Automatic Waveform Quality Control for Surface Waves Using Machine Learning Techniques

Chengping Chai[1] ⟨chaic@ornl.gov⟩, Jonas Kintner[2] ⟨jvk5803@psu.edu⟩, Kenneth M. Cleveland[3] ⟨mcleveland@lanl.gov⟩, Jingyi Luo[4] ⟨jl6zh@virginia.edu⟩,
Monica Maceira[1,5] ⟨maceiram@ornl.gov⟩, Charles J. Ammon[2] ⟨charlesammon@psu.edu⟩, Hector J. Santos-Villalobos[1] ⟨hsantos@ornl.gov⟩

1. Oak Ridge National Laboratory; 2. Department of Geosciences, Pennsylvania State University; 3. Los Alamos National Laboratory;
4. Data Science Institute, University of Virginia; 5. Department of Physics and Astronomy, University of Tennessee.

Introduction

Large seismic waveform data sets are common as a result of additional seismic station and seismic network deployments growing in frequency and size. Unfortunately, not all waveforms have sufficient signal-to-noise characteristics and contribute useful information when employed in specific analysis techniques. For this reason, waveform data are usually examined before they are included in analyses sensitive to noisy signals. Quality control of waveform data, however, requires substantial time and effort, that can be a significant burden with large datasets. The complexity of surface wave signals makes reliable automation of the quality control process a challenge. In some cases, data quality control becomes the most time-consuming part of the analysis.

We screened roughly 400,000 surface-wave waveforms and assigned quality levels (A, B, C etc.) to each waveform as part of efforts to improve earthquake locations in remote regions. The labeled dataset allows us to apply machine learning techniques such as logistic regression, support vector machine, K-nearest neighbors, random forests, and neural networks. These machine-learning algorithms can reduce time cost associated with waveform quality control and could speed up the processing of surface-wave waveforms especially for projects on the global scale.

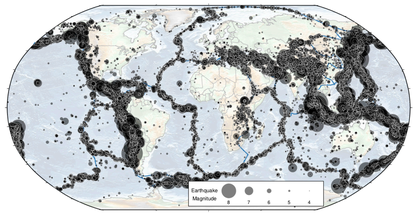

Fig. 1 Global seismic events with magnitude larger than 4.5 occurred between 1990 and 2019.

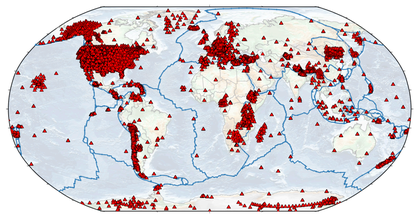

Fig. 2 Seismic stations available at data centers. Most of these stations are/were deployed temporarily.

Data & Method

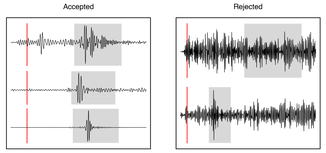

Fig. 3 Examples of accepted and rejected waveforms. The red vertical line indicates earthquake origin time. The grey box shows the expected arrival times associated with group velocities between 2.5 km/s and 5 km/s.

The quality labels were obtained from several earthquake relocation studies including Cleveland and Ammon (2013, 2015), Cleveland et al. (2018), Kintner et al. (2018, 2019) and unpublished results by K. M. Cleveland and J. A. Kintner. Surface-wave waveforms were downloaded from Incorporated Research Institutions for Seismology (IRIS) Data Management Center (DMC). The following machine-learning algorithms were applied to the dataset.

- Logistic regression (LR)
- Support vector machine (SVM, with linear and Gaussian kernels)
- K-nearest neighbors (KNN)
- Random forests (RF)
- Artificial neural networks (ANN)

Since surface-wave waveforms have variable lengths, we first compute statistical features from each waveform and then use these features to train machine learning models. Most features were computed in the time domain to reduce computational cost. Frequency domain features work equally well when designed properly.

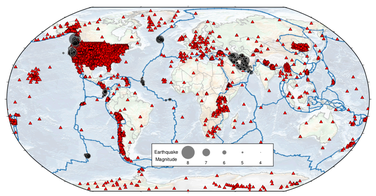

Fig. 4 A map showing earthquakes (circles) and seismic stations (triangles) used by this study. These earthquakes were relocated with a surface-wave double difference algorithm by several different projects. Quality labels were assigned to associated waveforms during the processing.

