PROCEEDINGS, 49th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 12-14, 2024
SGP-TR-227

# Data Driven Web Based Simulation for Geothermal Heat Pump Drilling Around Saxony, Germany

Berat Tuğberk Günel

ITU Ayazağa Maslak, Sarıyer/İstanbul, Türkiye

gunel18@itu.edu.tr

**Keywords:** geothermal drilling, data-driven model, machine learning, mechanical-specific energy, drillingautomation.streamlit.app

**ABSTRACT**

This research describes an innovative data-driven approach for estimating mechanical-specific energy (MSE) in the context of geothermal heat pump well drilling in Saxony, Germany. Due to their remarkable hardness, the geological formations in this region present particular challenges. A web-based automated simulation was constructed using datasets generously donated by Technische Universität Freiberg and real-time measurements obtained from the dedicated Wissenschaftliches Bohren in Sachsen (WiBOS) drilling rig. To improve MSE predictions, the simulation uses machine learning techniques such as uncertainty analysis. To accomplish this, this web-based software is divided into four independent phases: data pre-processing, interval automation, visualization, and energy prediction. Notably, the Data Pre-Processing phase detects and manages outliers using specialized algorithms and statistical models, whilst the Intervals Automation phase provides mechanical-specific energy values for depth intervals. Furthermore, this work investigates the complexities of energy prediction by taking geological formations, drilling depths, and MSE values into account. The technology enables drilling specialists to find correlations, interpret data, and make informed judgments in this difficult geological setting. This research contributes to more efficient and cost-effective drilling operations in Saxony by providing valuable insights for the geothermal heat pump well drilling business in this region by delivering a data-driven approach to MSE prediction and energy management.

## 1. INTRODUCTION

Drilling a well is the first fundamental step to initialize the fluid production from the subsurface. Such wells can be encircled around conventional hydrocarbon production purposes, or geothermal. In either way, drilling consumes tremendous amount of time, and money compared to other operations. That's why, several optimization methods have been developed by scientist, and still continue to be developing.

In this study, it was tried to construct a data processing pipeline following a data-driven model to predict mechanical-specific energy in the unit of bars for possible prospective geothermal wells. To achieve this, a web-based automated simulation that is empowered by statistical models, and machine learning methods was constructed. Such a web-based simulation can be coded by using various software languages. However, in this study, a Python library called Streamlit was utilized (Streamlit, 2024). This web-based software consists of four distinct pages as follows: Data Pre-Processing, Intervals Automation, Visualization, and Energy Prediction was built.

Moreover, a detailed study of various datasets relevant to geothermal heat pump wells was undertaken throughout the development and subsequent algorithmic selection procedure. The Institut für Bohrtechnik und Fluidbergbau of Technische Universität Freiberg generously contributed these datasets. The datasets included a wide range of real-time measures, including depth, temporal data, rate of penetration (ROP), torque, axial tension, air pressure, air flow rate, and among others. Furthermore, the use of the Wissenschaftliches Bohren in Sachsen (WiBOS) drilling rig, a specialized scientific research equipment owned by the university, aided in the collecting of these real-time measurements.

## 2. METHODOLOGY

In this section, the previously mentioned web pages will be explained in detail. However, this section is designed to be a guide for the usage of the web-based simulation. Nonetheless, a technical background concerning each implementation will be pointed out.

### 2.1 Data Pre-Processing

Data Pre-Processing enables users to upload comma-separated file (.csv) or Microsoft Excel file one at a time. When a proper dataset file is successfully uploaded, it consequently displays the name of the well, the summary statistics table, and the number of missing values if there are any. If any missing value is found within the dataset, it displays a table which demonstrates the number of missing (empty) values for each corresponding feature. Along with this table, two options will be appearing when any missing value exists: Dropping missing values, and imputing missing values. As the name implies, when the 'Drop NaN' option is selected, it drops all of the missing values from the dataset; whereas 'Impute NaN' selection tries to impute missing values by leveraging each feature's statistical behaviour. However, the 'Impute NaN' option is still under development.

**2.2 Intervals Automation**

When the first step, in this case the Data Pre-Processing, is prosperously completed, a user can proceed into this page. Then, it will be asked to fill the input properties as follows: Drill Pipe Length, Error Rate, Threshold, Drill Pipe Weight, Hammer Weight, Bit Information, Formation Information, Formation Water Depth. The first three input properties are mandatory, where as other ones are optional. However, the software requires all of the properties except the formation water depth to calculate the mechanical specific energy values for each interval. Thus, the formation information is particularly curial for the behalf of machine learning model as these pieces of information are used by models to correlate each formation along with its depth to estimate mechanical specific energy values. Immediately upon prosperous property entries, the software tries to detects outliers within the dataset. To do that, the software uses three different approaches: A Custom Model, Statistical Model, and Machine Learning Model.

