What is our a priori information? Generating representative earth models for ML training

Mark Roberts*, Olga Brusova, Keyla Gonzalez, Sean Crawley and Alejandro Valenciano
TGS

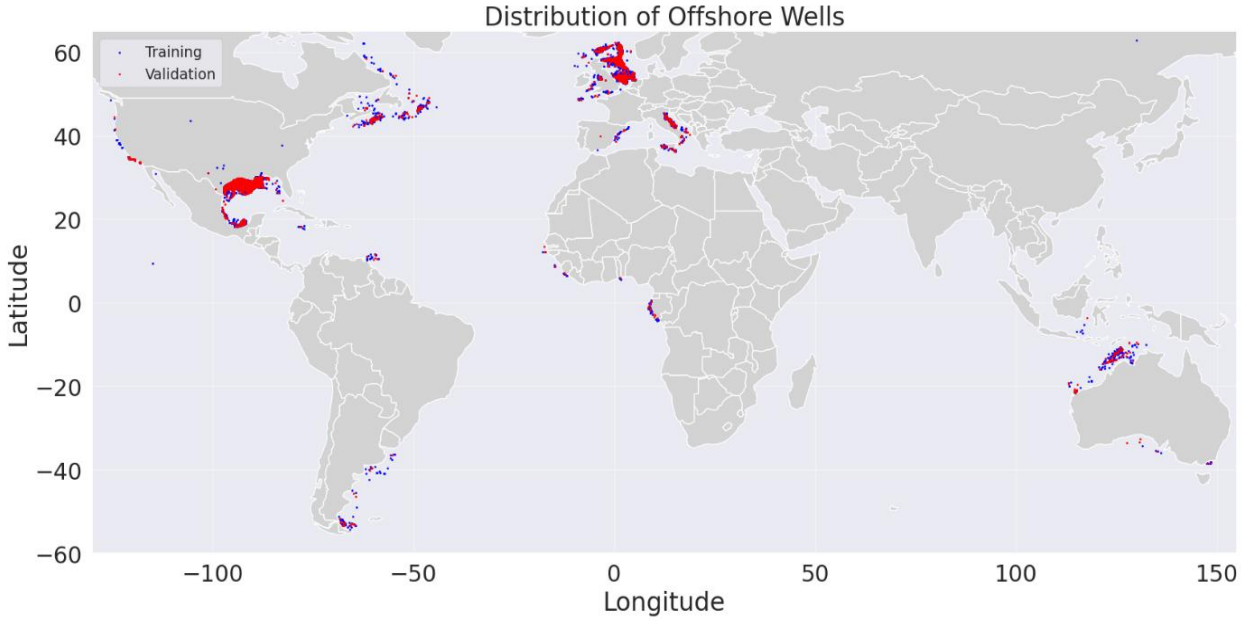
Abstract

This study presents a novel workflow for generating realistic 3D Earth representations tailored to machine learning (ML)-based velocity estimation tasks. By leveraging a comprehensive global dataset of 60,090 offshore wells integrated with seismic structural attributes, we address critical limitations in current synthetic datasets, which often lack representative geological complexity. Our methodology features three significant innovations: (1) An ML-driven approach that robustly predicts missing shear wave sonic logs, substantially broadening training data coverage and quantifying associated uncertainties; (2) Structural modeling through predictive painting that accurately extracts relative geological time from seismic data, ensuring geological consistency; and (3) A hybrid augmentation technique that combines pseudo-synthetic earth models with realistic salt masks and detailed lithological trends, significantly enhancing geological realism. The resulting training datasets effectively capture the diversity and complexity encountered in real-world scenarios, laying the groundwork for improving the robustness, accuracy, and generalizability of ML models used in full waveform inversion (FWI) and tomographic velocity modeling.