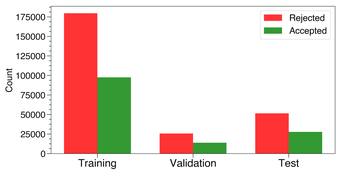

Fig. 5 Labels used by machine-learning algorithms. The initial labels contain either four (A, B, C, and D) or five levels (A, B, C, D, and F) depending on who labeled the waveforms. We have converted these initial labels to two levels (accept or reject) as a way to include more data and for generalization. The dataset were split into three parts including 70% for training, 10% for validation, and 20% for test.

Preliminary Results

Model Accuracy

We trained the machine learning models with increasing number of training samples. The test errors were computed with the test dataset with a size of 55,684 samples. As expected, the resulting models perform generally better when more training samples are included. The improvement for logistic regression and support vector machine with a linear kernel is small after 10,000 samples. K-nearest neighbors algorithm performs poorly when the training sample size is small. The support vector machine with a Gaussian kernel leads to better results than the linear kernel.

The random forests and artificial neural networks improve rapidly with large amount of training samples. The artificial neural networks outperform the random forests after 20,000 samples. For similar problems, our results suggest random forests algorithm is more suitable for training size less than 20,000 and artificial neural networks algorithm works best for training size larger than 20,000 samples. The smallest test error rate is 7.8%.

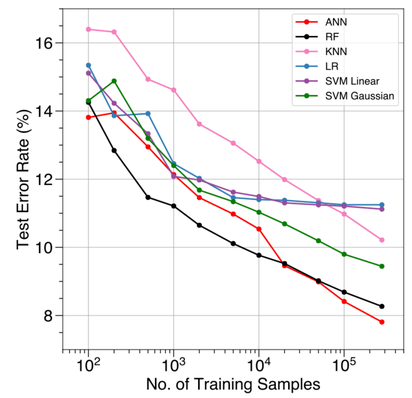

Fig. 6 A comparison of test error rates for different machine-learning models.

We compared the models trained with different algorithm using the entire train dataset. Artificial neural networks has the largest accuracy, recall, and precision while logistic regression has the smallest. The performance of random forests and artificial neural networks is noticeably better (around 2%) than other models. Support vector machine with a linear kernel performs almost the same as logistic regression. K-nearest neighbors and support vector machine with a Gaussian kernel have a performance in between. The best accuracy is 92.2% from the artificial neural networks. We prefer the model trained with artificial neural networks since it is the best for all three metrics (accuracy, recall, and precision).

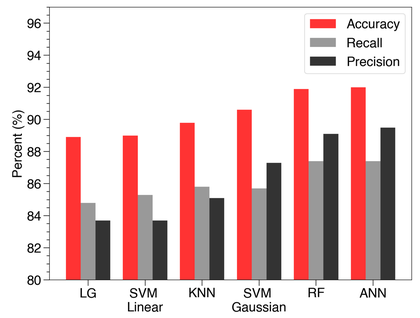

Fig. 7 A comparison of model performance using three different metrics (accuracy, recall, and precision).

Model Speed & Size

We also compared the efficiency of the machine-learning models. In terms of the model size, the logistic regression is the smallest (3.3KB) while the K-nearest neighbors is the largest (1.4GB). The random forests model has a size of 174MB. The model of artificial neural networks requires 4.8MB of storage. The support vector machine models have a size similar to random forests.

Assuming 5 seconds per waveform, it takes a human 18,205 seconds to label 3,641 waveforms. The machine learning models can be about 380 times faster. It only takes about 48 seconds to label the same dataset if we use random forests, artificial neural networks, or logistical regression models. Even for the slower models (K-nearest neighbors and support vector machine), the machine is almost 100 times faster than human.

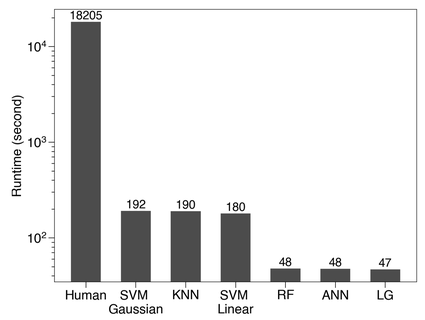

Fig. 8 A comparison of runtime for different machine-learning models and human. The runtime includes both feature extraction and model prediction.

