potential fault inferred from bedding truncation along a large northwest-trending linear bathymetric low that coincides with a weak magnetic anomaly (MacHutchon et al., 2020). A study by Stamatakos et al. (2024) suggests that the fault is potentially inactive, as it has likely been inactive since the Early Cretaceous, since no offsets are observed in the dykes that intruded the bedrock during that time.

Due to the absence of national or international network stations near the two 2020 epicenters (Figures 3A and 26B), the uncertainties surrounding their hypocenter locations are likely to be substantial. Furthermore, the CGS (2020) catalog exhibits a bias toward regions like Cape Town, which have more comprehensive historical records, as it includes both written historical and instrumental events. Historical events were documented in the areas immediately surrounding Cape Town, which was more densely populated at the time. In contrast, regions north of the Colenso fault may have little to no written historical records, resulting in an underrepresentation of past occurrences in those areas. This raises the possibility that seismic events may have occurred in these undocumented regions but were only noted in more monitored and densely populated areas.

Stamatakos et al. (2024) propose that the Colenso Fault has been inactive since the Miocene. Supporting this hypothesis, Claasen et al. (2024) conducted marine terrace investigations near Paternoster on South Africa's West Coast and found no vertical offsets in the calcrete caps of the Colenso Fault. Furthermore, Stamatakos et al. (2024) reference micro-seismic studies by Mulabisana (2023) and onshore fault mapping by Coppersmith et al. (2024), both of which show no recent geomorphological evidence of active faulting. Mulabisana (2023) involved a network of seven stations along the Colenso Faults to investigate seismic activity over six months, yielding five micro-seismic events. However, the low signal-to-noise ratio of the data, coupled with large uncertainties in locations, resulted in the omission of focal mechanism calculations and depth evaluations. Consequently, Stamatakos et al. (2024) concluded that this study could not be considered a reliable indicator of seismic activity for this fault system, given the waveform quality. The 1969 Ceres-Tulbagh earthquake (Krüger and Scherbaum, 2014; Smit et al., 2015), the nearest seismic event with a well-defined focal mechanism, occurred on a sinistral strike-slip fault. As a result, earthquakes in this area are unlikely to cause large vertical offsets. Additionally, geomorphological indicators of moderate-magnitude strike-slip earthquakes can be difficult to

detect. Therefore, the lack of visible evidence does not necessarily mean that seismic activity has been absent.

One of the key issues addressed in Chapter 1 is the large uncertainty regarding the locations of the 1809 and 1811 events. Hartnady (2003) used historical accounts as part of his evidence to propose a Milnerton fault origin for these events. Our results (Figure 26), indicate no evidence of heightened seismicity on the Milnerton fault.

Based on the heightened seismicity shown in Figure 26B and the findings of Stein and Liu (2009) that enhanced seismicity in stable continental regions frequently accompanies large historical earthquakes, it is possible that the 1809 and 1811 events took place near the region of the Colenso Fault. But this evidence by itself is insufficient to prove this beyond a reasonable doubt. Von Buchenroder (1830) accounts support a Milnerton Fault hypocenter, including seismic liquefaction. Seismic liquefaction could likely also occur with a Colenso Fault hypocenter, particularly if the event exceeds a magnitude of 6.3. Seismic liquefaction as outlined by Sekac et al., (2016) depends on various factors such as the local geology and the source parameters of the earthquake. Evaluation of the geology and its liquefaction potential depends on the consolidation status of the rocks (Sekac et al, 2016). The relative grade of consolidation is inversely proportional to the level of amplification with unconsolidated sediments resulting in the highest amplification of seismic waves and thus in turn increases the probability of liquefaction (Sekac et al., 2016). In Figure 3B, the closest branch of the Colenso Fault in the study area terminates in the deformed lower greenschist facies Malmesbury group (MG) (Kisters and Belcher, 2018; Kisters et al., 2002; Rozendaal et al., 1999). The proposed Milnerton Fault (and the surrounding Milnerton) region lies in a region covered with consolidated Mesozoic rocks and younger loose sediments (Flint et al., 2011). Liquefaction is strongly associated more with recent loose deposits as opposed to areas with extensive bedrock outcrop (i.e., MG) (Fouché, 2020). Seismic waves propagating from an earthquake on the Colenso Fault within the MG will likely not result in liquefaction given that the rocks of this group are consolidated (Sekac et al., 2016). As seismic waves propagate into the younger unconsolidated sediments, liquefaction may occur. This phenomenon is proposed to have occurred in 1809 near the Milnerton area, which explains the relatively heightened seismic activity observed near the regions of the Colenso Fault compared to other regions in the study area (Figure 26B).

