Data Preprocessing for Machine Learning in Seismology

Vladimir Chernykh [a], Andrey Stepnov [b]and Olga Lukyanova [a]

[a] *Computing Center of Far-Eastern Branch, Russian Academy of Sciences, 65 Kim Yo Cheng st., Khabarovsk, 680000, Russia*

[b] *Geophystech LLC, 1b Nauki, Yuzhno-Sakhalinsk, 693022, Russia*

Abstract

The problem of preliminary data processing on P, S arrivals of seismic waves has been formulated. Data preprocessing was carried out for further classification using machine learning models. A comparative analysis of the following neural networks has been carried out: GPD, EQTransformer, and PhaseNet. Demonstrated the automation process for machine learning methods of seismic waves detection.

Keywords 1

Machine learning, earthquake, time series, waveform data

1. Introduction

Current day earthquake detection and analysis require the necessity of staff involvement, trained in visual detection of different seismic waves in a continuous stream of data from local seismic networks.

Local earthquakes generate different types of seismic waves, which travel away from the source. The fastest among these are P-waves and S-waves (primary and secondary). Accurate detection of P and S waves is used in earthquake source location by computing source parameters: coordinates, hypocenter depth, and origin time.

The continuous growth of seismic networks causes an increase in trained human staff demand.

Threshold methods [1] are very popular as an incomplete approach to seismic events detection automation. However, these methods have proven ineffective in low-magnitude earthquake detection, especially in noisy environments.

The machine learning approach has shown an ability to achieve detection accuracy compared to (or even surpassing) which of trained staff [2].

Preprocessing of seismic data is a first and critical step in full automation of classification of seismic wave arrival times. The present paper demonstrates seismic data preprocessing for subsequent use in machine learning methods of earthquake detection and describes the method employed to automate machine learning methods of seismic waves detection.

2. Data description

For neural-networks training and evaluation purposes, we used a dataset of hand-picked local earthquake data from the Sakhalin island seismic network. Dataset consists of 3045 P-arrivals, 3737 S-arrivals, and 3045 noise fragments collected from 2014 to 2021.

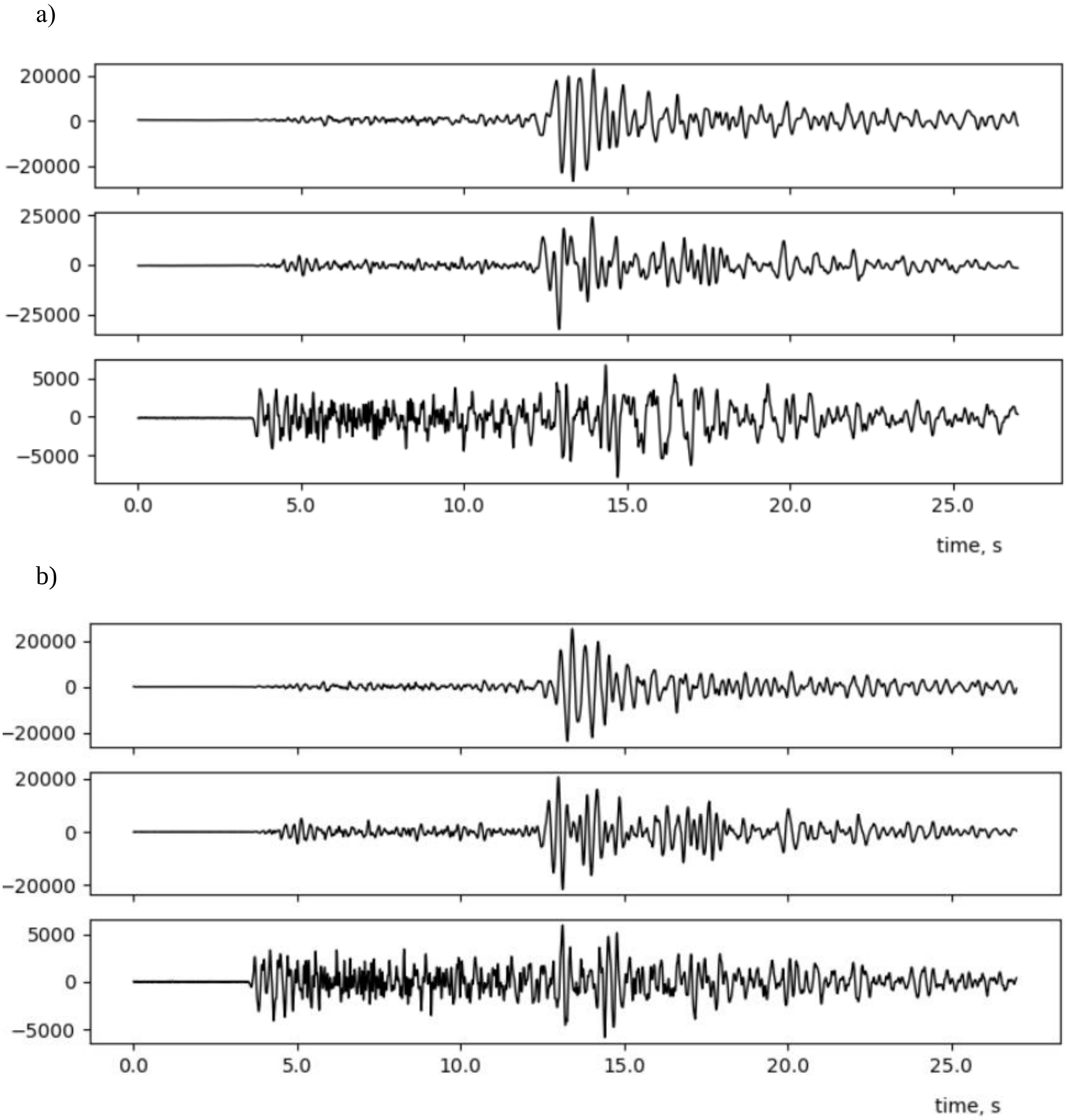
Each seismic record is a 3-component (North, East, and vertical components) 4 seconds slice of ground movement information with a sampling rate of 100 Hz. Continuous seismic data streams are usually stored as day-length entries with gaps for station offline times. Data gathered from stations

