PROCEEDINGS, 47th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 7-9, 2022
SGP-TR-223

A Probabilistic Approach to Model and Optimize Geothermal Drilling

Robert Rallo[1], Rolando Carbonari[2], Dang Ton[3], Rahmat Ashari[4], Pradeepkumar Ashok[4], Alain Bonneville[5], Daniel Bour[6], Trenton Cladouhos[7], Geoffrey Garrison[8], Roland Horne[3], Eric van Oort[4], Susan Petty[7], Adam Schultz[9], Carsten F Sørlie[10], Ingolfur Orn Thorbjornsson[11], Matt Uddenberg[12], Leandra Weydt[13]

[1]Pacific Northwest National Laboratory; [2]Hebrew University of Jerusalem; [3]Stanford University; [4]University of Texas, Austin; [5]Geosciences Consulting LLC; [6]Bour Consulting; [7]Cyrq energy; [8]AltaRock Energy; [9]Oregon State University; [10]Equinor; [11]Iceland GeoSurvey (ISOR); [12]Stravan Consulting; [13]Technical University of Darmstadt

E-mail: robert.rallo@pnnl.gov, rolando.carbonari@mail.huji.ac.il, dangton@stanford.edu, rahx@utexas.edu, pradeepkumar@austin.texas.edu, alain.bonneville@geosciences-consulting.com, Daniel@bourconsult.com, trenton.Cladouhos@cyrqenergy.com, ggarrison@altarockenergy.com, horne@stanford.edu, vanoort@austin.texas.edu, susan.petty@cyrqenergy.com, Adam.Schultz@oregonstate.edu, cso@equinor.com, ingolfur.thorbjornsson@isor.is, muddenberg@stravan.co, weydt@geo.tu-darmstadt.de

Keywords: probabilistic models, optimization, process mining

ABSTRACT

Data-driven approaches are key for modeling and optimizing the operation of geothermal wells. However, data collected at different drilling sites are usually not standardized and contain missing or erroneous values which hinder the development of reliable models that can be used across a broad range of conditions and geological contexts. In this work, we present an end-to-end workflow for analyzing, modeling, and optimizing geothermal drilling based on probabilistic methods to account for data uncertainty and heterogeneity. The components of the workflow include: 1) self-organizing maps to visualize well operation data and process trajectories, 2) process mining to analyze drilling event logs and discover process models, 3) Bayesian models to predict relevant operation and performance metrics from incomplete and uncertain drilling data, and 4) algorithms to optimize well drilling under specific operational constraints. The workflow has been implemented as a web-based tool that facilitates geothermal drilling planning and operation tasks. Models used in the workflow have been trained and tested on the database of 113 geothermal wells representing various geological settings which was built in the framework of the EDGE project.

1. INTRODUCTION

The cost per foot in geothermal drilling is more expensive than the cost of onshore oil and gas drilling. Relative to oil and gas drilling, the main factors driving geothermal cost include harsh downhole conditions that shorten equipment's life, the use of larger hole diameters that require more expensive casing, and indirect costs due to fluid re-injection. Accordingly, reducing drilling cost and the risk of well failure are key for cost reduction in geothermal energy, particularly for Enhanced Geothermal Systems (EGS), where 60-80% of the total cost is in the wellfield.

The EDGE project (Carbonari et al., 2021), funded by the U.S. Department of Energy (DOE) Geothermal Technologies Office, is an international research coordination network aimed at developing machine learning strategies to improve geothermal drilling efficiency, including cost reduction and early well failure identification. The EDGE research team includes three U.S. Universities, a DOE National Laboratory and four Geothermal and Oil and Gas companies from several countries (Iceland, Norway, USA). The aim of the project is to use data from different geothermal fields to develop a continuous optimization framework for geothermal drilling. The resulting optimization scheme will be able to continuously improve as new data are ingested into the database. The project is structured in two yearlong phases with specific tasks focusing on data collection and management, model development and drilling optimization, and identification and mitigation of well failures.

During the first year, the project team focused on collecting data from more than 100 wells from different companies and geothermal fields and developing a curated data repository including data, computational codes, analysis workflows and models. Exploratory data analysis (EDA) was used to assess both the quality and the structure of the data (missing data, outliers, errors, correlations). After preprocessing, a subset of data was used to identify suitable machine learning approaches for developing predictive models.

The second year of the project focuses on developing and validating models to predict well operation parameters, non-productive time, and drilling cost. Models are integrated into a multi-objective optimization framework that enables the identification of optimal operation parameters to minimize cost and non-productive time. In addition, drilling data and machine learning techniques are being used to identify main causes of well failure and to develop mitigation strategies aimed at reducing the impact of such failures.

