A Multi-Survey Dataset and Benchmark for First Break Picking in Hard Rock Seismic Exploration

Pierre-Luc St-Charles **Bruno Rousseau** **Joumana Ghosn** **Jean-Philippe Nantel**
Applied Machine Learning Research Team
Mila, Québec AI Institute
{firstname.lastname}@mila.quebec

Gilles Bellefleur **Ernst Schetselaar**
Geological Survey of Canada
Natural Resources Canada
{firstname.lastname}@nrcan-rncan.gc.ca

Abstract

Seismic surveys are a valuable source of information for mineral exploration activities. We introduce a reflection seismic survey dataset acquired at four distinct hard rock mining sites to stimulate the development of new seismic data interpretation approaches. In particular, we provide annotations as well as a sound benchmarking methodology to evaluate the transferability of supervised first break picking solutions on our dataset. We train and evaluate a baseline solution based on a U-Net and discuss potential improvements to this approach.

1 Introduction

The application of machine learning techniques and methodologies in geoscience and geophysics is an active research area that is becoming increasingly important [1]. Seismology in particular has a high potential for impactful contributions, as seismic data is often voluminous and hard to comprehensively analyze, even for experts. Recent efforts have been made to gather and study seismic datasets using machine learning [2–4], but the unavailability of data is a major impediment for researchers. This is especially true for reflection seismic data acquired on-land as part of long-term exploration and commercial mining activities. As a result, in this field, the growth of machine learning applications is slowed, and published works can rarely compare different interpretation methods without reimplementing them entirely.

Our primary goal is to provide a large reflection seismic dataset to the community in order to foster the development of new machine learning applications for deep mineral exploration. Specifically, we introduce a dataset of land seismic surveys captured across multiple mining sites, and provide a benchmark for the evaluation of solutions to a fundamental preprocessing problem in the analysis and interpretation of seismic data: first break picking [5]. To the best of our knowledge, this is the first public contribution of a curated multi-survey seismic dataset focused on hard-rock environments. The size of our dataset, which contains millions of seismic traces, will allow researchers to assess how well their predictive models generalize across varied conditions. Each survey is also acquired in 3D, meaning that seismic gathers can be analyzed in conjunction with 2D geopositioning data. This should encourage the development of a new generation of machine learning approaches that exploit the full richness of 3D surveys. Finally, we provide baseline evaluation results following our benchmarking protocol for popular first break picking approaches, and provide new ideas on how to interpret this data as well as how to incorporate prior knowledge to improve these baselines.

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

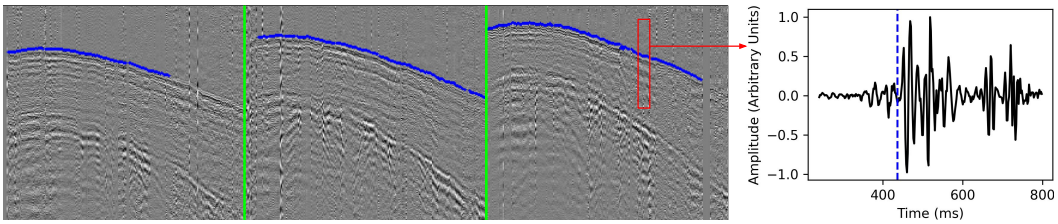

Figure 1: Example of a shot gather taken from a 3D seismic survey that is part of our dataset. The sections separated by green vertical lines correspond to changes from one receiver line to the next, and are named "line gathers". The dome-like pattern that repeats across all line gathers is typical; it is caused by the variation in distance between receivers and source location. For each trace, the location of the first arrival that should be "picked" by predictive models as part of our benchmark is drawn in blue. The amplitudes of a single trace (highlighted in red) near the first arrival are plotted on the right.