Preliminary Results

Waveform Examination

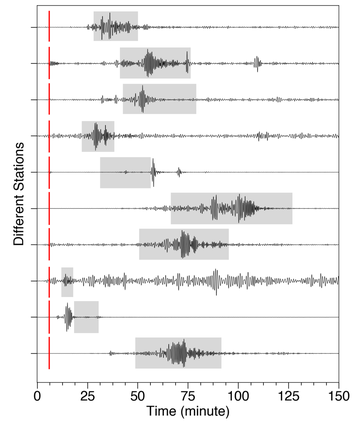

Fig. 9 Example waveforms accepted by human but rejected by the ANN model.

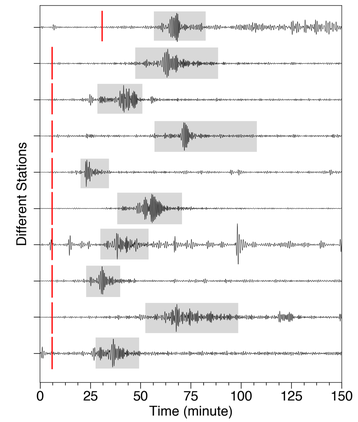

Fig. 10 Example waveforms rejected by human but accepted by the ANN model.

We examine individual waveforms in the validation and test dataset using both static and interactive visualizations. As shown in Fig. 9, many of the human-accepted waveforms may be mislabeled. Some examples show contaminations of surface waves from another seismic event. Several examples contain complicated wave trains that may due to complex subsurface structure. In Fig. 10, we can see many of the human rejected waveforms may be acceptable.

We also use Google Earth to visualize all the waveforms in the validation and test dataset interactively. By comparing the ANN-predicted labels against human-picked labels, we found many of the incorrectly predicted labels may be due to human errors. The ANN model performs reasonably well compare to human. Independent tests with earthquakes outside of the labeled dataset also show reasonable results.

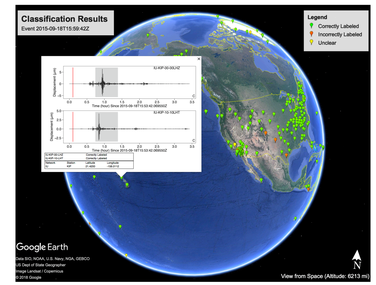

Fig. 11 Screenshot of an interactive interface based on Google Earth that we used to visually inspect the results.

Directions & Acknowledgements

Regardless of the details, machine learning approaches, tuned to be conservative and avoid discarding in-determinant signals, can dramatically reduce the burden of waveform-quality assessment. We will measure human error in the labels and compare it with machine learning models. Additional machine learning algorithms will be tested. We will also try to balance the data to have equal amount of accepted and rejected waveforms. The results will be compared with traditional techniques.

Acknowledgements Thanks to the individuals who work hard to collect high quality seismic data and share it openly. Thanks also to the developers of GMT, Obspy, Matplotlib, Scikit-learn, Keras, and Google Earth. The facilities of the IRIS Data Services, and specifically the IRIS Data Management Center, were used for access to waveforms and related metadata required for waveform data. We thank helpful discussions with Kipton Barros, Singanallur Venkatakrishnan, and Derek Rose.

References

Cleveland, K. M., & Ammon, C. J. (2013). Precise relative earthquake location using surface waves. Journal of Geophysical Research: Solid Earth, 118(6), 2893-2904. https://doi.org/10.1002/jgrb.50146

Cleveland, K. M., & Ammon, C. J. (2015). Precise Relative Earthquake Magnitudes from Cross Correlation. Bulletin of the Seismological Society of America, 105(3), 1792-1796. https://doi.org/10.1785/0120140329

Cleveland, K. M., Ammon, C. J., & Kintner, J. A. (2018). Relocation of Light and Moderate-Magnitude (M 4-6) Seismicity Along the Central Mid-Atlantic. Geochemistry, Geophysics, Geosystems. https://doi.org/10.1029/2018GC007573

Kintner, J. A., Ammon, C. J., Cleveland, K. M., & Herman, M. (2018). Rupture processes of the 20132014 Minab earthquake sequence, Iran. Geophysical Journal International, 213(3), 1898-1911. https://doi.org/10.1093/gji/ggy085

Kintner, J. A., Wauthier, C., & Ammon, C. J. (2019). InSAR and Seismic Analyses of the 201415 Earthquake Sequence near Bushkan, Iran: Shallow Faulting in the Core of an Anticline Fold. Geophysical Journal International, 1011-1023. https://doi.org/10.1093/gji/ggz065