At the end of the project. the main deliverables will include the EDGE Data Repository that will provide access to all public data products developed during the project, and the EDGE Dashboard that will provide a common interface to access the models and the multi-objective optimization framework. This paper provides an overview of diverse elements of the geothermal well optimization framework, including exploratory data analysis and visualization with self-organizing maps, process model discovery from drilling logs, probabilistic Bayesian models to estimate key features of the drilling process, and drilling process optimization. The workflow has been implemented as a web-based tool that facilitates planning and operation tasks. Models included in the dashboard have been trained and tested on a database of 113 geothermal wells representing various geological settings.

2. DATA AND METHODS

2.1 Drilling Data

Data gathered in the EDGE project includes both proprietary and public well data. Proprietary information on well operation has been provided by EDGE project's industrial partners - AltaRock (USA), Cyrq Energy (USA), Equinor (Norway), Iceland GeoSurvey-ISOR (Iceland)-, while the public data come from the Utah FORGE and Fallon project. Raw data and curated datasets, containing information related to drilling operations at different temporal scales, were annotated with geological information to facilitate data analytics and machine learning tasks.

2.2 Self-Organizing Map Analysis

The Self-Organizing Map (SOM) algorithm performs a topology preserving mapping from a high-dimensional input space onto a low dimensional output space formed by a regular grid of map units (Kohonen, 1990). From a functional point-of-view, SOM resembles Vector Quantization which approximates, in an unsupervised way, the probability density functions of high-dimensional input data with a finite set of reference vectors that describe class borders with a nearest-neighbor rule. In contrast, SOM units are organized over the space spanned by a regular grid where the adaptation process affects a predefined topological neighborhood producing both, a vector quantization and a low-dimensional ordered representation of the original high-dimensional input data. In addition, since each SOM unit has well-defined low-dimensional coordinates over the map grid, the SOM can also be used as a dimensional reduction algorithm.

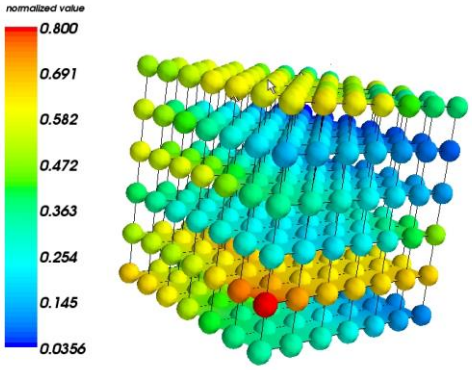


Figure 1: SOM structure. Horizontal layers correspond to projections of a single variable whereas vertical rows represent individual components of the codebook vector for each SOM unit.

The SOM algorithm is based on unsupervised competitive learning. The training process is entirely data-driven and SOM units compete to become specific detectors of certain features in data. Map units are represented by a *n*-dimensional weight vector, with *n* equal to the input data dimension. As in vector quantization, every reference vector describing a class is named a codebook. Each SOM unit has a topological neighborhood determined by the shape and structure of the SOM lattice which can be either rectangular or hexagonal. The number of SOM units and their topological relations are defined during algorithm initialization. The size of the map will ultimately determine its accuracy and generalization capabilities.

SOMs are used to extract and visualize relationships among data in high dimensional spaces. There exist a variety of SOM-based visualization techniques that exploit its ability to project data in two dimensions. Figure 1 shows the layered organization of a SOM, where every unit provides a partial view of the whole data set for each variable at the corresponding horizontal layers. Codebook vectors can be visualized using the component planes (or c-planes), i.e., horizontal layers in Figure 1. In this representation, the SOM is considered as a set of stacked layers in which each component plane forms a horizontal layer in this structure while each codebook vector corresponds to column. Component planes are visualized by taking the values of the respective components from the codebook vectors and plotting them with a color code over the SOM grid. This representation provides valuable information on the distribution of each feature in the dataset. By visualizing several component planes simultaneously, it is possible to infer relationships between features.

Clusters are formed by groups of units with similar codebook vectors. The clustering structure of the input space can be visualized over the SOM grid by displaying the distances between neighboring units. The visualization of these clusters could be enhanced by labeling the map with auxiliary data. Since it is difficult to detect and quantify clusters just by visual inspection, SOM's reference vectors can, in turn, be clustered to detect coherent sets of units with similar structural characteristics by using simpler clustering algorithms such as K-means.

New data samples can be projected over the SOM grid by mapping to their corresponding unit. Dynamic processes (i.e., time-dependent) can also be analyzed and visualized by the SOM. The procedure to study the behavior of a transient process uses the locations of the units corresponding to a sequence of system states over the process dynamics. The evolution of a dynamical system can be then characterized as a trajectory corresponding to a sequence of system states displayed on the SOM. Labeling of the SOM permits the characterization of these trajectories and the identification of interesting regimes in the evolution of a dynamical system (process).

2.3 Analysis of Process Operation

Relevant information on the drilling process and geothermal well operation is collected by plant operators in the form of discrete event logs that include sequential information related to relevant events (Table 1) and their precise timestamps. The analysis of these events allows inferring process models that can be used to identify deviations with respect to normal operation parameters that can serve as early warning of potential faults.

