Seismic Activity and Oil Production Analytics

Shinelle Barretto Abigail Bradshaw Xavier Mares Charlie Tran

Abstract

Hydraulic fracturing (fracking) has transformed the oil and gas industry. It has also established the United States of America as one of the top leaders in energy production. However, if sustainability is not prioritized, this process has also been known to pose environmental risks, such as induced earthquakes. This project assesses the environmental impact of fracking and optimizes well site selection by balancing production efficiency and environmental risks. Traditional supervised learning models such as Support Vector Machine (SVM) and Random Forest Classifier (RFC) were used to perform as a baseline performance metric to evaluate how well machine learning could predict optimal well locations. These were then compared to deep learning models such as Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) to compare respective model predictive capabilities and their accuracies. The RNN model proved the most effective, offering high accuracy in optimizing well placement while mitigating seismic risks.

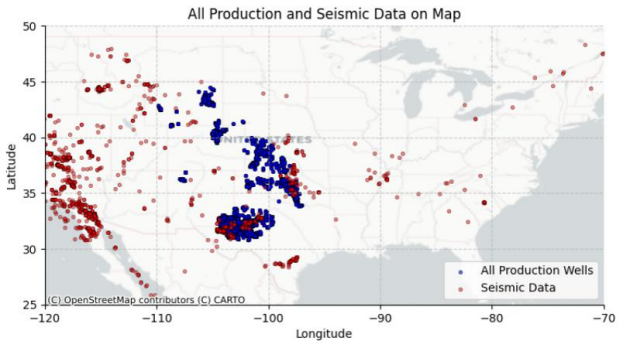


Fig 1. All Production and Seismic Data on Map

Introduction

The oil and gas industry has been a cornerstone of economic growth since the mid-19th century, beginning with the first commercial oil well in Pennsylvania in 1859. Technological advancements, such as hydraulic fracturing (fracking) and horizontal drilling, have dramatically transformed the sector. Introduced in the 1940s, fracking involves injecting high-pressure fluids into underground rock formations to release trapped oil and gas, unlocking vast reserves of previously inaccessible resources. This has positioned the United States as a global leader in energy production.

One of the primary strategies in well placement is offset drilling, where new wells are placed near existing, successful ones. This approach leverages known geological formations, reducing initial risks and maximizing efficiency. However, offset drilling inherently limits exploration to areas close to productive sites and fails to identify untapped optimal locations. Moreover, it can exacerbate environmental risks, such as induced seismicity (man-made earthquakes), if placed in seismically sensitive regions.

The environmental challenges of fracking are significant. For example, between 2010 and 2015, Oklahoma experienced a dramatic increase in earthquakes due to wastewater injection wells associated with fracking. Public awareness of these risks has led to growing regulatory scrutiny and a demand for more sustainable and responsible practices in well placement.

Historically, oil and gas well placement relied on geological surveys, historical production data, and expert judgment. While effective, these traditional methods often lack the precision and adaptability needed to address modern extraction challenges. Environmental risks and the need for

sustainability require more innovative and data-driven approaches.

This project aims to address these issues by optimizing well placement using machine learning. By integrating production data—quantifying annual liquid, gas, and water outputs—and seismic activity data, we developed a comprehensive scoring system. This system evaluates the trade-off between maximizing production and minimizing environmental risks, providing actionable insights for sustainable and safer well placement, even in offset drilling scenarios.

To achieve these objectives, we implemented four machine learning models:

Support Vector Machines (SVM): Known for robust classification and regression performance, SVM provided a baseline but required extensive parameter tuning and preprocessing.

Random Forest: An ensemble learning method that effectively handled non-linear relationships while offering insights into feature importance, such as identifying the dominant impact of liquid production.

Long Short-Term Memory Networks (LSTM): A variation of RNN, LSTM proved adept at handling temporal dependencies in seismic activity but required higher computational resources.

Recurrent Neural Networks (RNN): Ideal for modeling sequential data, RNN excelled in capturing spatial dependencies between production and seismic events.

This data-driven approach represents a critical step toward sustainable resource extraction, offering the potential to redefine well placement strategies in the oil and gas industry. By identifying locations that balance productivity and environmental safety, the project aligns with the industry's growing focus on sustainable and responsible practices.