VI International Conference Information Technologies and High-Performance Computing (ITHPC-2021), September14–16, 2021, Khabarovsk, Russia

EMAIL: syler1983.9@gmail.com (A. 1); a.stepnov@geophystech.ru (A. 2)
ORCID: 0000-0002-8383-8620 (A. 1); 0000-0003-3401-6687 (A. 2)

with sampling rates different from 100 Hz were resampled to 100 Hz using the Fourie method. Accelerometer data were integrated by time to convert it to seismograms.

Seismic events were filtered by a minimal magnitude of 1 and maximum distance to an earthquake source of 300 km. The data first were detrended and high-pass filtered above 2 Hz and then normalized by the absolute maximum amplitude on any of the three components. Figure 1 illustrates preprocessing on an actual earthquake from 01.04.2021.



a)

b)

Figure 1: Seismic data before (a) and after (б) detrend and filtering

Also, we used a larger dataset of southern California seismic network records [2], composed of 4 773 750 3-component records with an equal number of P-wave arrivals, S-wave arrivals, and noise patches. Data preprocessing is similar to formerly described methods used on the Sakhalin data.

We used day-long continuous data from 3-component seismographs to evaluate the program integration of machine learning methods.

3. Models

The integration is designed for classification neural networks and supports output classes number, labels, and positive characteristic (seismic event vs. noise) customization.

In this study, following models was evaluated: GPD [2], EQTransformer [3], and PhaseNet [4].

GPD input is a 3-component 4 seconds long record (with a sampling rate of 100 Hz), in other words, an array of data with a shape 400x3. Model output is a set of three probabilities corresponding to the likelihood of each respective class: P-wave, S-wave, and noise. EQTransformer and PhaseNet input is a 3-component record of 60 and 90 seconds length, respectively.

For PhaseNet and EQTransformer evaluation, we reconstructed datasets to meet new input data shape requirements. New datasets composed of the same P and S waves arrivals and employed the same preprocessing methods as described in section 2. Data description.

All models were trained on southern California data and evaluated on 20% of Sakhalin data (table 1, pre-trained), followed by fine-tuning on 80% of Sakhalin data with evaluation on 20% of Sakhalin data (table 1, fine-tuned). GPD displayed the best results and thus was chosen as a target model for the automation process of seismic events detection.

Table 1
Models evaluation results on the local seismic events data

Model	Accuracy, pretrained	Accuracy, fine-tuned	F1 score, pretrained			F1 score, Fine-tuned		
			P	S	N	P	S	N
GPD	**0.87**	**0.94**	**0.89**	**0.87**	**0.86**	**0.94**	**0.95**	**0.93**
EQTransformer	0.51	0.81	0.34	0.20	0.34	0.86	0.89	0.34
PhaseNet	0.61	0.49	0.69	0.66	0.66	0.60	0.61	0.50

4. Automation of seismic waves detection

Program integration is designed to work with seismogram databases generated by SEISAN [5] software.

SEISAN software is a software package for analyzing earthquakes. The system provides the means to maintain the database containing the configuration of the seismic station network, earthquake records, data stream archives from the seismic station network.