Traditionally, event logs have been analyzed using probabilistic approaches such as Markov chain models. A Markov chain model is a transition probability matrix where each (*i,j*) element represents the probability that a drilling operation transitions to state *j* from state *i*. Using this formalism, the system is modeled as a sequence of states and, as time goes by, it moves in between states with a specific probability. The transitions between states are conditioned, or dependent, on the state you are in before the transition occurs.

More recently, process mining (Aalst et al., 2004; Aalst, 2016) has emerged as a promising approach for the analysis of operational processes from event logs. Process mining represents a collection of methods, algorithms, and tools to gain a better understanding of the execution of a process, by means of analyzing operational execution data (i.e., event logs). This approach can be used for process discovery, conformance checking, and performance analysis. Discovery is the process of automatically generating a model from event logs that can explain the logs themselves without any prior knowledge. There are several algorithms that can be used for this discovery process including alpha, heuristic and inductive miners. Conformance checking compares an event log with a given process model and identifies discrepancies (anomalies and faults) that can result in degraded performance. Finally, process models can be annotated with additional data (e.g., cost, time) to detect bottlenecks and inefficiencies.

Table 1: Operation and Operation Group Codes included in the EDGE Data Repository

Code	Description	OpsGroup	Code	Description	OpsGroup
CUTC	Cut and Pull Casing	ABAND	CASE	Running Casing	CASING
PLUG	Plugging Operations		DPIPE	Driving Pipe	
BOPND	BOP Nipple Down	BOPOPS	CMTD	Drilling Cement/Shoe	CEMENT
BOPNU	BOP Nipple Up		CMTP	Primary Cement Operations	
BOPO	Other BOP Operations		CMTPL	Cement Plug Operations	
COOL	Circulate and Cool Well	COMPLETE	CMTS	Secondary Cement Operations	
GRAV	Gravel Packing		FRPLUG	FRPLUG	
PERF	Perforating		WOC	Waiting On Cement	
STIM	Stimulation		CIRC	Circulate/Condition Mud	DRILL
TUBG	Running Production Tubing		CLCM	Clean out cement	
CORE	Coring	EVALUATE	CLEANO	Clean Out	
EVAL	Well Evaluation		DIR	Directional Work	
FIT	Leak Off Test		DRIL	Drilling Ahead w/ Connections	
LDTOOL	LDTOOL		DRILR	Drilling - Rotating	
LOG	Wireline Logging		DRILS	Drilling - Sliding	
PTEST	PTEST		FLOW	Flow Check	
SLK LN	Slick line operations		MAGFLX	Magna Flux Pipe	
SLWL	SLWL		MIX	Mud Mixing	
TEST	Testing Operations, DST		MIXMUD	Mix Mud	
INJ	Injection, Water	INJ	MUBHA	Make Up Bottom Hole Assembly	
COCMT	Clean out cement	LOST	OPEN	Opening Hole	
ANCH	Anchoring Operations	MOB	PUMP	Pumping Drilling Fluids	
JDN	Jacking Down Operations		RDFREQ	RDFREQ	
JUP	Jacking Up Operations		REAM	Reaming/Underreaming	
MOB	Mob/Demob		RUFREQ	RUFREQ	
MOVE	Move Rig		SURV	Running Survey Tools	
PRELD	Pre-loading Rig		WASH	Washing Down	
RIGD	Rigging Down		ACID	ACID	OTHER
RIGU	Rigging Up		ALIFT	ALIFT	
SKID	Skid Rig		BW	Blow Well	
FISH	Fishing Operations	PROBLM	CUT	Slip and cut drilling line	
KILL	Well Kill Operations		CUTDL	Cut and Slip Drill Line	
LOST	LosingCirc./Pumping LCM		DP	PU/LD Drill Pipe	
MILL	Milling		IDLE	Well idle	
POHREP	Trip Out for Repairs		Maintn	Non Down Time Repairs	
REGS	Regulatory Problems		OTHER	Other Activity	

REPR	Rig Repairs	
REPS	Service Company Repairs	
RIHREP	Trip In following Repairs	
STUCK	Stuck Pipe Operations	
STUK	Stuck Pipe Operations	
WOE	Waiting on Equipment	
WOO	Waiting on Orders	
WOW	Waiting on Weather	
SFTYMT	Safety meeting	SAFE
TEST1	TEST1	TEST
MAG	Magna Glow BHA	TOOL
RIGMOV	Rig Move	
SAFE	Safety Meeting	
SERV	Rig Service	
SSEA	Sub Sea Operations	
ST.BY	Stand by	
WELD	Welding Operations	
BHAOP	BHA Operations	TRIP
POH	Pull Out of Hole	
RIH	Run In Hole	
TOOLS	P/U--L/D TOOLS	
TRP	Trippingpipe	
TRPI	Trippingin	
TRPO	TrippingOut	
WIPE	Wiper Trip	