Related Work

Predicting suitable locations for hydraulic fracturing (fracking) has been a focus of machine learning research, aiming to optimize resource extraction while minimizing risks. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models are effective for analyzing time-series data, such as seismic activity and fluid injection rates, capturing complex temporal dependencies. Support Vector Machines (SVMs) excel in classifying geological formations, offering robust performance with high-dimensional data. Random Forests provide valuable insights through feature importance analysis and classification, identifying key geophysical factors like porosity and depth. This project combines these methods to develop a comprehensive framework for fracking site prediction, leveraging the unique strengths of each model to improve accuracy and decision-making.

Objectives

Optimize Well Placement:
Identify the most productive well locations that also demonstrate high environmental safety, ensuring a balance between maximizing resource extraction and minimizing environmental risks.

Develop a Comprehensive Scoring System:
Create a scoring framework that integrates production data (annual liquid, gas, and water outputs) with environmental data (seismic activity, magnitude, and proximity) to rank well locations effectively.

Leverage Machine Learning Models:
Implement and compare machine learning algorithms such as Support Vector Machines (SVM), Random Forest, Long Short-Term Memory (LSTM) and Recurrent Neural

Networks (RNN) networks to analyze and predict optimal well sites.

Enhance Decision-Making:
Provide actionable insights that improve the industry's ability to balance economic productivity with environmental stewardship, moving beyond traditional offset drilling methods.

Address Environmental Concerns:
Mitigate risks associated with fracking, such as induced seismicity and groundwater contamination, by prioritizing environmentally safer locations.

Enable Sustainability:
Contribute to the industry's shift toward sustainable practices by adopting data-driven methodologies that align resource extraction with safety and regulatory standards.

Evaluate Model Performance:
Compare the performance of various machine learning models to determine which algorithm best balances predictive accuracy, efficiency, and scalability for this specific application.

Propose Future Improvements:
Lay the groundwork for incorporating additional factors, such as water contamination risks and greenhouse gas emissions, and expand the model's applicability to diverse geographic and geological conditions.

Proposed Method

The methodology for this project involves a systematic approach to data collection, preprocessing, scoring, and machine learning model implementation to predict optimal fracking locations.

Data Collection: Two datasets were utilized to develop the prediction model, Production data and Seismic data. The Production dataset includes information about annual liquid, gas, and water production from active wells. Additional attributes, such as well location, drilling direction, and production status, were also included to provide context for site productivity. The Seismic dataset captures seismic events in terms of their geographical coordinates (latitude and longitude) and magnitudes. Higher magnitudes were considered indicative of greater environmental risk, providing crucial information for assessing the environmental impact of fracking sites.

Data Preprocessing: To ensure the datasets were suitable for machine learning, the following preprocessing steps were applied. First we cleaned the production data. Non-numeric attributes, such as drilling direction and production status, were standardized. For example, drilling directions were mapped to numeric values ("Horizontal" as 1 and "Vertical" as 0). Only active wells were considered, as inactive sites were excluded from the scope of the project. Next we handled missing values. Missing data points in both datasets were addressed using the SimpleImputer technique, which replaced missing values with the mean of the respective columns. Then we merged datasets, production and seismic datasets were merged based on proximity. The distance between seismic events and wells was calculated using Euclidean distance, ensuring that seismic activity relevant to each well was accurately associated. Finally we did feature scaling. Numerical features, such as production metrics and seismic magnitudes, were normalized to ensure they had comparable scales for effective model training.

Scoring System: A scoring system was developed to balance productivity and environmental risks. Two separate scores were calculated for each well, Production score and Envoirnmental score. The production score was computed as a weighted sum of the normalized annual production metrics. Liquid, gas, and water production were assigned weights of 0.5, 0.3, and 0.2, respectively, reflecting their relative

contributions to overall productivity. The environmental score was inversely proportional to the magnitude of seismic activity near each well. Wells located closer to high-magnitude seismic events received lower environmental scores, reflecting higher risks. The final Combined Score for each well was calculated as the product of the Production Score and Environmental Score, providing a comprehensive measure that balances production potential with environmental impact.

Model Training: The project employed four machine learning models to predict the Combined Score and rank well locations, Support Vector Machines (SVM), Random Forest, RNN, and LSTM.

Support Vector Machines (SVM): SVM was used as a baseline model to classify wells based on their scores. It demonstrated strong performance on smaller datasets but required significant parameter tuning to achieve optimal results.