Introduction

Recent advances in machine learning can potentially transform geophysical workflows, particularly in velocity model building. ML-based tomographic updates (Crawley et al., 2024) demonstrate enhanced velocity estimation through learned data relationships, while diffusion-based approaches (Taufik et al., 2024) show promise for regularizing full waveform inversion (FWI). However, these techniques are critically dependent on training data that adequately represent real-world geological variability-a requirement not met by current synthetic datasets like OpenFWI (Deng et al., 2022), which lack realistic a priori earth model distributions. We address this limitation through three interconnected innovations: First, we extend the ARLAS methodology (Gonzalez et al., 2023) to predict shear wave logs (including associated uncertainties) across a global well dataset, effectively addressing data scarcity issues. Second, we integrate seismic structural information using predictive painting (Fomel, 2010) to create geologically consistent 3D models. Third, we develop a hybrid augmentation approach that combines real salt masks with pseudo-synthetic structural models, overcoming challenges in complex geological settings.

This paper first details our machine-learning framework for log prediction and uncertainty quantification. We then present our seismic-to-model conversion workflow using relative geological time attributes. Finally, we demonstrate how synthetic model augmentation enhances training data diversity while maintaining geological realism. The resulting methodology provides a robust foundation for ML-based velocity estimation in data-rich and challenging subsurface environments.



Method - Wells: An ML model that predicts missing shear wave sonic logs

Our dataset comprises 60,090 offshore wells from various basins worldwide. The locations of the wells are illustrated in Figure 1. The dataset features an uneven distribution across global regions, with the Gulf of Mexico representing the majority (88%). While nearly half of the wells have acoustic measurements, only 5% include shear wave information. The lack of shear wave data highlights the necessity for reliable prediction methods. We use 75% of the data for model training and 25% for model validation. Training and validation data are randomly sampled from different global basins. Figure 1 illustrates the distribution of training and validation wells globally.

Before machine learning training, the raw well-log data underwent thorough cleaning and quality control processes. These steps were essential for ensuring data accuracy and consistency and for reliable model training and inference. The clean-up workflow included automated tasks such as log categorization, depth alignment, normalization, and removing low-quality data. A key innovation in the clean-up process is the introduction of synthetic curve comparisons for sonic log classification. This method uses Faust and Castagna's equations to calculate synthetic compressional and shear slowness logs from resistivity data. Comparing actual sonic logs to these synthetic models offers a robust and automated approach for classifying sonic curves, even when log names and descriptions are unreliable.

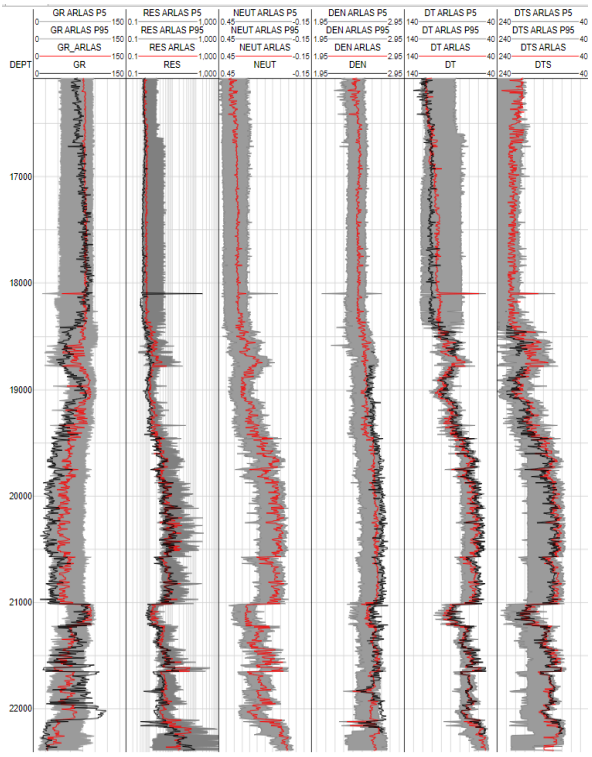


Figure 2 Example of ARLAS predictions (red curves) as compared to measured data (black curves). The plot shows: (1) gamma ray, (2) deep resistivity, (3) neutron porosity, (4) bulk density, (5) compressional slowness, (6) shear slowness. P5-P95 interval in grey.

The core of the ARLAS methodology (Gonzalez et al., 2023) is the Gradient Boosting Tree (GBT) machine learning model. GBT utilizes an ensemble learning approach that creates a predictive model by sequentially combining multiple decision trees and iteratively reducing the error between predictions and actual values. This technique excels at handling complex data patterns and achieving high prediction accuracy.