2 Data description

Reflection seismology is a method used for geophysics exploration that is based on the analysis of the propagation of elastic waves through solid or fluid materials. Here, we focus on land surveys where the seismic energy sources are explosive charges. These sources produce shockwaves that travel into the subsurface and that are affected by variations in acoustic impedance which correspond to changes in lithology. The measurement of the reflected and refracted shockwaves at the surface using receivers (or "geophones") is used for characterizing and imaging interfaces between different types of subsurface media. This imaging is fairly accurate in comparison with the results of other geophysical exploration techniques, as it offers a cell resolution in the order of a few meters. It can also cover surveys that stretch over tens of kilometers with reflection data up to several kilometers deep. Seismic data however requires many processing steps before being used for subsurface interpretation, one of which is first break picking, described in Section 3.

In its raw form, the seismic data we provide consists of energy measurements acquired following the activation of individual sources one at a time, where each activation is called a "shot". The series of measurements recorded by a receiver for a single shot is called a "trace", and the combination of traces across multiple receivers is called a "shot gather". Since receivers are usually arranged in straight lines across survey areas, shot gathers can be visualized as 2D images where the horizontal axis corresponds to the spatial extent of the receiver line and the vertical axis is time. In the case of 3D surveys, multiple receiver lines are set up in parallel, and shots are recorded across different lines simultaneously; an example of a shot gather for multiple receiver lines is shown in Figure 1.

Our proposed dataset is composed of four 3D surveys acquired at unique mining sites in three provinces of Canada: these are referred to as "Lalor", "Brunswick", "Halfmile" (short for "Halfmile Lake"), and "Sudbury". For one of the three sites (Lalor), the sampling rate of the receivers is 1 ms, and it is 2 ms for the three other sites. For Sudbury and Lalor, the traces contain 1001 samples, and for Halfmile and Brunswick, they contain 751 samples. Considering only the data that is usable for our proposed benchmark (described in Section 4), the surveys vary in size, with 690 to 1,541 shots and 8 to 28 receiver lines. When combined, they provide us with a total of 3913 shot gathers covering an average of 10.38 receiver lines, for a total of 8.37M traces. The spatial coordinates of the sources and receivers are also provided. Links to the raw data as well as additional statistics related to each survey dataset are available online[1].

Related works. Apart from the wiki-based list of [6], there does not seem to be a compendium or review of seismic datasets published in the literature. This is not surprising, as seismic surveys are rarely published due to their costly nature and due to the fact that they are mostly centered around private commercial projects. Besides, we must note that seismic datasets can be used for numerous tasks, and few datasets actually contain the raw ("pre-stack") data required to train models on tasks such as first break picking. For example, the benchmarks of [3, 7] focus on facies classification and only provide a labeled seismic volume to be used for training and evaluation. Besides, even fewer public datasets are land acquisitions, and the only ones listed in [6] ("Teapot dome", "Stratton") are oilfield surveys. This means that our contribution to the literature of four new land surveys over crystalline hard-rock terrain is quite unique.

[1] `https://github.com/mila-iqia/hardpicks`

3 First break picking

The workflow required to turn raw seismic measurements into interpretable information regarding subsurface interfaces is complex and involves multiple labor intensive steps by domain experts [8]. One of the first steps of this workflow is dubbed "first break picking" and consists in indicating where the useful signal begins for every recorded trace. This requires distinguishing the onset of the arriving seismic wave from the omnipresent background noise. Automated tools have long been used to assist annotators in identifying first breaks based on trace statistics [5]. However, these tools perform poorly on noisy traces due to their lack of contextual awareness. First break picking is labor intensive, tedious, error-prone, and sometimes ambiguous, as shown in Figure 1. Correspondingly, manual picking can vary in quality and affect the outcome of downstream data processing. This task is therefore ideal for the development and testing of supervised machine learning approaches.

Related works. The interpretation of shot gathers as images, as shown in Figure 1, is quite convenient for the training of predictive models based on Convolutional Neural Networks (CNNs). In fact, the majority of recently published works on first break picking have opted for this approach [9–14]. Small CNNs have also been specifically shown to be more robust than other simple network architectures when trained on the gathers of a survey of an ore deposit [15]. For works that relied on larger CNNs to increase the amount of contextual information available for picking, U-Net architectures [16] seem to be the most popular and successful design choice [11–14]. This highlights the prominent position of image processing approaches for first break picking, and entails that we should train and provide a CNN as a baseline in our proposed benchmark. We must however stress that there is no evidence that non-CNN approaches cannot compete with CNNs. We hope that the introduction of our multi-survey dataset will allow new types of models to be designed and trained, and these would ideally take the available geopositioning data into full consideration.