The design of the KNPS was done such that it can withstand an M_L 7 earthquake without rupture (De Beer, 2016). Liquefaction studies have previously been conducted for the nuclear power plant, which has been conditionally screened out (Eskom, 2023). There is no clear heightened cloud of seismic activity that the KNPS finds itself in from our results and that of (CGS, 2022; USGS, 2022). Given that it can withstand M_L 7 events along with the liquefaction measures implemented, there is no conclusive hazard associated with the KNPS. However, the region surrounding it, as well as the greater Cape Town area, has seismic hazards due to the extensive younger loose sedimentary cover and the potential activity of the Colenso Fault.

6.1. Manual Evaluation vs STA/LTA vs Machine Learning

From the results, all methods failed to identify all true positives present in the dataset (see Table 10 and Figure 23). Manual review of the waveform data resulted in the highest number of true positives with a precision of 100%. ML takes second place for the most effective method since it also has a precision of 100%, but the number of false negatives is 12 (Figure 23). The high false-negative rate may be related to the additional filters added to the ML algorithm to remove false positives. This means that there is a choice when applying ML to a dataset whereby one either aim for an algorithm with 100% precision by analysing the detections and creating a false positive filter based on a single day's data or one can aim for a filter-free algorithm that will allow false positives at the expense of decreasing the false negatives.

Manual evaluation outperforming machine learning on a small dataset agrees with previous studies (Ross et al., 2018; Stepnov et al., 2021; Woollam et al., 2019). When the dataset is small, the identification and classification of phases by machine learning has not yet reached the same level as that of expert seismologists (Ross et al., 2018; Stepnov et al., 2021; Woollam et al., 2019). However, the field is rapidly approaching a point where the performance of machine-learning algorithms will be comparable to that of human analysts (Ross et al., 2018; Stepnov et al., 2021; Woollam et al., 2019). When the datasets are large, manual review of seismograms for phase identification is no longer practical due to resources and time, making machine learning the ideal method for this scenario (Woollam et al., 2019).

The STA/LTA algorithm takes third place for the most efficient method since it has a true positive rate of 80 %, but the number of false positives from this method is 1354 ultimately resulting in a method precision of 3 % for our dataset (Figure 23). STA/LTA shows a lower accuracy compared to human analysts in seismic event detection and classification (Ayub and SanLinn, 2021; Ross et al., 2018; Woollam et al., 2019). The algorithm struggles to distinguish between noise and seismic events, whereas human analysts can recognize both P- and S-waves to distinguish earthquakes from noise disturbances (Huang and Wu, 2019; Ross et al., 2018; Ross, Yue et al., 2019; Woollam et al., 2019).

Ahmed et al. (2021) also did a comparative analysis between STA/LTA and machine learning and found that machine learning outperformed STA/LTA and other traditional methods using borehole data to detect micro-seismic events. Moreover, it is well established that STA/LTA is extremely robust for signals with high signal-to-noise-ratio (SNR) and not so effective for low SNR regions (Ahmed et al., 2021; Sugondo and Machbub, 2021; Trnkoczy, 1999). The number of false positives recorded by this technique in this project is therefore directly related to the low SNR and noise sources present.