The ARLAS model utilizes various features, including compressional sonic logs, gamma-ray logs, resistivity logs, neutron porosity, bulk density, true vertical depth, and well-location data. This multi-dimensional approach enables the model to capture complex relationships between log properties and their spatial variations. Here, we extend the method to include shear logs and offshore wells. The trained ARLAS model was evaluated on a held-out dataset, demonstrating good alignment with measured data in areas with sufficient training data (Figure 2). However, challenges arise in shallow sections and regions with limited data availability, underscoring the need for further model refinement. The P5 and P95 curves (representing the 5th and 95th percentiles of the model predictions) were generated to assess model uncertainty and reliability. These bounds provide a range in which the true value is expected to fall with 90% confidence. Significant discrepancies between P5 and P95 indicate potential overfitting and decreased prediction confidence.

Method - 3D model generation

The method presented here for extracting structural information from seismic data is based on the work of Fomel (2010). "Plane-wave destruction" utilizes the concept of local plane waves, represented as simple linear equations that depict seismic events with varying slopes. This technique entails predicting each trace in a seismic section from its neighbor based on estimated slopes and subtracting the prediction from the original trace, yielding a residual that represents non-planar wave components. A local operator is employed to propagate each trace along the estimated dominant slopes, which are determined by minimizing the prediction residual through regularized least-squares optimization. The technique can also be extended to three dimensions by applying it to both inline and crossline directions, as is done here. "Predictive painting" then advances this concept further, facilitating the recursive propagation of a reference trace to distant neighbors, while following the local structure of seismic events, effectively "painting" the information from the reference trace onto neighboring traces. This process enables a more comprehensive understanding of the seismic data structure.

By applying predictive painting to a reference trace that contains only time values, we can derive an attribute called