4 Benchmark methodology

Our benchmark is centered on the supervised training and evaluation of predictive models for the task of first break picking. Along with our multi-survey dataset, we provide first break pick annotations for 72.3% of traces in line gathers that are deemed valid by an expert based on both signal strength and annotation quality. The annotations themselves are based on the predictions of simple picking tools (based e.g. on STA/LTA ratios) that are then manually corrected by experts.

The main driver for our experiments is to determine whether trained models can generalize their knowledge across survey sites. To reach such conclusions, we split the sites at our disposal into different cross-validation folds. We believe that this is the ideal way to split multi-survey data, as shuffling and splitting across the gathers of all surveys may lead to misleading results if the trained models manage to overfit to the spatial or subsurface characteristics of particular regions. We elected to use two sites in the training set, one site in the validation set and one site in the test set. To limit the computational requirements of our experiments, we restricted this to four folds by requiring that each site appears once in validation and once in test. We refer to these folds below using the name of the sites used for validation and testing, and provide the full description of these folds on our website (the link is provided in Section 2).

For the performance metrics, we consider the *Mean Average Error* (MAE), the *Mean Bias Error* (MBE), and the *Root Mean Squared Error* (RMSE), where all errors are measured at the sample scale (i.e., in terms of "pixels" in the gather images). We also rely on the *Hit Rate at δ samples* (HR@δpx), defined as the fraction of annotated traces where the prediction error is smaller than δ samples. For a given dataset fold, we conduct a random hyperparameter search over 50 trials. Each trial, we train on two sites and evaluate on a third (the validation site). After 50 trials, the hyperparameter configuration leading to the model with the highest validation HR@1px score is withheld, and 10 models are trained once again with this configuration, starting from different random seeds. These 10 models are finally evaluated on the test site to give a range of performance.

The baseline we propose is based on a U-Net architecture [16] where the encoder and decoder are composed of stacks of fully convolutional blocks. The input to the model is a line gather which is treated as an image with multiple channels; its first channel contains the seismic amplitude, and it is complemented with three more channels containing the source-receiver distance and the distances to the closest previous and next receivers on the same line. These distance channels are constant

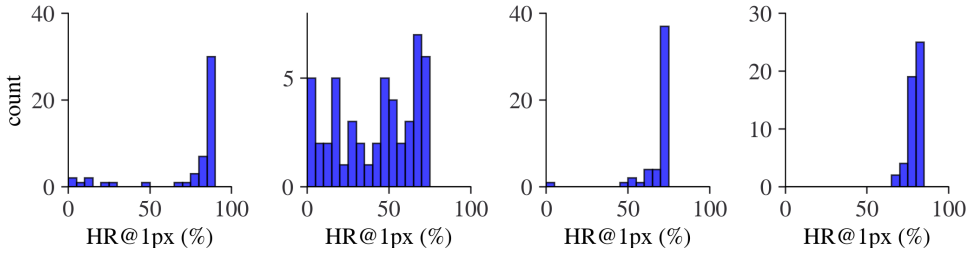

Figure 2: Distribution of HR@1px scores obtained on the validation site of each fold across 50 trials. From left to right, the histograms correspond to Brunswick, Lalor, Sudbury, and Halfmile. Note the more inconsistent results on Lalor, which is the only site with a 2 ms sampling rate.

Table 1: Metrics computed on the test site of each fold for the trial with the highest HR@1px metric found on the validation site. Ten runs with different random seeds were performed, and the mean plus-or-minus the standard deviation over these runs is shown. The *Hit Rates* (HR) are in percentage, and the errors are in number of samples.