Random Forest: This ensemble learning model was employed to handle non-linear relationships and to identify the key features influencing the scores. Random Forest provided valuable insights into feature importance, aiding the interpretability of results.

Recurrent Neural Networks (RNN): The RNN model was designed to analyze sequential data, making it particularly suitable for capturing temporal dependencies in seismic activity. This model effectively accounted for how seismic risks evolve over time.

Long Short-Term Memory (LSTM): As an extension of RNNs, LSTM networks were implemented to manage long-term dependencies in the data. Their ability to handle complex temporal patterns made them highly effective for predicting Combined Scores and identifying optimal well placements.

Each model was trained and validated using an 80:20 train-test split. The deep learning models (RNN and LSTM) utilized scaled and reshaped data to align with their input requirements. Hyperparameter tuning was conducted for all models to maximize performance.

Experiments

The experiments began with preparing the datasets for training and evaluation. The combined dataset, consisting of production and seismic data, was split into training and testing subsets using an 80:20 ratio. This split ensured that the models were trained on a substantial portion of the data while reserving a segment for unbiased evaluation. Preprocessing steps included scaling numerical features to standardize the range of values, which is essential for ensuring that machine learning models perform optimally. Sequential data, such as the temporal patterns in seismic activity, were reshaped into three-dimensional arrays required for deep learning models like Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks.

Evaluation Metrics: To assess the performance of each model, multiple evaluation metrics were employed. For regression tasks, metrics included Mean Squared Error (MSE). These metrics were chosen because they provide insight into the models' accuracy and ability to generalize to unseen data. MSE quantifies the average squared difference between predicted and actual values.

Training the Models: Each of the four machine learning models—Support Vector Machines (SVM), Random Forest, RNN, and LSTM—was trained using the processed dataset. SVM was configured as a baseline model due to its simplicity and interpretability. Hyperparameter tuning was conducted to optimize the SVM's kernel and

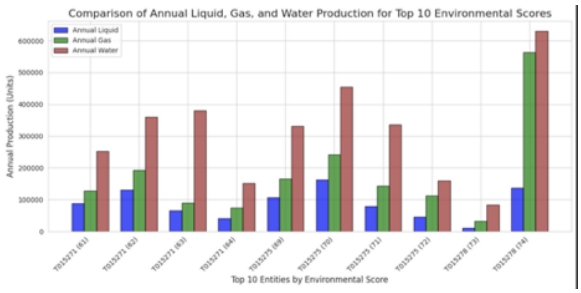
regularization parameters. Random Forest was trained using a combination of decision trees, with the number of trees and depth being optimized to prevent overfitting. For RNN and LSTM models, the architecture included multiple layers to handle temporal dependencies, with hyperparameters such as the number of neurons, dropout rates, and learning rates adjusted to achieve the best performance. The Adam optimizer was employed for both RNN and LSTM models, using Mean Squared Error as the loss function. Each model was trained for up to 100 epochs, and early stopping was implemented to prevent overfitting during training.

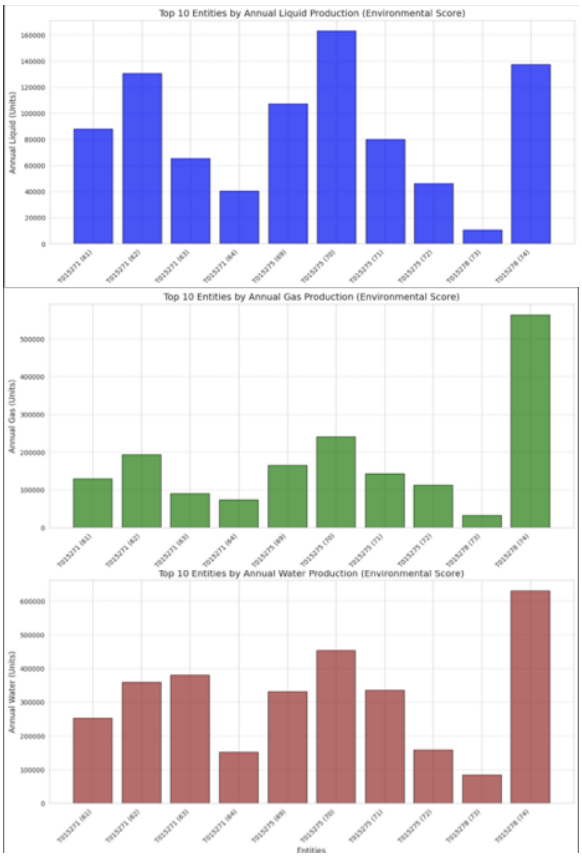
Visualization and Insights

To better understand the models' performance and the data distribution, various visualizations were created. Geospatial scatterplots displayed the geographical distribution of wells and seismic events, providing a clear depiction of production and risk hotspots.
Seismic data on top of wells