2.4 Bayesian Models and Optimization

A Bayesian network (BN) is a probabilistic graphical model for representing knowledge about an uncertain domain where each node corresponds to a random variable and each edge represents the conditional probability for the corresponding random variables (Pearl, 1988). Due to dependencies and conditional probabilities, a BN corresponds to a directed acyclic graph (DAG) without loops or self-connections. One of the key steps while developing a Bayesian network is using the proper representation of the causal structure of the modeled domain. The domain representation is embedded in the relationships between nodes in the model. There exist four main algorithmic approaches to learn the structure of a Bayesian network from data:

- Constraint-based algorithms, which use conditional independence tests to learn conditional independence constraints from data. The constraints in turn are used to learn the structure of the Bayesian network under the assumption that conditional independence implies graphical separation (so, two variables that are independent cannot be connected by an edge).

- Score-based algorithms, which are general-purpose optimization algorithms (e.g., hill-climbing, tabu search) that rank network structures with respect to a specific goodness-of-fit score.

- Hybrid algorithms that combine aspects of both constraint-based and score-based algorithms, as they use conditional independence tests (usually to reduce the search space) and network scores (to find the optimal network in the reduced space) at the same time.

- Pairwise Mutual Information algorithms, that learn approximate network structures using only pairwise mutual information.

Once trained, the Bayesian network can be used to estimate the posterior probability distribution of a set of query variables given an observed event represented by the values of a set of evidence variables. Given its computation complexity, exact inference becomes an intractable problem and is only feasible in small and simple networks. For large and more complex networks, approximate inference methods are commonly used to obtain reasonable estimates of the posterior probabilities. The accuracy of these stochastic approximation techniques largely depends on the number of samples generated. In this work we use *likelihood weighting* which is an approximate inference algorithm based on Monte Carlo sampling. Examples of queries that can be answered by a Bayesian network developed from geothermal drilling data include:

- *What is the most probable ROP given a WOB of 20,000 lbs. when drilling at a depth of 4000 ft in a geothermal well located in a metasedimentary reservoir?*

- *What is the probability of having a drilling cost in the range of $500/ft to $750/ft when drilling at an ROP of 15 ft/hr and with a mud flow average of 200 gal/min in an extended crust geological regime?*

Multiparameter and multi-objective optimization approaches combined with process models provide a framework to assess the effects of several variables on the minimization of a given cost function. In this work we have used the built-in R function *optim* coupled with various process models to optimize relevant geothermal drilling operation parameters such as rate of penetration (ROP). Some of the parameters that impact the ROP include bit properties (e.g., bit type, bit tooth wear, bit hydraulics), weight on bit (WOB), rotational speed (RPM), and the geological context. Multiparameter optimization can be applied for example to identify the optimum WOB and RPM that maximize ROP for a given geological context.

Complex cost functions, involving more than one optimization variable, require using multi-objective optimization techniques. Multi-objective optimization has been applied in fields where optimal decisions need to be taken in the presence of trade-offs between two or more conflicting objectives. Typically, outputs (or objectives) are conflicting and unique solutions where all objectives are minimized at once don’t exist. In this situation, the goal is to identify a set of optimal solutions know as Pareto set. Multi-objective optimization has been implemented using the *GPareto* package in R (Binois and Picheny, 2019) which provides Gaussian-Process based sequential strategies to solve this type of optimization problems.

3. RESULTS AND DISCUSSION

3.1 Exploratory Analysis of Drilling Data using Self-Organizing Maps (SOM)

We illustrate the use of the SOM methodology with data provided by ISOR corresponding to the operation of drilling wells located in different locations across Iceland. The data collected for these wells includes the drilling features listed in Figure 2. After a data preprocessing stage that included filtering erroneous and incomplete data followed by normalization and scaling, the set of 17 features were used to train a self-organizing map using a hexagonal lattice without periodic boundary conditions. The SOM analysis is based on a Matlab implementation (SOM Toolbox, http://www.cis.hut.fi/somtoolbox).

Feature	Units
Depth	m
ROP	m/h
WOB	ton
TDH	m
RPM	rpm
Torque	daNm
tot. rpm	rpm
SPP	bar
tot. pump	l/s
temp.down	°C
temp. ret.	°C
diff.temp	°C
Kill-line	bar
Drill_Time	hrs
Outer_Diam	inches
Inner_Diam	inches
Drill_bit_mm	mm