Test Site	HR@1px	HR@3px	HR@5px	HR@7px	HR@9px	RMSE	MAE	MBE
Sudbury	73.1 ± 0.5	93.9 ± 0.6	96.2 ± 0.6	97.5 ± 0.5	98.2 ± 0.5	35.1 ± 12.3	2.8 ± 1.2	1.2 ± 1.1
Brunswick	87.6 ± 1.4	96.4 ± 0.6	97.8 ± 0.6	98.3 ± 0.6	98.6 ± 0.6	50.2 ± 13.8	4.5 ± 2.3	3.8 ± 2.5
Halfmile	83.8 ± 0.5	92.6 ± 0.5	95.9 ± 0.6	97.9 ± 0.6	98.8 ± 0.6	35.2 ± 33.3	3.8 ± 4.3	2.9 ± 4.1
Lalor	76.3 ± 1.8	80.0 ± 1.7	82.7 ± 1.7	86.4 ± 1.9	89.0 ± 2.0	460 ± 74	124 ± 40	123 ± 40

along the time axis, and act as simple geospatial priors that the model can exploit. We use four data augmentation operations during training. Firstly, gather images are randomly cropped so that they have between 512 and 1024 samples on their time axis. Secondly, some traces inside each gather may be randomly dropped (as if there was a gap in the receiver line) or added with null amplitudes (as if an extra receiver was present but "dead"). Thirdly, we nullify the amplitudes of roughly 8% of all traces in the gathers: this forces the model to rely more on contextual information. Finally, we flip gather images along their receiver axis in order to increase the (perceived) diversity of the datasets. Other augmentation operations were originally considered, but their impact was later found to be marginal. The parameter values for the selected operations were found empirically on a small subset of the available data before conducting the large-scale hyperparameter search.

For the optimization, we rely on Adam [17] and pick base learning rates uniformly across a logarithmic scale of $[10^{-5}, 5 \cdot 10^{-3}]$. We train for a maximum of 20 epochs and either never modify the learning rate, or reduce it by multiplying it with a factor of 0.1 after either 5 or 10 epochs. The batch size is fixed for all experiments at 16 line gathers. Early stopping is performed if the HR@1px on the validation site does not improve for more than 4 consecutive epochs. For the encoder, we evaluate ResNet blocks [18] with two different depths (18 and 34 total layers), and EfficientNet blocks [19] under their "b0", "b2", and "b4" configurations. For the decoder, we explore three different levels of complexity by scaling the number of feature maps carried over from the encoder. Specifically, we use $[256, 128, 64, 32, 16]$, $[512, 256, 128, 64, 32]$, or $[1024, 512, 256, 128, 64]$ feature maps, where each number corresponds to the input depth for each of the five decoder blocks. Note that across all potential hyperparameter configurations, our smallest model has roughly 14M trainable parameters while the biggest model has 49M.

Figure 2 shows the distribution of model performance found on different validation sites during our hyperparameter search, showing that many configurations can perform similarly well. The performance of the best trial for each fold computed on the test site is presented in Table 1. We can observe that performance on the Lalor test site is worse than for the test sites of other folds. This could be explained by the different sampling rate used at Lalor which the predictive model did not experience during training and validation. This shows that the transferability of a simple U-Net can still be improved.

5 Conclusion

We present a multi-site dataset of raw seismic traces and human generated annotations as well as benchmark results for first break picking based on the U-Net architecture. By distributing this dataset, we seek to promote the use of a standard evaluation methodology for trained models and to promote the design and development of new models specifically tailored to process seismic data. In particular, it is unclear whether the interpretation of seismic data as images is ideal, and hope researchers will deviate from this approach if needed. In fact, there may already exist architectures designed for graph or point cloud data processing (e.g., [20]) that could be better suited for first break picking.

Broader impact

We present a multi-survey dataset as well as a rigorous methodology for evaluating first break picking algorithms. Our hope is that these contributions will drive the development of better models which will help seismologists and geophysicists improve their toolset for the processing and interpretation of seismic data. In the short term, this may save them a substantial amount of manual labor (or significant outsourcing costs) when working with new surveys. Our efforts could also incentivize some researchers to revisit older surveys with modern tools, and in the long term, revitalize existing mining sites due to the discovery of new economical ore deposits. Our efforts to improve first break picking solutions could also have indirect impacts on applications involving statics correction, velocity inversion, traveltime tomography, and hazard assessment.