6.2. Recommendations

The uncertainty of epicenters and discrepancies between different catalogs such as USGS and CGS, can be reduced by installing a permanent local dense array of broadband seismometers. This will also facilitate continuous seismic monitoring. Additionally, a temporary array of broadband seismometers should be deployed around the Colenso Fault to achieve a higher degree of certainty regarding seismic activity on this fault system and its active parts.

The targeted nature of the network means that detection limits will vary across the various mapped and proposed fault systems in the area. Given that the historical record shows that there is an active fault capable of producing damaging earthquakes in the near vicinity of a major population centre and a nuclear power plant, it would be sensible to maintain a reasonably dense network for a significant period of time to produce an unbiased view of the distribution of micro-seismicity.

Suggest re-establishing on-site liquefaction studies or conducting numerical modelling of potential 6-7 M_L events on the potentially active Colenso Fault to study the impact on the KNPS. The waveform data collected from this project can be used to investigate local attenuation studies to determine kappa-values, which are used in seismic risk assessments. Additionally, ground prediction modelling should be performed, focusing near the Colenso Fault to understand its effects on the Cape Basin and, to some extent, on the KNPS.

A thorough investigation into the seismicity associated with the Cape Canyon should be conducted using ocean-bottom seismometers to constrain the seismicity in this region, as the literature is limited regarding seismic activity related to this area, especially considering its proximity to the KNPS. Moreover, techniques such as earthquake location using envelope cross-correlation should be applied to the waveform data from this project to more precisely determine the locations of the increased offshore seismicity.

Machine learning is effective at identifying micro-seismic events. Training a neural network with local seismic data exclusively might remove the need for additional filters as was done in this project and consequently decrease the false negative count. Upon successful completion and training, this can then be implemented in the proposed local permanent dense array of broadband seismometers.

7. Conclusions

In relation to the aims outlined in Chapter 1, the results and research have led us to the following conclusions:

(a) The region surrounding the KNPS has 2 regions of seismicity. The first region can be identified as an offshore region of elevated seismicity. The second region of elevated seismicity occurs between the Colenso and the hypothetical Milnerton Fault.
(b) The Colenso fault is based on the results provided in this thesis a potential active fault due to elevated seismicity recorded on the fault system, but more research is required to confirm this with a higher degree of certainty.
(c) Manual review for earthquakes in waveform data results in a more precise and complete catalogue when compared to machine learning and STA/LTA algorithms. This is only the case for small datasets as it is not feasible to do manual review on large datasets as it will be time intensive. For large datasets (i.e., datasets composed of years to decades as opposed to months), machine learning is the best option when compared to STA/LTA as it is possible to some extent to decrease the number of false positives and increase the precision of the method at the cost of increasing the number of false negatives. STA/LTA is the third most effective method of phase identification although it has a significantly lower precision than the other two methods. Lastly, the results from this thesis indicate that machine learning models are transferable from one region to the other even if most of the training dataset of the supervised machine learning algorithm is different from the local region to which you are applying it.

References

Ahmed, S. et al. (2021) 'GrONingEnNET: Deep learning for Low-Magnitude earthquake detection on a Multi-Level sensor network,' Sensors, 21(23), p. 8080. https://doi.org/10.3390/s21238080.

Ayub, M. and SanLinn, I. (2021) 'A comparative analysis of machine learning models for first-break arrival picking,' International Journal of Advanced Computer Science and Applications, 12(1). https://doi.org/10.14569/ijacsa.2021.0120157.

Barrientos, S. and National Seismological Center (CSN) Team, 2018. The seismic network of Chile. Seismological Research Letters, 89(2A), pp.467-474.

Begg, G.C. et al. (2009) 'The lithospheric architecture of Africa: Seismic tomography, mantle petrology, and tectonic evolution,' Geosphere, 5(1), pp. 23–50. https://doi.org/10.1130/ges00179.1.

Bergen, K.J. et al. (2019) 'Machine learning for data-driven discovery in solid Earth geoscience,' Science, 363(6433). https://doi.org/10.1126/science.aau0323.