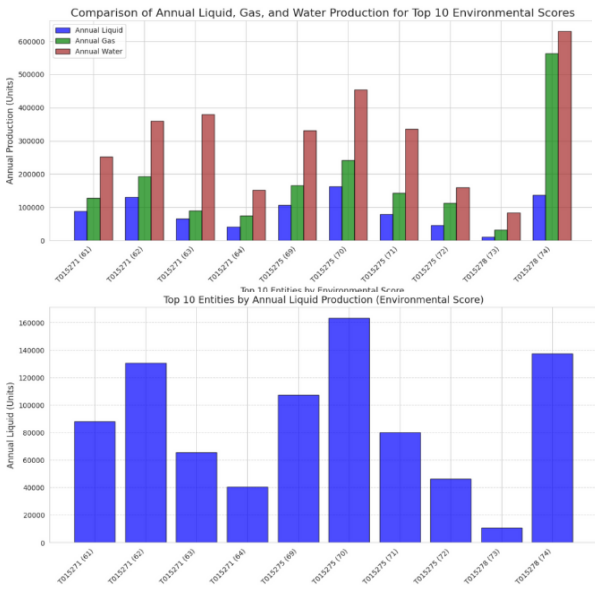
Histograms were used to illustrate the distribution of production scores, environmental scores, and combined scores, revealing underlying trends in the data.