2.2.1 Custom Model

To be able to construct a custom algorithm to detect outliers within the uploaded dataset, the field observations were used for behalf of this algorithm. The measured features that are found to be beneficial for the behalf of the custom model are depth, and air pressure. To expand this, it was found that if depth measurements are less than zero or less than 0.001 meter, it indicates non-drilling operation. Following this phenomenon, it was observed that if air pressure is less than 2 bars, again, it is an indication of non-drilling operation. At the end, the algorithm locates the corresponding indexes of outliers, counts the number of the outliers, and then drops them from the dataset.

2.2.2 Statistical Model

In order to detect outliers by leveraging the statistical methods, appropriate features should be identified as the outlier investigation will be performed from the distribution of these features. According to this, air pressure, air flow rate, and revolution per minute (RPM) measurements were chosen to be investigated. To investigate the possible outliers, the model uses three sigma rule Equation (1) to capture 99.73% of the measurements from RPM feature. In Equation (1), $\mu$ and $\sigma$ symbolize the mean and standard deviation of a feature respectively.

$$Lower\ Bound = \mu - 3\sigma$$
$$Upper\ Bound = \mu + 3\sigma \quad (1)$$

On the other hand, z-score test is used on the features air pressure, and air flow rate. A Z-score quantifies how many standard deviations a value deviates from the mean of a dataset assumed to follow a normal distribution. The use of Z-scores enable the conversion of a value's original distance from the mean into standard deviation units (Khare, Khare, Nema, & Baredar, 2023). That's why, the model leverages the benefit of Z-score values from Equation (2) to establish a comparison between resulted Z-score values and specified threshold value. $x$ term in Equation (2) stands for the value of a feature. Lastly, it was found to be important to mention that the best threshold value is 3 as it obeys the three-sigma rule.

$$Z = \frac{x - \mu}{\sigma} \quad (2)$$

2.2.3 Machine Learning Model

To benefit from a machine learning model to detect outliers from the given observations, the rate of penetration, and delta time measurements were chosen for investigation as it is quite straightforward to locate outliers from these observations. On the other hand, the isolation forest algorithm was utilized to perform the outlier's detection operation from these measurements. Utilizing the Isolation Forest algorithm for outlier detection in drilling measurements, such as ROP and delta time, offers several key advantages. It efficiently handles high-dimensional data, making it suitable for the substantial volume of drilling data. Isolation Forest is scalable and adaptable to real-time streaming data, ensuring rapid outlier identification for improved operational efficiency and safety. Its robustness to multimodal data and independence from distribution assumptions make it a versatile choice in the complex drilling environment. Additionally, its interpretable anomaly scores aid in prioritizing and investigating outliers, facilitating informed decision-making for enhanced drilling performance.

Following the outlier detection and removal, the software deletes the outliers and their corresponding rows from the dataset; then, it tries to specify the drilling intervals. Fundamentally, the intervals indicate the number of drill pipes that are being used during a drilling operation. Then the algorithm performs data processing and statistical analysis on drilling measurements. The function begins by dynamically generating column names for statistical results based on the specified columns to be processed. It also identifies and includes additional columns based on specific patterns, enhancing the flexibility of data analysis. Next, it extracts relevant data, including depth and delta time, from the input DataFrame and calculates the number of intervals based on depth data, interval size, and an error percentage. Additionally, the code evaluates the derivatives of depth measurements concerning time, ensuring continuous changes over time are considered in the analysis. The error rate, as specified in the code, plays a crucial role in constructing confidence intervals. This approach helps in managing and handling outliers effectively while generating meaningful statistical results within the defined confidence bounds.