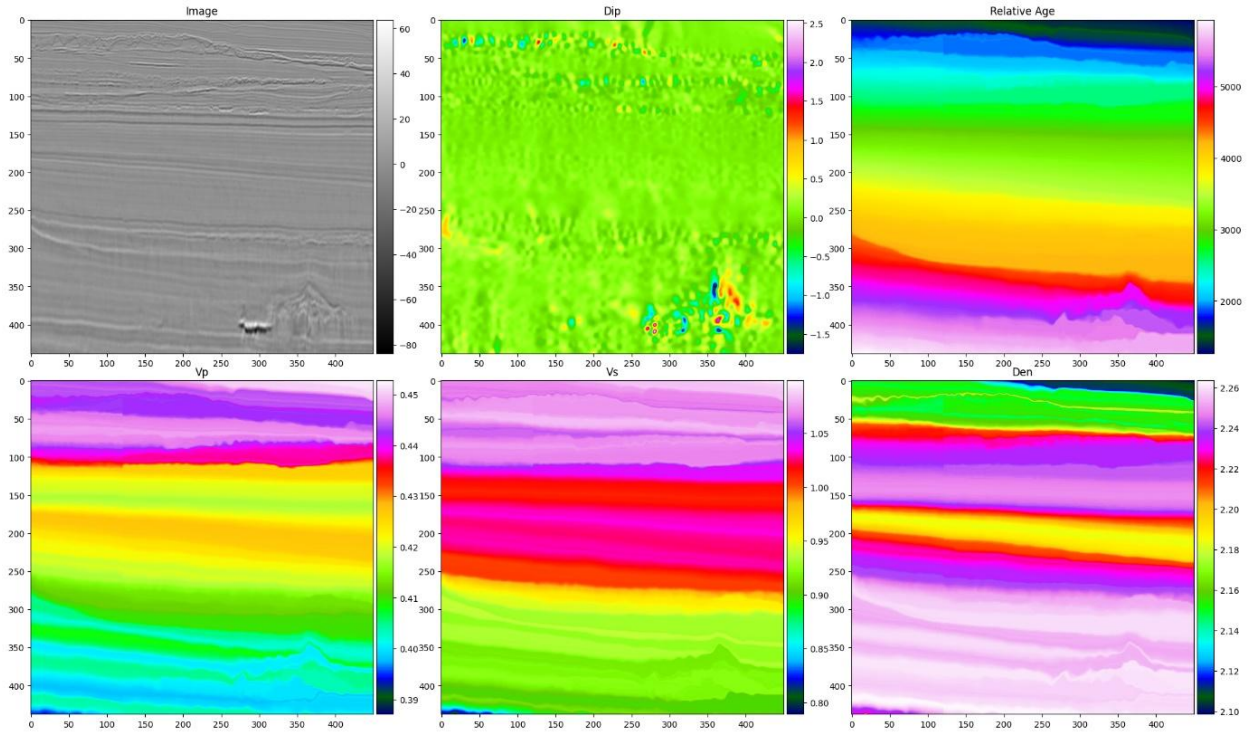


Figure 3 This shows the progression from a stacked seismic image to a representative Vp, Vs, and density volume. Top row: seismic image, estimated dip, relative geological time. Bottom row: Vp, Vs, and density volumes.

"relative geologic age." This attribute, as defined by Stark (2004), indicates the time shift between a given trace and the reference trace, effectively showing how much older or younger the geology at that trace is in relation to the reference point.

Finally, a representative 3D elastic models can be generated based on a relative geological time volume and an ARLAS well-log. Several QCs are made to ensure the well log has a long enough section of usable data to generate a meaningful representative model (length, absence of salt). If selected, the logs are filtered with a smoothing operator (the smoothing length is also randomly chosen) to generate models with different frequency content. Figure 3 outlines the complete flow for one seismic image. The top left shows a patch of the image taken from the data library, followed by the structural dip and resulting relative age volume. The bottom row contains the models taken after a lookup from the well data.

Method - 3D pseudo-synthetic model augmentation

While the approach outlined above works well in well-image areas, relative-geological-time estimates can become challenging in more complex areas such as around salt. As a result, a complementary approach has been taken, combining pseudo-synthetic structural models. Synthoseis (Merrifield et al., 2022) is a package designed to generate realistic and diverse synthetic 3D seismic models for training deep learning applications in geophysics, addressing the need for large quantities of labeled data. Synthoseis can generate different faulting styles and simple salt bodies, with the resulting models including 3-way and 4-way closures and on-lap (stratigraphic) closures. It generates layered earth models with shale and sand layers with varying net-to-gross ratios and can model fan-shaped features in the layers. However, the package has limitations, including simplistic lithological trends and blob-like salt models. In this paper, we have modified Synthoseis to randomly look up real interpreted salt masks from a data library (also used for Salt Segmentation training; see Roberts et al., 2024 for details). To provide a greater diversity of lithological trends, a substitution has been performed to utilize observed properties from the actual well logs. Another advantage over the previous approach is that we have integrated with the Synthoseis lithological trends for the very shallow and deep sections where there is a lack of actual well-log data. The results of this workflow can be seen in Figure 4.