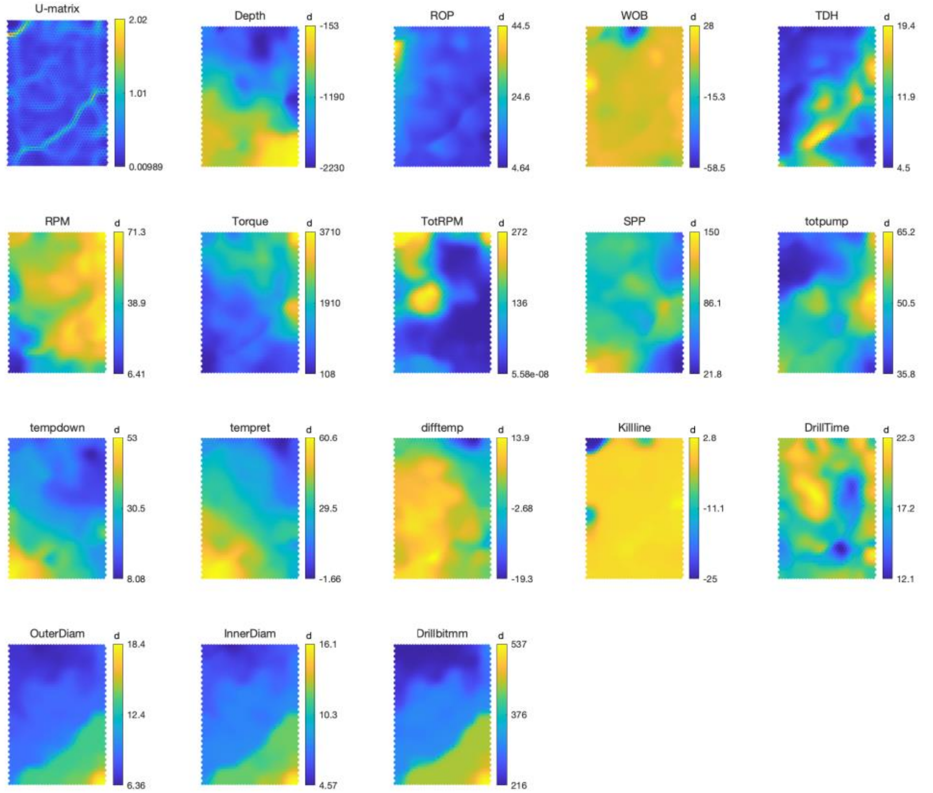


Figure 2: Self-Organizing Map Analysis of ISOR drilling data. (Left) Features included in the analysis. (Right) Visualization of SOM cluster structure (U-matrix) and component planes.

The analysis of the component planes (c-planes) facilitates the identification of relationships among the drilling process features. The c-planes in the last row of figure 2 that correspond to outer diameter, inner diameter and the size of the drill bit provide a simple example that illustrates how SOMs can be used as a visual analytics tool to discover relationships among features. The component plane corresponding to depth of the well highlights the presence of three main operation regimes (Figure 3). The inspection of the c-planes also highlights the dependency of the temperatures (*temp. down* and *temp. ret.*) with depth.

SOM projections were also used to analyze the operation trajectories of individual wells. The analysis reveals that wells operating in different geologic contexts have different operation patterns that produce diverse operation trajectories. The SOM, trained with data from all wells in the dataset, provides a common framework for the comparative analysis and visualization of individual operation. Figure 4 depicts the operation trajectories of a single well where each point in the trajectory corresponds to a given operation state and lines connect the sequence of individual states. Trajectories can be overlapped with relevant operation features to understand specific details of well operation. For instance, during the initial drilling stages of the well analyzed in Figure 4 (shallow deep) the well operates at low ROP. As depth increases, operation regions move for a short period of time to an area characterized by very high ROP (>40 m/h). As the drilling progresses and depth continues increasing the operation transitions to lower ROP ranges (~10 -15 m/h).

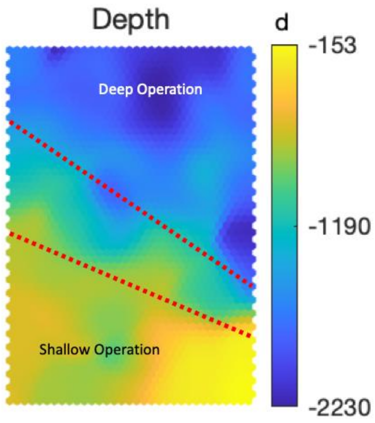


Figure 3: Drilling operation regimes based on depth.