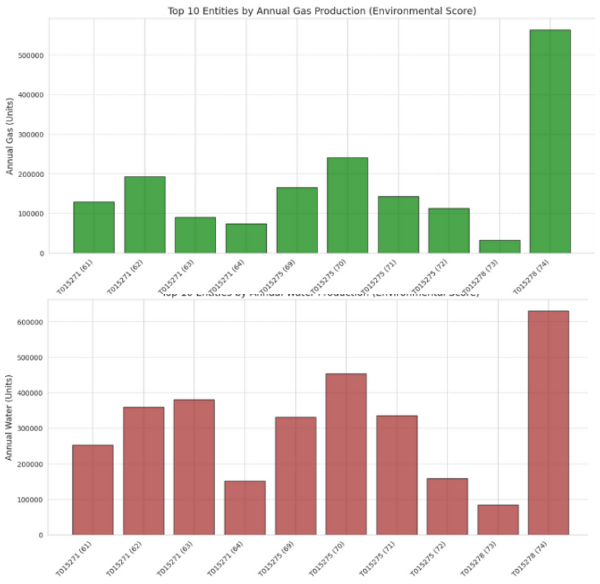
LSTM

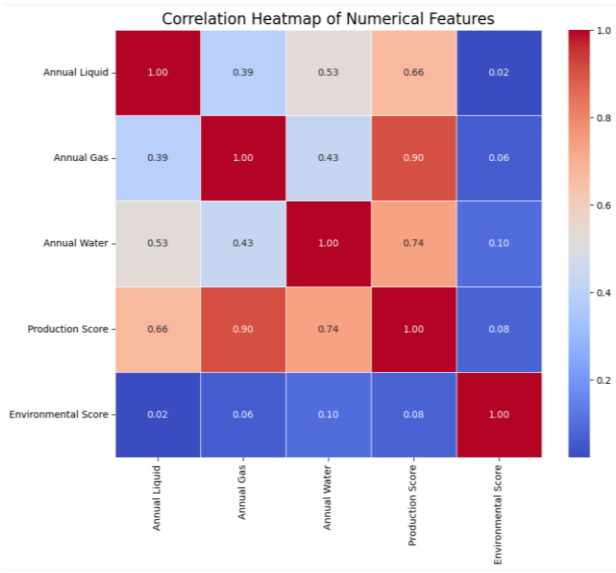




RNN

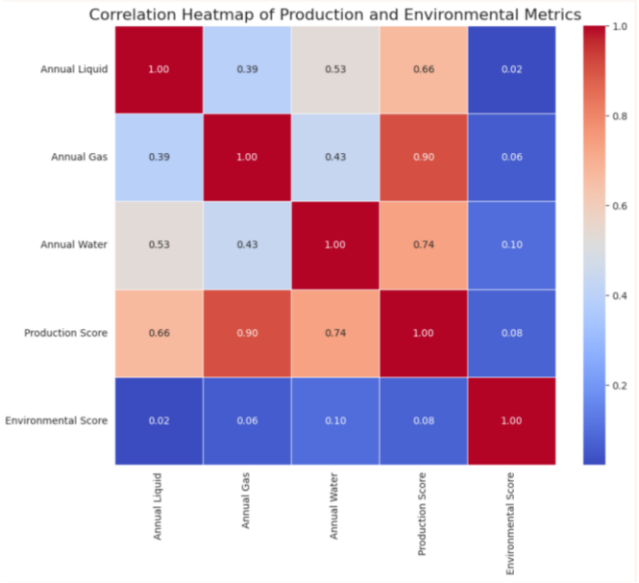




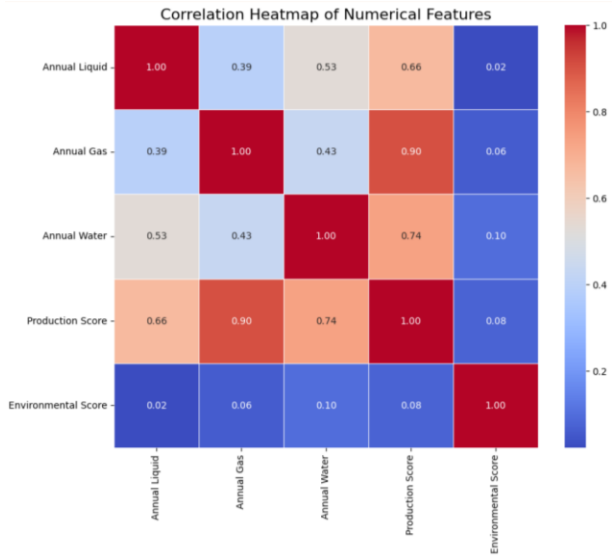


Furthermore, heatmaps were generated to analyze the correlation between features, offering valuable insights into relationships such as the inverse link between seismic magnitude and environmental scores.

LSTM

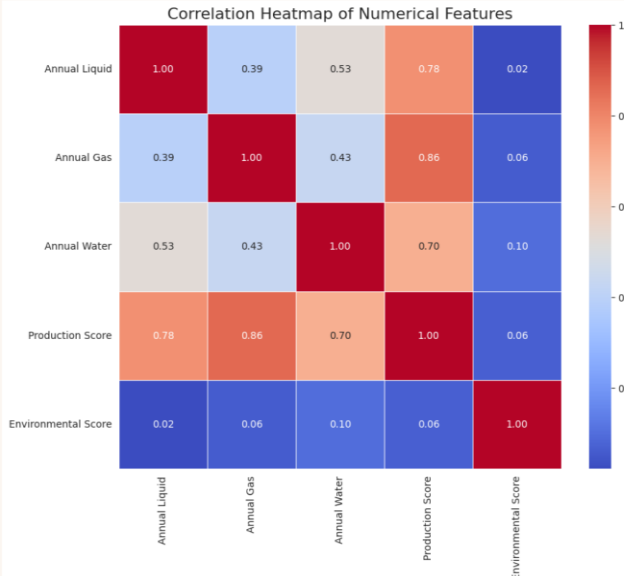


SVM



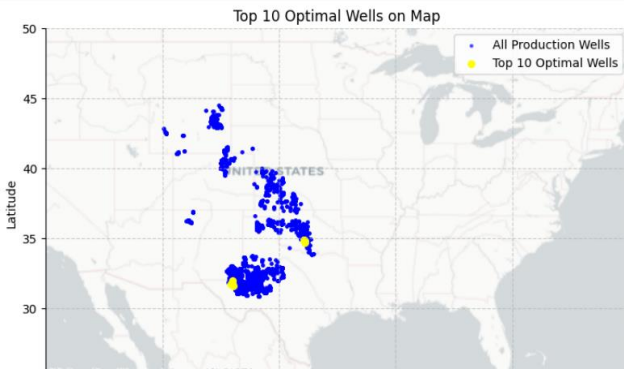
Random Forest

RNN



The top 10 optimal wells were identified by RNN, showing high normalized production scores and low seismic risks.