Beyreuther, M. et al. (2010) 'OBSPY: a Python toolbox for seismology,' Seismological Research Letters, 81(3), pp. 530–533. https://doi.org/10.1785/gssrl.81.3.530.

Bishop, C.M. (2006) Pattern recognition and machine learning. Springer Verlag.

Blewett, S.C.J., Phillips, D. and Matchan, E. (2019) 'Provenance of Cape Supergroup sediments and timing of Cape Fold Belt orogenesis: Constraints from high-precision 40Ar/39Ar dating of muscovite,' Gondwana Research, 70, pp. 201–221. https://doi.org/10.1016/j.gr.2019.01.009.

Bommer, J.J. et al. (2015) 'A SSHAC Level 3 Probabilistic Seismic Hazard Analysis for a New-Build Nuclear Site in South Africa,' Earthquake Spectra, 31(2), pp. 661–698. https://doi.org/10.1193/060913eqs145m.

Brandt (2011) Seismic hazard in South Africa. Council for Geoscience Report number:2011-0061.

Burchell W.J. (1822). Travels in the Interior of Southern Africa: Vol. 1. The Batchworth Press, London.

Bumby, A.J., Eriksson, P. and Van Der Merwe, R. (2004) 'The early Proterozoic sedimentary record in the Blouberg area, Limpopo Province, South Africa; implications for the timing of the Limpopo orogenic event,' Journal of African Earth Sciences [Preprint]. https://doi.org/10.1016/j.jafrearsci.2004.07.056.

Calais, E. et al. (2016) 'A new paradigm for large earthquakes in stable continental plate interiors,' Geophysical Research Letters, 43(20). https://doi.org/10.1002/2016gl070815.

Chauhan, N.S. (2022) DBSCAN Clustering Algorithm in Machine Learning - KDNuggets. https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html (Accessed: December 6, 2022).

Council for Geoscience. CGS earthquake catalogue. 2022.

Dames and Moore (1976). Geologic Report. Koeberg Power Station, Cape Province, South Africa. Report prepared for The Electricity Supply Commission. Job No.: 9629-014-45.

Dames and Moore, 1981. Preliminary coal and power investigation – Moretele 2/Matahanjana District,for the Department of Economic Affairs, Republic of Bophuthatswana, Volume 1, Sections I and II.

De Beer, J. (2016) The history of geophysics in Southern Africa. AFRICAN SUN MeDIA.

Dertat, A. (2018) 'Applied Deep Learning - Part 4: Convolutional neural networks,' Medium, 19 June. https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2.

Eskom Holdings SOC Ltd (2023) Safety Case for Long-Term Operation of Koeberg Nuclear Power Station. 331–618. https://nnr.co.za/wp-content/uploads/2024/01/K-29731-E-Attachment-7-331-618-Safety-Case-for-Long-Term-Operation-of-Koeberg-Nuclear-Power-Station-Revision-3-1.pdf (Accessed: November 15, 2023).

Ferraro, M. and Giordani, P. (2019) 'Soft clustering,' Wiley Interdisciplinary Reviews: Computational Statistics, 12(1). https://doi.org/10.1002/wics.1480.

Flint, S.S. et al. (2011) 'Depositional architecture and sequence stratigraphy of the Karoo basin floor to shelf edge succession, Laingsburg depocentre, South Africa,' Marine and Petroleum Geology, 28(3), pp. 658–674. https://doi.org/10.1016/j.marpetgeo.2010.06.008.

Fouché, N. (2020) 'The liquefaction potential of the upper quaternary sands of the Cape Flats, Western Cape, South Africa,' Journal of the South African Institution of Civil Engineers, 62(2), pp. 22–30. https://doi.org/10.17159/2309-8775/2020/v62n2a3.

Gaol, Y.H.L. et al. (2021) 'Preliminary Results of Automatic P-Wave Regional Earthquake Arrival Time Picking Using Machine Learning with STA/LTA As the Input Parameters,' IOP Conference Series, 873(1), p. 012060. https://doi.org/10.1088/1755-1315/873/1/012060.