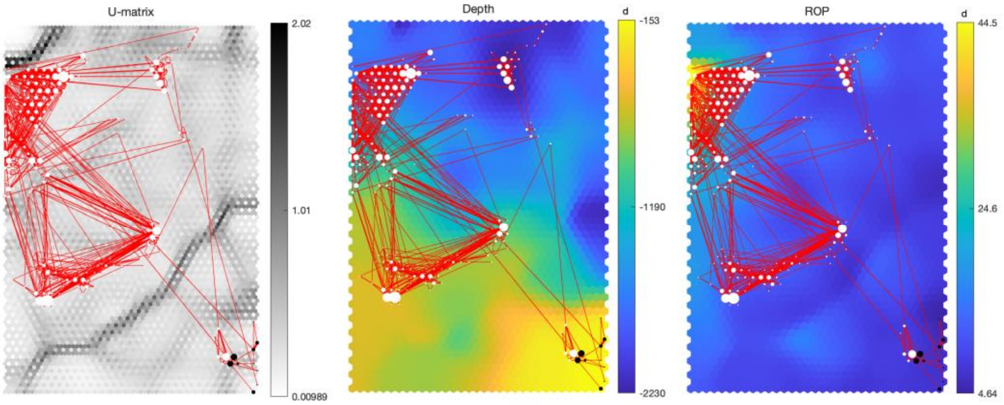


Figure 4: Projection of the operation trajectories over the SOM U-matrix and on top of the depth and ROP c-planes.

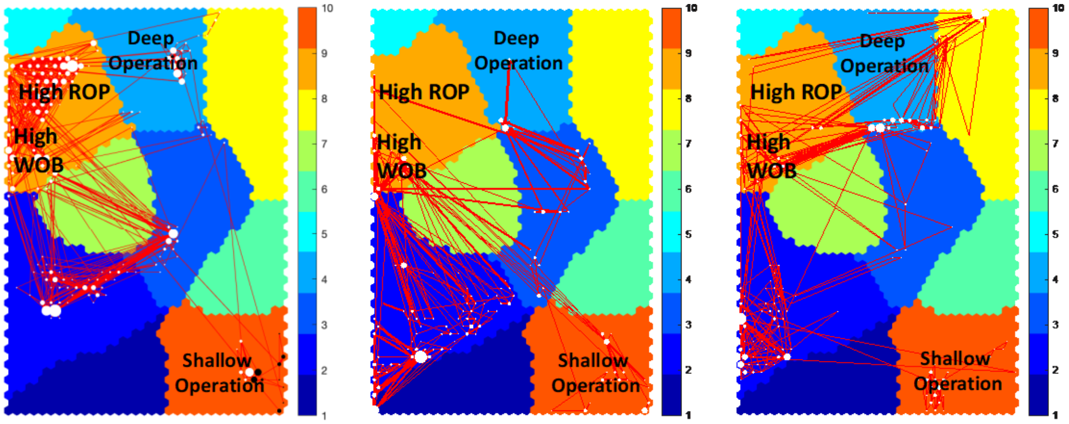


Figure 5: SOM clustering and projection of the operation trajectories corresponding to three wells located in different geological contexts. Each color patch corresponds to one of the ten operating regimes identified by the clustering process

Clustering and labeling the SOM improves the interpretability of the map and facilitates the identification of operation regimes. We used k-means clustering with k=10 to generate 10 partitions corresponding to distinct operation regimes. Figure 5 depicts the clustering results (where each colored region corresponds to a k-means cluster) and the operation trajectories of three wells located in different geothermal fields. The visual inspection of the trajectories highlights the different drilling patterns in each well. All trajectories originate in a shallow operation regime (i.e., beginning of the drilling process) but exhibit different dynamics as depth increases. Once clustered and labeled, the SOM can be used to identify abnormal operation regimes linked to well failures that may result in non-productive time or equipment damage. When used with real time data, the SOM can provide operators with an early detection mechanism of well failures allowing the implementation of mitigation actions.

3.2. Analysis of Process Operation

The analysis of sequences of operation data provides information on the probability of transitioning from a given drilling operation state to an operation leading to non-productive time. For this analysis we have used a dataset provided by AltaRock-Cyrq that includes a sequence of timestamped operation codes that describe the different states of the drilling process. The dataset also contains a flag that indicates whether or not the operation sequence has resulted in non-productive time (NPT). The operation codes are classified in 16 groups, including the 'PROBLM' category that corresponds to operation codes that resulted in problems during the drilling process (see Table 1). Figure 6 summarizes the number of instances of operations in the PROBLM category that are present in the dataset and that caused non-productive times. Equipment failures (i.e., rig repairs) and fishing operations resulting from the recovery of downhole equipment after a stuck pipe are the main causes of non-productive time. Recovering from these failure conditions has significant impact on the overall process cost. For instance, it is estimated that operation costs while fishing can increase up to 75%.