Entity	Latitude	Longitude
23999	31.687732	-103.985304
22402	34.695031	-97.740893
22381	34.842076	-97.832113
16971	34.782294	-97.772416
18914	34.769894	-97.751868
18932	31.717306	-104.152967
22391	34.841883	-97.840063
2280	31.687715	-103.988632
2281	31.594981	-103.885770
22403	31.969978	-103.998075



Results and Discussion

Algorithm	Normalized MSE	Strengths	Weakness
SVM	1.06799	Simple, interpretable	Poor scalability, hyperparameter sensitivity
Random Forest	0.01693	Handles High-Dimensional Data	Overfitting on Noisy Data
LSTM	0.004993	Handles temporal relationships well	Computationally intensive
RNN	0.00382	Sequential data handling	Vanishing exploding gradient

The project demonstrated that RNN outperformed SVM, Random Forest, and LSTM in predictive accuracy.

MSE SCORE FOR RNN

Normalized MSE	RNN
Production Score	0.0001
Environmental Score	0.3593
Combined Score	0.0038

This success is attributed to RNN's ability to effectively capture sequential relationships and temporal dependencies, making it

particularly well-suited for datasets involving production metrics and seismic activity. Additionally, RNN's simpler architecture compared to LSTM allowed it to achieve high performance with reduced computational overhead, resulting in faster training and inference times. However, like other deep learning models, RNN relies on efficient preprocessing pipelines to ensure data quality and optimize performance.

While the scoring system offered actionable insights into well placement, future enhancements could include integrating additional environmental factors, such as groundwater usage and greenhouse gas emissions. This would further align well placement strategies with broader sustainability and environmental stewardship goals.

Conclusion

This project highlighted the potential of machine learning in optimizing fracking operations. By integrating production and seismic data, the developed scoring system identified well locations that balanced productivity and environmental safety. RNN emerged as the most effective algorithm, providing the highest predictive accuracy. Future iterations will focus on expanding environmental considerations and improving model scalability to drive sustainable resource extraction practices.

Limitations

Limited Environmental Variables:
The environmental score was exclusively based on seismic activity, which, while significant, excludes other critical environmental factors. For instance, risks related to groundwater contamination, greenhouse gas emissions, and soil degradation were not considered. Incorporating these variables could provide a more holistic evaluation of environmental risks.

Data Coverage:
The dataset used in this project was geographically constrained, focusing on specific regions. This limitation may impact the generalizability of the model to other areas with different geological or environmental characteristics. A more diverse dataset could improve the robustness of the scoring system and its applicability across various locations.

Model Complexity and Computational Demands:
While Long Short-Term Memory (LSTM) models effectively capture temporal dependencies, their high computational demands can limit scalability, especially for larger datasets. Training LSTM models requires significant computational resources and careful hyperparameter tuning to avoid overfitting. Recurrent Neural Networks (RNN) mitigated some of these issues but still require well-structured preprocessing and optimization.

Reliance on Preprocessed Data:
All models relied heavily on preprocessed data, including imputed missing values and standardized ranges. This preprocessing introduces potential biases and assumptions, which may affect model accuracy if the raw data quality varies significantly in real-world applications.

Interpretability Challenges:
Deep learning models like RNN and LSTM, while powerful, lack interpretability compared to simpler models like Random Forest. This limitation can make it harder to explain the model's decisions to stakeholders, which is crucial in high-stakes industries like oil and gas.

Future Work

Expand the environmental score to include metrics like air quality, water usage, and greenhouse gas emissions.
Develop ensemble models combining RNN with other algorithms to enhance predictive accuracy.
Scale the model to accommodate larger datasets, ensuring its applicability across diverse geological conditions.

References

U.S. Geological Survey. (2023). Earthquake Hazards Program. Retrieved from USGS Website
IHS Markit. (2023). Energy and Natural Resources Data. Retrieved from IHS Markit Website