When the intervals are successfully calculated, the algorithm starts to calculate the pseudodrillibility index from Equation (3) where ROP is the penetration rate (m/h), N the rotational speed (RPM), W the thrust (kN), α the pseudodrillibility index (kN/mm), and D the bit diameter (cm) (Khare, Khare, Nema, & Baredar, 2023). It was called as pseudodrillibility index as these measurements were not downhole measurements, rather they are surface measurements. Moreover, the pseudodrillibility index is calculated to correlate the surface measurement with downhole measurements if there is any substantial relation exists. The demonstration of possible correlations can be examined from Visualization page of the web-based simulation software.

$$ROP = 3.35\frac{NW}{\alpha D} \tag{3}$$

Furthermore, if the given dataset contains pieces of information concerning axial tension, and torque; and the input parameters such as drill bit information, drill pipe information, and hammer information were given, the algorithm tries to calculate mechanical specific energy (MSE) in psi from Equation (4). Thus, Equation (4) benefits from drilling units. From an operational perspective, this is valuable as it serves as a benchmark for assessing efficiency. When the observed MSE closely aligns with the recognized strength of confined rock, it indicates that the drilling bit is operating efficiently. However, if there is a significant deviation, it implies that energy is being wasted (Okuchaba, 2008). Lastly, WOB stands for weight on bit applied on the drilling bit in lbm, A is the surface area of the bit in $ft^2$, T is the torque in ft-lbf, ROP is the rate of penetration in ft/h.

$$MSE = \frac{WOB}{A} + \frac{120\pi NT}{AxROP} \tag{4}$$

On the other hand, in order to account for the energy usage of the drilling rig, so called WiBOS, Equation 5 is utilized, and adapted for the hydraulic power calculation to compensate energy consumption (Bourgoyne, Millheim, Chenevert, & Young, 1986). In Equation (5), $P_H$ stands for hydraulic horse power in hp, $p$ is the pressure in psi, and $q$ is the flow rate in gal/min.

$$P_H = \frac{\Delta pq}{1714} \tag{5}$$

In conclusion, the software uses a Custom Model, Statistical Model, and Machine Learning Model to discover and manage outliers, hence they improve the data quality. The algorithm computes the pseudodrillibility index and mechanical specific energy to estimate the drilling efficiency. It also accounts energy consumption via Equation (5).

### 2.3 Visualization

Data visualization is critical in today's data-driven businesses, such as drilling operations in the subject of drilling engineering. Streamlit, a Python package for constructing interactive web applications, is used in this application to create a user-friendly platform for visualizing drilling data. This section highlights the application's key components and features and as well as its prospective impact on the drilling sector.

#### 2.3.1 Correlation Matrix

Creating and displaying of correlation matrix visualizations is one of the application's key functions. Users can choose which variables to include and whether to include formation information. The application computes and shows correlation coefficients, revealing correlations between drilling parameters. This function assists engineers and analysts in discovering important correlations that may have an impact on drilling efficiency. The math behind this correlation matrix is the determining simple linear relationship coefficients between each feature to display their relation. Minus one dictates a perfect negative correlation whereas positive one indicates perfect positive correlation.

#### 2.3.2 Feature Investigation

The "Feature Investigation" section allows users to utilize scatter plots to investigate correlations between two variables. Users can customize the x- and y-axes variables, include formation information, and even add linear regression curves for further research. Additionally, when applicable, users can visualize drilling depths and formation water depths, boosting their understanding of drilling profiles.

#### 2.3.3 Distribution Analysis

The software has the capability for investigating the distribution of drilling parameters. Users can select a feature of interest and choose whether or not to include formation information. The tool generates histograms and violin plots, which provide information on the statistical distribution and variability of drilling data.

### 2.4 Energy Prediction

The approach and technical specifics of the program created for energy prediction in drilling operations are presented in this section. The software is designed to anticipate mechanical specific energy (MSE) at various depth intervals during drilling. It includes data pre-processing, the building of machine learning models, and the study of uncertainty.

Mechanical-specific energy plays pivotal role in the optimization of the drilling process. Because, it indicates how efficient the drilling operation is. If the consumed mechanical-specific energy is close to the confined stress of the formation, the drilling process is said to be efficient. So, during the drilling operations, engineer should monitor difference between predicted mechanical-specific energy, and consumed mechanical-specific energy. Moreover, the ultimate aim should be keeping the difference as small as possible to improve the drilling efficiency. The consumed mechanical-specific energy can be altered by weight on bit, surface area of the bit, revolution per minute, or torque.

2.4.1 Data Pre-Processing

Data preparation is the first critical stage in the energy prediction process. The software accepts datasets comprising depth interval information, potential formations, and MSE values. The following are important phases in data pre-processing:

**Data Validation:** The application checks provided datasets to ensure they fulfil specified format standards. It looks for crucial columns. Users are encouraged to provide appropriately formatted data if incorrectly structured datasets are found.