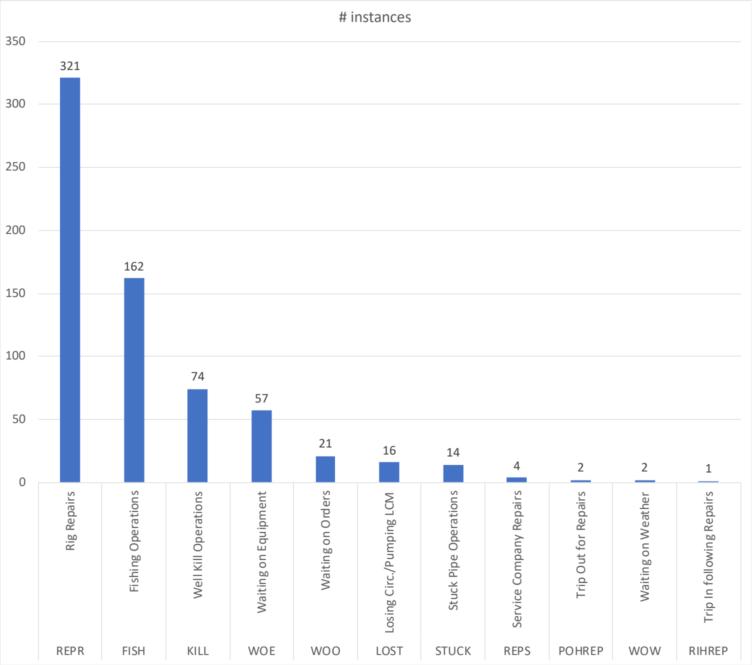


Figure 6: Distribution of the operation codes corresponding to the class 'PROBLM' (see Table 1) that resulted in non-productive times.

We have used the process mining techniques implemented in the PM4Py (https://pm4py.fit.fraunhofer.de/) python package to analyze the drilling logs and to generate process models that can help monitor and optimize well operation. Figure 7 depicts the process model obtained from the operation data of a single well exhibiting low incidence of non-productive time events. The process model provides a state transition diagram that relates all types of operation observed during the drilling process. In this specific case, the only non-productive time observed is due to waiting (WOO) between tripping in/out (TRPI/TRPO) operations. The sequence of drilling ahead (DRIL), rig service (SERV) and other generic activities (OTHER) are the most frequent observed operation states in this well and describe the expected normal operation. Process models provide an integrated view of the well operation and can be helpful to identify desired (i.e., cost-effective) and undesired (i.e., non-productive time) operation patterns that can provide relevant information to develop strategies to decrease drilling time (actual drilling and tripping) and reduce total costs.

The same approach can be used to integrate all the event log data in a geothermal field. Using combined data, the process model provides a holistic view of the operations across the whole field. Figure 8 presents the process model obtained after the integration of all well data. This model captures different well dynamics and therefore is more complex than the single well model discussed in figure 7. The most frequent categories of operations in this model include DRILLING, REPAIRS and TRIPPING.

Developing specific process models for individual wells facilitates the identification of common patterns of operation and allows the categorization of well operation regimes. The use of conformance checking algorithms allows the direct comparison of well operations and facilitates the identification of deviations with respect to expected/optimal behaviors.

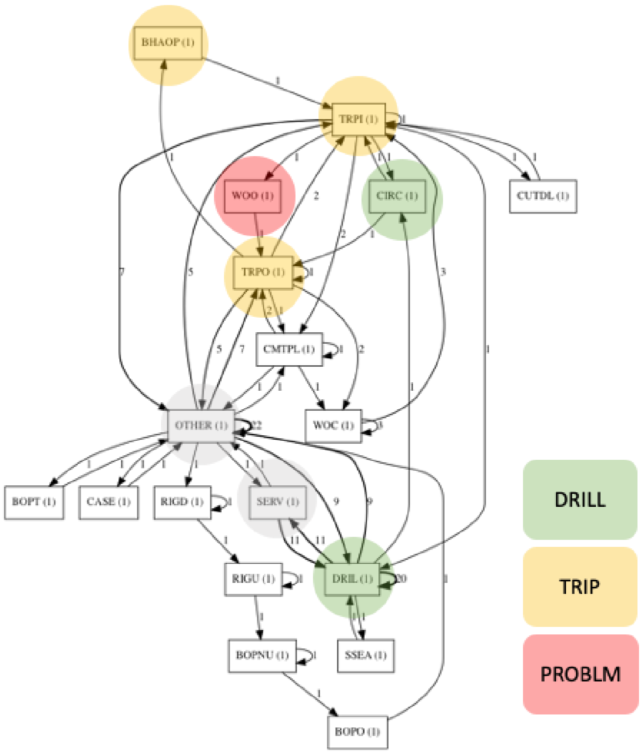


Figure 7: Process model corresponding to a geothermal well with very low incidence of non-productive time