Gesret, A. et al. (2014) 'Propagation of the velocity model uncertainties to the seismic event location,' Geophysical Journal International, 200(1), pp. 52–66. https://doi.org/10.1093/gji/ggu374.

Grigoli, Francesco, Luca Scarabello, Maren Böse, Bernd Weber, Stefan Wiemer, and John F. Clinton. "Pick- and waveform-based techniques for real-time detection of induced seismicity." Geophysical Journal International, 213, no. 2 (2018): 868-884.

Gresse P.G., von Veh, M.W., and Frimmel, H.E., (2006). Namibian (Neoproterozoic) to Early Cambrian Successions, in M.R. Johnson, C.R. Anhaeusser and R.J. Thomas (eds), The Geology of South Africa, pg 395-421.

Hartnady, C.J.H., 2003. Cape Town earthquakes: review of the historical record.

Hauksson, E. et al. (2020) 'Caltech/USGS Southern California Seismic Network (SCSN) and Southern California Earthquake Data Center (SCEDC): data availability for the 2019 Ridgecrest sequence,' Seismological Research Letters, 91(4), pp. 1961–1970. https://doi.org/10.1785/0220190290.

Huang, T. and Wu, Y. (2019) 'A robust algorithm for Automatic P-Wave Arrival-Time picking based on the local Extrema scalogram,' Bulletin of the Seismological Society of America, 109(1), pp. 413–423. https://doi.org/10.1785/0120180127.

Johnston, S.T. (2000) 'The Cape Fold Belt and Syntaxis and the rotated Falkland Islands: dextral transpressional tectonics along the southwest margin of Gondwana,' Journal of African Earth Sciences, 31(1), pp. 51–63. https://doi.org/10.1016/s0899-5362(00)00072-5.

Kanaujia, J., Kumar, A. and Gupta, S.C. (2015) '1D Velocity Structure and Characteristics of Contemporary Local Seismicity around the Tehri Region, Garhwal Himalaya,' Bulletin of the Seismological Society of America, 105(4), pp. 1852–1869. https://doi.org/10.1785/0120140306.

Kijko, A., Vermeulen, P.J. and Smit, A. (2021) 'Estimation Techniques for Seismic Recurrence Parameters for Incomplete Catalogues,' Surveys in Geophysics, 43(2), pp. 597–617. https://doi.org/10.1007/s10712-021-09672-2.

Kisters, A.F.M. et al. (2002) 'Timing and kinematics of the Colenso Fault: The Early Paleozoic shift from collisional to extensional tectonics in the Pan-African Saldania Belt, South Africa,' South African Journal of Geology, 105(3), pp. 257–270. https://doi.org/10.2113/1050257.

Kisters, A.F.M. and Belcher, R.N. (2018) 'The Stratigraphy and Structure of the Western Saldania Belt, South Africa and Geodynamic Implications,' in Regional geology reviews. Springer International Publishing, pp. 387–410. https://doi.org/10.1007/978-3-319-68920-3_14.

Kong, Q. et al. (2018) 'Machine Learning in Seismology: Turning Data into Insights,' Seismological Research Letters, 90(1), pp. 3–14. https://doi.org/10.1785/0220180259.

Kong, Q. et al. (2019) 'Machine Learning in Seismology: Turning Data into Insights,' Seismological Research Letters, 90(1), pp. 3–14. https://doi.org/10.1785/0220180259.

Krüger, F. and Scherbaum, F. (2014) 'The 29 September 1969, Ceres, South Africa, earthquake: Full waveform moment tensor inversion for point source and kinematic source parameters,' Bulletin of the Seismological Society of America, 104(1), pp. 576–581. https://doi.org/10.1785/0120130209.