**Concatenation:** When different datasets are uploaded, the software merges them to form a single comprehensive dataset. This stage ensures that information from multiple sources is pooled for the analysis.

**User Inputs:** Users are requested to enter crucial parameters such as "Target Depth" and "Drill Pipe Length." The depth intervals for MSE prediction are determined by these parameters.

**Formation Intervals:** Users can determine likely formation intervals by providing depth ranges and assigning formations. This data is essential for correlating depth intervals to plausible geological formations.

2.4.2 Machine Learning Model Development

The development of machine learning models to forecast MSE values is at the heart of the energy prediction software. The following are the steps in model development:

**Feature Engineering:** The program employs dataset features such as "Teufe [m] Mean" and "Formation." The "Formation" feature is one-hot encoded to transform category data into numerical form. That means, it assigns unique numbers for each formation type.

**Data Scaling:** The "Teufe [m] Mean" feature is standardized using StandardScaler, which ensures that it has a mean of 0 and a standard deviation of 1. In Equation 2, x represents original value of a feature, μ compensates mean of the training samples, and σ accounts for the standard deviation of the training samples (sklearn.preprocessing.StandardScaler, n.d.). In machine learning, standardization is critical to bringing features to a similar scale and preventing one feature from dominating the others during model training. This increases the convergence of optimization techniques as well as the overall performance of models that use distance-based calculations, such as Support Vector Machines and k-means clustering.

**Model Selection:** For MSE prediction, a variety of regression models are examined, including Linear Regression, Random Forest Regressor, Support Vector Regressor (SVR), Gradient Boosting Regressor, and Elastic Net. Using cross-validation, a grid search technique is used to optimize hyperparameters for each model.

**Model Training:** The best-performing model is determined using cross-validation performance. The hyperparameters are tweaked, and the chosen model is trained on the training data.

**Model Evaluation:** The trained model is tested on a hold-out test dataset using evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. These metrics give information on the projected accuracy of the model.

2.4.3 Uncertainty Analysis

The software does uncertainty analysis using bootstrapping to account for uncertainty in MSE predictions:

**Input Data Transformation:** The input data is adapted to meet the model training format, including depth intervals and projected shapes. In a single pass, the input data is encoded and scaled.

**Bootstrap Resampling:** The input data is resampled using replacement to achieve bootstrapping. For each resampled dataset, the trained model predicts MSE values.

**Uncertainty Bounds:** Based on the bootstrapped MSE predictions, uncertainty bounds, including lower and upper bounds by referring Equation (1), are determined. These bounds show the range of MSE values that can be obtained for each depth interval.

**2.4.4 RESULTS AND DISCUSSION**

The software presents the results in a clear and informative manner:

**Machine Learning Model Results:** To evaluate the prediction ability of the best-performing model, measures such as mean squared error (MSE), root mean squared error (RMSE), and $R^2$ are provided. Hyperparameters utilized for the specified model are also displayed if appropriate.

**Model Prediction:** There are predicted MSE values for each depth interval, as well as corresponding uncertainty ranges. This data helps drilling engineers make educated judgments about energy management during drilling operations.

**Uncertainty Analysis Plot:** The anticipated MSE with uncertainty intervals is visualized in a graphic. It assists users in comprehending the diversity in MSE predictions at various depths and forms.

Furthermore, the energy prediction tool combines data pre-processing, machine learning model development, and uncertainty analysis to produce accurate and actionable drilling projections. Drilling engineers can utilize machine learning approaches to enhance energy management, possibly cutting operational costs and minimizing environmental impact.

**REFERENCES**

Bourgoyne, A. T., Millheim, K. K., Chenevert, M. E., & Young, F. S. (1986, January). *Applied Drilling Engineering.* doi:10.2118/9781555630010

Khare, V., Khare, C., Nema, S., & Baredar, P. (2023). Chapter 2 - Data visualization and descriptive statistics of solar energy system. In V. Khare, C. Khare, S. Nema, & P. Baredar (Eds.), *Decision Science and Operations Management of Solar Energy Systems* (pp. 33-75). Academic Press. doi:https://doi.org/10.1016/B978-0-323-85761-1.00002-0

Okuchaba, B. J. (2008). Development of a model to calculate mechanical specific energy for air hammer drilling systems. *Master's Thesis*. Texas A&M University. Retrieved from https://hdl.handle.net/1969.1/ETD-TAMU-2804

*sklearn.preprocessing.StandardScaler*. (n.d.). Retrieved from Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

*Streamlit*. (2024). Retrieved from Streamlit: https://streamlit.io