SEISAN database includes the following directories:

- REA – earthquake readings and full epicenter solutions in a database
- WOR – the users work directory
- DAT – default and parameter files, system configuration files
- WAV – digital waveform data files
- archives – database of continuous seismic data from stations split into day-long files

The product of the automation development is a program that analyzes SEISAN database files and searches for earthquakes on the data stream from specified stations. The program analyzes the network configuration, including information on active stations from the DAT directory, and searches daily archives of the continuous stream from seismic stations.

Currently, the automation is not used in real-time, rather daily analysis of the new seismic data from the specified monitoring stations is performed. The workflow of the earthquake detection automation is displayed in figure 2.

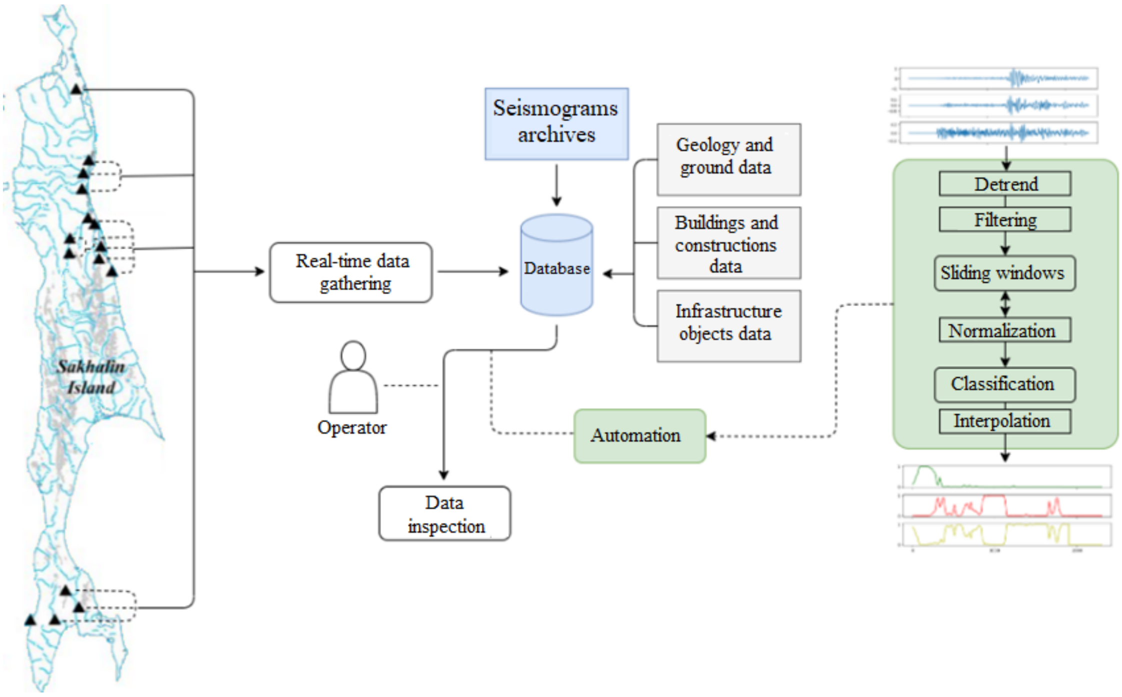


Figure 2: Earthquake detection automation workflow

The automation employs the following workflow: program scans SEISAN database structure and then picks correct seismogram archive for analysis. Next, preliminary data processing is performed: detrend and high-pass filtering above 2 Hz. The seismic data stream is then split by a sliding window with a length of 4 seconds and a step of 0.1 seconds; each window is normalized and used as an input for the target neural network prediction.

Class predictions were then restored to input data frequency (from 10 Hz to 100 Hz) using linear interpolation, resulting in three probability curves: P-wave curve, S-wave curve, and noise probability curve. Probability curves example displayed in figure 3,b alongside with raw input data (figure 3,a).