Lienert, B.R., Berg, E. and Frazer, L.N. (1986) 'HYPOCENTER: An earthquake location method using centered, scaled, and adaptively damped least squares,' Bulletin of the Seismological Society of America, 76(3), pp. 771–783. https://doi.org/10.1785/bssa0760030771.

Liermann, V. and Stegmann, C. (2019) The Impact of Digital Transformation and FinTech on the Finance Professional. Springer Nature.

Lock, B. E. (1980). Flat-plate subduction and the Cape Fold Belt of South Africa. Geology, 8, 35-39.

MacHutchon et al. (2020) 'What the marine geology of Table Bay, South Africa can inform about the western Saldania Belt, geological evolution and sedimentary dynamics of the region,' Journal of African Earth Sciences, 162, p. 103699. https://doi.org/10.1016/j.jafrearsci.2019.103699.

Malservisi, R. et al. (2013) 'How rigid is a rigid plate? Geodetic constraint from the TrigNet CGPS network, South Africa,' Geophysical Journal International, 192(3), pp. 918–928. https://doi.org/10.1093/gji/ggs081.

Manzunzu, B. et al. (2017) 'The aftershock sequence of the 5 August 2014 Orkney earthquake (ML 5.5), South Africa,' Journal of Seismology, 21(6), pp. 1323–1334. https://doi.org/10.1007/s10950-017-9667-z.

Manzunzu, B. et al. (2019) 'Seismotectonics of South Africa,' Journal of African Earth Sciences, 149, pp. 271–279. https://doi.org/10.1016/j.jafrearsci.2018.08.012.

Martín-González, F. et al. (2023) 'Understanding seismicity and seismotectonics in a stable continental region (NW Iberian Peninsula): Implications for the nature of intraplate seismicity,' Global and Planetary Change, 227, p. 104177. https://doi.org/10.1016/j.gloplacha.2023.104177.

Midzi, V. et al. (2010) '1-D velocity model for use by the SANSN in earthquake location,' Seismological Research Letters, 81(3), pp. 460–466. https://doi.org/10.1785/gssrl.81.3.460.

Miyamoto, S., Ichihashi, H. and Honda, K. (2008) 'Introduction,' in Springer eBooks, pp. 1–7. https://doi.org/10.1007/978-3-540-78737-2_1.

Miyamoto, T. et al. (2022) 'Characteristics of Fault Rocks Within the Aftershock Cloud of the 2014 Orkney Earthquake (M5.5) Beneath the Moab Khotsong Gold Mine, South Africa,' Geophysical Research Letters, 49(14). https://doi.org/10.1029/2022gl098745.

Moen, H.F.G. (1999) 'The Kheis tectonic subprovince, Southern Africa; a lithostratigraphic perspective,' South African Journal of Geology, 102(1), pp. 27–42. http://sajg.geoscienceworld.org/content/102/1/27.

Mooney, W.D. (2007) 'Crust and lithospheric structure – global crustal structure,' in Elsevier eBooks, pp. 361–417. https://doi.org/10.1016/b978-044452748-6.00011-0.

Mousavi, S.M. et al. (2020) 'Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking,' Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-17591-w.

Münchmeyer, J. (2022) Machine learning for fast and accurate assessment of earthquake source parameters. PhD Dissertation. Humboldt-Universitat zu Berlin.

Münchmeyer, J. et al. (2022) 'Which Picker Fits My Data? A Quantitative Evaluation of Deep Learning Based Seismic Pickers,' Journal of Geophysical Research: Solid Earth, 127(1). https://doi.org/10.1029/2021jb023499.

Münchmeyer, J. (2024) 'PyOcto: A high-throughput seismic phase associator,' Seismica, 3(1). https://doi.org/10.26443/seismica.v3i1.1130.

Nkosi, N.Z. et al. (2022) 'Physical property studies to elucidate the source of seismic reflectivity within the ICDP DSeis seismogenic zone: Klerksdorp goldfield, South Africa,' International Journal of Rock Mechanics and Mining Sciences, 155, p. 105082. https://doi.org/10.1016/j.ijrmms.2022.105082.