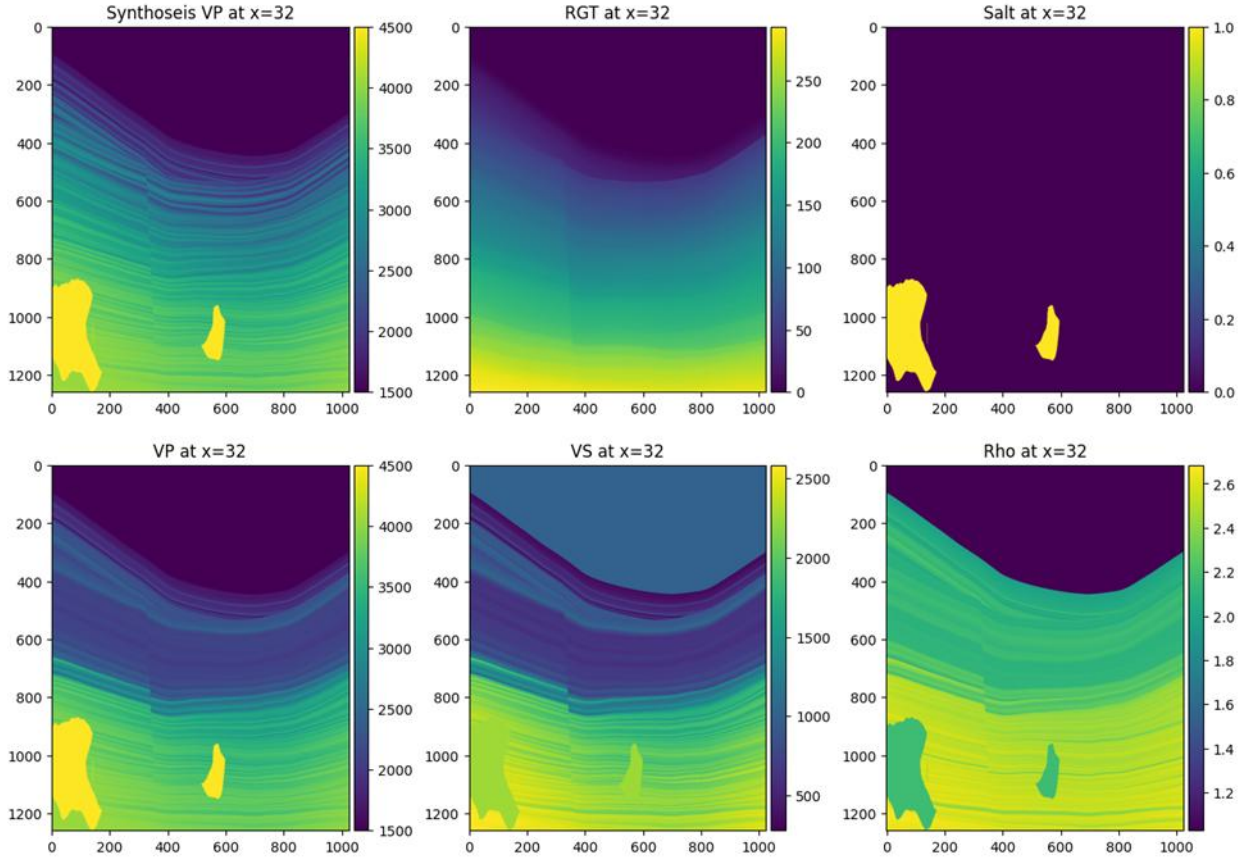


Figure 4 An example of a representative Vp, Vs, and density volume generated from a pseudo-synthetic structural model. Top row: Synthoseis Vp model, relative geological time, salt mask. Bottom row: Vp, Vs, and density volumes incorporating well-log information.

Discussion and Future Work

In this abstract, we demonstrate how to leverage a vast library of seismic and well-log data to generate diverse earth models that reflect the petrophysical and structural diversity observed in the earth. No single approach adequately represents the real-world geological variability of earth models; therefore, combining the discussed methods is necessary. Well-log data is utilized wherever possible but supplemented in shallow and deep sections by petrophysical trends. In areas with high image quality, the structure is estimated from seismic images and enhanced with synthetics to address complex visuals. Interpreted salt masks are used to augment synthetic structures, providing realistic geobodies. These diverse models will be employed to train an ML-based tomographic model (Crawley et al., 2024) and enhance its generalizability and robustness to unseen scenarios.

Conclusions

This work presents a comprehensive strategy that combines advanced machine learning techniques, seismic structural analysis, and hybrid synthetic augmentation to generate realistic earth models. The ARLAS model effectively predicts missing shear wave sonic logs with quantified uncertainty, addressing critical data gaps common in global well datasets. Our seismic-to-model workflow, utilizing predictive painting and relative geological time attributes, enables the direct translation of seismic structures into accurate elastic models. Furthermore, integrating pseudo-synthetic earth models with real-world salt masks and lithological information effectively enhances geological realism, particularly in structurally complex areas. These combined innovations significantly improve the representativeness and diversity of training data, which will translate into substantial benefits for ML-based tomographic updates and regularizing Full Waveform Inversion (FWI) models.

Acknowledgements

We are thankful to TGS management for permission to show the data. The authors thank Burak Bitlis and Seet Li Yong for their help in preparing the seismic data.