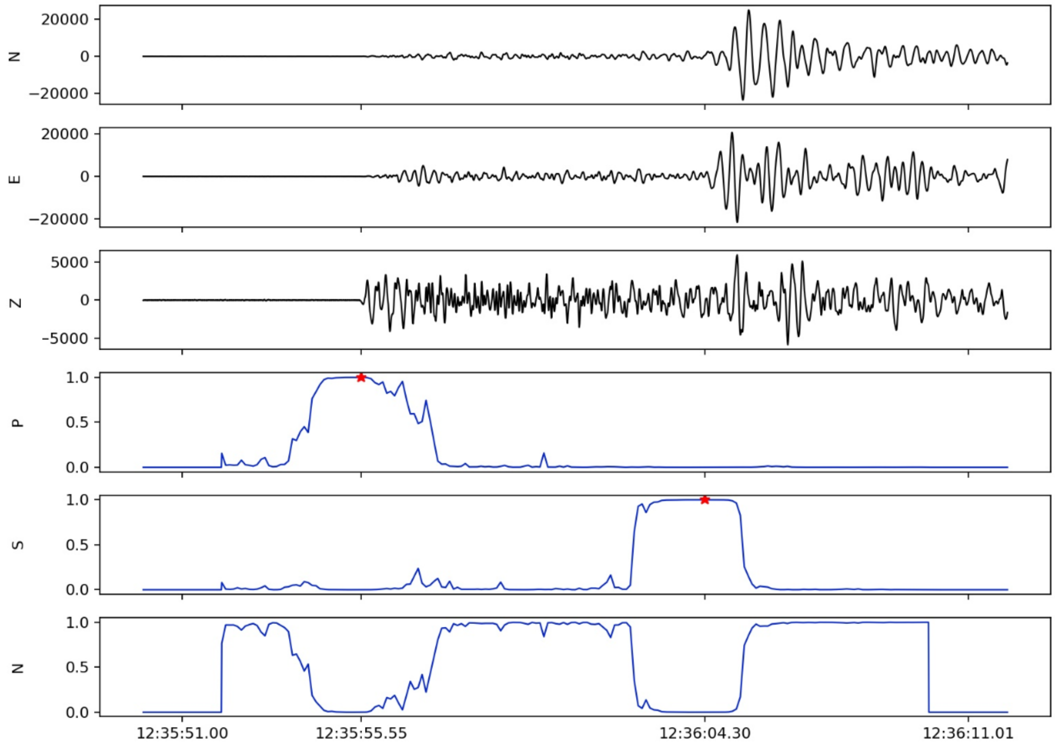


Figure 3: Probability curves example for an earthquake prediction (N, E, Z – input components; P, S – seismic waves probability, N – noise probability)

Positive wave arrivals predictions are marked as red “stars” on P and S probability curves.

For each positive class (P and S waves), probability curve peaks are found with the minimal allowed distance between adjacent peaks in 10 seconds and the threshold probability value is 0.95.

Then, the mean value in a 1-second span (a quarter of the window length) around the peak position is calculated for every class probability curve. Finally, the values are compared, and if the mean value belonging to the peak is highest, then the peak position is assumed as a positive prediction.

Positives are then outputted in a text file in order of occurrence with corresponding information about positives time, probability, type (P-arrival, S-arrival), and seismic station.

In addition, the program supports data and predictions visualization in the form of graphs, including scores visualization (figure 3), preprocessed data plotting, and raw data plotting (figure 3).

Also, launch options for performance evaluation of entire automation and only neural network computation times were implemented to provide means for future models comparative analysis.

5. Conclusion

The development of the present study yielded the application for the automation of seismic waves detection using machine learning methods.

Also, during the automation evaluation, new classification neural network training flaws were revealed, which may lead to further studies and improvements.

Extensive effort was put into data gathering and processing for model training and evaluation, which may be used for future projects and new classification neural networks. Also, a program package was developed for data (P and S waves and noise records) gathering from SEISAN databases. The package also includes the ability to filter out events by magnitude, source depth and distance, and seismic monitoring station properties (such as the number of components, instrument types).

6. Acknowledgements

The work was done with partial support by the Russian Foundation for Basic Research, project No. 18-29-03196. The studies were carried out using the resources of the Center for Shared Use of Scientific Equipment "Center for Processing and Storage of Scientific Data of the Far Eastern Branch of the Russian Academy of Sciences" [6], funded by the Russian Federation represented by the Ministry of Science and Higher Education of the Russian Federation under project No. 075-15-2021-663.

7. References

[1] Withers M., Aster R., Young C., Beiriger J., Harris M., Moore S. and Trujillo, J.. A Comparison of Select Trigger Algorithms for Automated Global Seismic Phase and Event Detection. Bulletin of the Seismological Society of America, Vol. 88, No. 1, pp. 95-106, February 1998.

[2] Ross Z., Meier M., Hauksson E., Heaton T.. Generalized Seismic Phase Detection with Deep Learning. Bulletin of the Seismological Society of America, Vol. 108, pp. 2894-2091, 2018.

[3] Mousavi, S.M., Ellsworth, W.L., Zhu, W. et al. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nat Commun 11, 3952 (2020). https://doi.org/10.1038/s41467-020-17591-w.

[4] Zhu, W., & Beroza, G. C. (2018). PhaseNet: A Deep-Neural-Network-Based Seismic Arrival Time Picking Method. arXiv preprint arXiv:1803.03211.

[5] SEISAN - earthquake analysis software. URL: http://seisan.info.

[6] A. A. Sorokin, S. V. Makogonov, S. P. Korolev, The information infrastructure for collective scientific work in the Far East of Russia, Scientific and Technical Information Processing 44(4) (2017) 302-304, doi:10.3103/S0147688217040153.