Ottemöller et al. (2015) Seisan earthquake analysis software for Windows, Solaris, Linux and MacOsX.

Paton, D., Macdonald, D.I.M. and Underhill, J.R. (2006) 'Applicability of thin or thick skinned structural models in a region of multiple inversion episodes; southern South Africa,' Journal of Structural Geology, 28(11), pp. 1933–1947. https://doi.org/10.1016/j.jsg.2006.07.002.

Raith, J.G. et al. (2003) 'New Insights into the Geology of the Namaqua Tectonic Province, South Africa, from Ion Probe Dating of Detrital and Metamorphic Zircon,' The Journal of Geology, 111(3), pp. 347–366. https://doi.org/10.1086/373973.

Raschka, S. and Mirjalili, V. (2019) Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2, 3rd Edition.

Rehman, S.U. et al. (2014) DBSCAN: Past, present and future. https://doi.org/10.1109/icadiwt.2014.6814687.

Ross, Z.E. et al. (2019) 'PhaseLink: a Deep learning approach to seismic Phase Association,' Journal of Geophysical Research: Solid Earth, 124(1), pp. 856–869. https://doi.org/10.1029/2018jb016674.

Ross, Z.E., Meier, M. and Hauksson, E. (2018) 'P Wave Arrival Picking and First-Motion Polarity Determination With Deep Learning,' Journal of Geophysical Research: Solid Earth, 123(6), pp. 5120–5129. https://doi.org/10.1029/2017jb015251.

Ross, Z.E. and Zhu, W. (2023) 'Neural mixture model association of seismic phases,' arXiv (Cornell University) [Preprint]. https://doi.org/10.48550/arxiv.2301.02597.

Rozendaal, A. et al. (1999) 'Neoproterozoic to early Cambrian crustal evolution of the Pan-African Saldania Belt, South Africa,' Precambrian Research, 97(3–4), pp. 303–323. https://doi.org/10.1016/s0301 9268(99)00036-4.

Schoch, A.E., (1975). The Darling Granite Batholith, Annals of the University of Stellenbosch (A1), 1, pg 1-104.

Schubert, E. et al. (2017) 'DBSCAN Revisited, Revisited,' ACM Transactions on Database Systems, 42(3), pp. 1–21. https://doi.org/10.1145/3068335.

Schultz, R. et al. (2020) 'Hydraulic Fracturing-Induced Seismicity,' Reviews of Geophysics, 58(3). https://doi.org/10.1029/2019rg000695.

Sekac, T.N. (2016) 'A GIS Based Approach into Delineating Liquefaction Susceptible Zones Through Assessment of Site-Soil-Geology-A Case Study of Madang and Morobe Province in Papua New Guinea (PNG),' International Journal of Innovative Research in Science Engineering and Technology/International Journal of Innovative Research in Science, Engineering and Technology, 5(5), pp. 6616–6629. https://doi.org/10.15680/ijirset.2016.0501003.

Shearer, P.M. (2000) 'Introduction to seismology,' Choice Reviews Online, 37(08), pp. 37–4521. https://doi.org/10.5860/choice.37-4521.

Smit, L. et al. (2015) 'Microseismic activity and basement controls on an active intraplate Strike-Slip fault, Ceres–Tulbagh, South Africa,' Bulletin of the Seismological Society of America, 105(3), pp. 1540–1547. https://doi.org/10.1785/0120140262.

Stamatakos et al. (2024) Enhanced SSHAC Level 2 Probabilistic Seismic Hazard Analysis for the Duynefontyn nuclear site, Western Cape Province, South Africa., Council for Geoscience. CGS Report 2024-0001 Rev.0. Council for Geoscience. https://nnr.co.za/wp-content/uploads/2024/04/19.-Duynefontyn-SSHAC-EL2-PSHA-CGS-2024-0001-R0-compressed-PART-2_1.pdf (Accessed: March 31, 2024).