Acknowledgments

We thank E. Adam, S. Cheraghi, and A. Malehmir for providing first break picks for the 3D seismic surveys. We thank First Quantum Minerals, Glencore, and Trevali Mining for providing access to the field seismic data.

References

[1] Anuj Karpatne, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1544–1554, 2018.

[2] Vincent Dumont, Verónica Rodríguez Tribaldos, Jonathan Ajo-Franklin, and Kesheng Wu. Deep learning on real geophysical data: A case study for distributed acoustic sensing research. *arXiv preprint arXiv:2010.07842*, 2020.

[3] Yazeed Alaudah, Patrycja Michałowicz, Motaz Alfarraj, and Ghassan AlRegib. A machine-learning benchmark for facies classification. *Interpretation*, 7(3):SE175–SE187, 2019.

[4] Fabrizio Magrini, Dario Jozinović, Fabio Cammarano, Alberto Michelini, and Lapo Boschi. Local earthquakes detection: A benchmark dataset of 3-component seismograms built on a global scale. *Artificial Intelligence in Geosciences*, 1:1–10, 2020.

[5] Françoise Coppens. First arrival picking on common-offset trace collections for automatic estimation of static corrections. *Geophysical Prospecting*, 33(8):1212–1231, 1985.

[6] Society of Exploration Geophysicists (SEG). Open data compendium, 2021. Wiki page available online at `https://wiki.seg.org/wiki/Open_data`.

[7] Dimitri Bevc, Adam Halpert, Felix Herrmann, Bruce Power, Cengiz Esmersoy, and Sergey Fomel. SEG machine learning interpretation challenge, 2020. Dataset and description available online at `https://public.3.basecamp.com/p/JyT276MM7krjYrMoLqLQ6xST`.

[8] Edip Baysal, Dan D Kosloff, and John WC Sherwood. Reverse time migration. *Geophysics*, 48(11):1514–1524, 1983.

[9] Sanyi Yuan, Jiwei Liu, Shangxu Wang, Tieyi Wang, and Peidong Shi. Seismic waveform classification and first-break picking using convolution neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(2):272–276, 2018.

[10] Lionel J. Woog, Anthony Vassiliou, and Rodney Stromberg. Acquisition/Processing: AI-complemented first-break picking for field low-S/N seismic data. *The Leading Edge*, 40(6):460–463, 2021.

[11] Yuanyuan Ma, Siyuan Cao, James W Rector, and Zhishuai Zhang. Automated arrival-time picking using a pixel-level network. *Geophysics*, 85(5):V415–V423, 2020.

[12] Chiel Fernhout, Paul Zwartjes, and Jewoo Yoo. Automatic first break picking with deep learning. *IOSR Journal of Applied Geology and Geophysics*, 8(5):24–36, 2020.

[13] Pengyu Yuan, Shirui Wang, Wenyi Hu, Xuqing Wu, Jiefu Chen, and Hien Van Nguyen. A robust first-arrival picking workflow using convolutional and recurrent neural networks. *Geophysics*, 85(5):U109–U119, 2020.

[14] Jing Zheng, Jerry M. Harris, Dongzhuo Li, and Badr Al-Rumaih. SC-PSNET: A deep neural network for automatic P- and S-phase detection and arrival-time picker using 1C recordings. *Geophysics*, 85(4):U87–U98, 2020.

[15] Tasman Gillfeather-Clark, Tom Horrocks, Eun-Jung Holden, and Daniel Wedge. A comparative study of neural network methods for first break detection using seismic refraction data over a detrital iron ore deposit. *Ore Geology Reviews*, page 104201, 2021.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[20] Chaoyun Zhang, Marco Fiore, Iain Murray, and Paul Patras. CloudLSTM: A recurrent neural model for spatiotemporal point-cloud stream forecasting. *arXiv preprint arXiv:1907.12410*, 2019.