Stein, S. and Liu, M. (2009) 'Long aftershock sequences within continents and implications for earthquake hazard assessment,' Nature, 462(7269), pp. 87–89. https://doi.org/10.1038/nature08502.

Stepnov, A., Chernykh, V. and Konovalov, A., 2021. The seismo-performer: a novel machine learning approach for general and efficient seismic phase recognition from local earthquakes in real time. Sensors, 21(18), p.6290.

Strasser, F.O., Albini, P., Flint, N.S. and Beauval, C. 2015. Twentieth century seismicity of the Koffiefontein region (Free State, South Africa): consistent determination of earthquake catalogue parameters from mixed data types. Journal of Seismology, 19, 915–934.

Strasser, F.O. et al. (2015) 'Twentieth century seismicity of the Koffiefontein region (Free State, South Africa): consistent determination of earthquake catalogue parameters from mixed data types,' Journal of Seismology, 19(4), pp. 915–934. https://doi.org/10.1007/s10950-015-9503-2.

Sugondo, R.A. and Machbub, C. (2021) 'P-Wave detection using deep learning in time and frequency domain for imbalanced dataset,' Heliyon, 7(12), p. e08605. https://doi.org/10.1016/j.heliyon.2021.e08605.

Thamm, A.G., Johnson, M.R., Anhaeusser, C.R. and Thomas, R.J., 2006. The cape supergroup. The Geology of South Africa, pp.443-460.

Theron, J.N., Gresse, P.G., Siegfried, H.P., and Rogers, J., (1992). The Geology of the Cape Town Area, Geological Survey of South Africa. Explanation of sheet 3318, 140 pg.

Trnkoczy, A. (1999) Understanding and parameter setting of STA/LTA trigger algorithm.

United States Geological Survey. USGS earthquake catalogue. 2022.

Vaezi, Y. and Van Der Baan, M. (2015) 'Comparison of the STA/LTA and power spectral density methods for microseismic event detection,' Geophysical Journal International, 203(3), pp. 1896–1908. https://doi.org/10.1093/gji/ggv419.

Van Niekerk, H.S. (2009) The origin of the Kheis Terrane and its relationship with the Archean Kaapvaal Craton and the Grenvillian Namaqua province in Southern Africa. PhD dissertation. University of Johannesburg.

Viola, G., Kounov, A., Andreoli, M. A. G., and Mattila, J. (2012). Brittle tectonic evolution

Waldeland, A.U. et al. (2018) 'Convolutional neural networks for automated seismic interpretation,' The Leading Edge, 37(7), pp. 529–537. https://doi.org/10.1190/tle37070529.1.

Von Buchenroder W.L. (1830). An account of the earthquake which occurred at the Cape of Good Hope, during the month of December 1809; etc. S. Afr. Quart. Jnl 1, 18-25

Wigley, R.A. and Compton, J.S. (2005) 'Late Cenozoic evolution of the outer continental shelf at the head of the Cape Canyon, South Africa,' Marine Geology, 226(1–2), pp. 1–23. https://doi.org/10.1016/j.margeo.2005.09.015.

Woollam, J. et al. (2019) 'Convolutional Neural Network for Seismic Phase Classification, Performance Demonstration over a Local Seismic Network,' Seismological Research Letters, 90(2A), pp. 491–502. https://doi.org/10.1785/0220180312.

Woollam, J. et al. (2022) 'SeisBench—A Toolbox for Machine Learning in Seismology,' Seismological Research Letters, 93(3), pp. 1695–1709. https://doi.org/10.1785/0220210324.

Zhu, W. et al. (2022) 'Earthquake Phase Association Using a Bayesian Gaussian Mixture Model,' Journal of Geophysical Research: Solid Earth, 127(5). https://doi.org/10.1029/2021jb023249.

Zhu, W. and Beroza, G.C. (2018) 'PhaseNet: a Deep-Neural-Network-Based seismic arrival time picking method,' Geophysical Journal International [Preprint]. https://doi.org/10.1093/gji/ggy423.