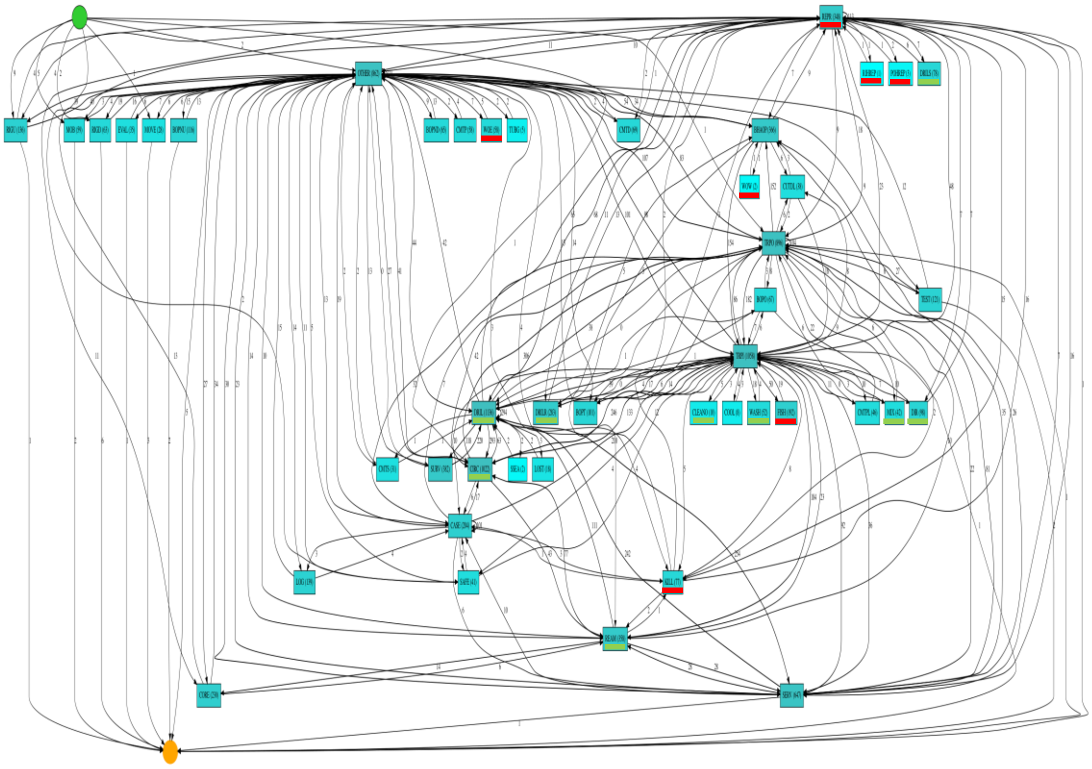


Figure 8: Model process for all the wells in a geothermal field. States marked in red correspond to the PROBLM category whereas the ones in green represent normal drilling operations (see Table 1).

3.3. Bayesian Models, Optimization, and the EDGE Dashboard

Probabilistic modeling using Bayesian networks allows capturing the details of the drilling operation while accounting for uncertainty and missing data. Since data corresponding to geothermal drilling includes a combination of discrete/categorical variables (e.g., geological context), and continuous features (e.g., operational parameters), Bayesian networks need to incorporate data heterogeneity. In this work we use a hybrid Bayesian network model that learns different types of conditional probability distributions for discrete and for continuous variables. The model used is implemented in the R package *bnlearn* (Scutari, 2010).

During structure learning and the parameterization of the Bayesian model for drilling, we have developed a secondary data resource that includes the full reconstruction of the missing data present in the EDGE datasets. We used the R package *mice* (van Buuren and Groothuis-Oudshoorn, 2011) for multivariate missing data imputation. A set of random forest models were trained to capture the interrelationships among all operation parameters, and then used to impute missing values. The new dataset has been included in the EDGE data repository to support additional model development efforts. Figure 9 compares the data distributions corresponding to observed and imputed data for each variable containing missing information.

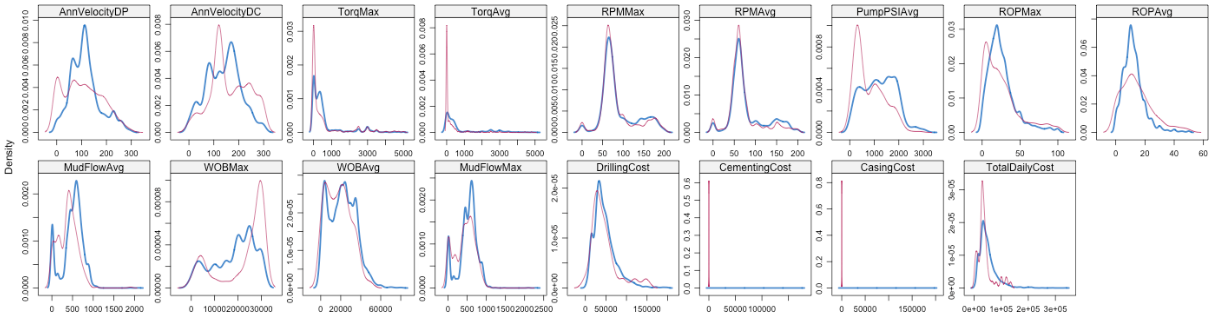


Figure 9: Missing data reconstruction. Blue lines represent the density of the observed data whereas the magenta lines represent imputed data.

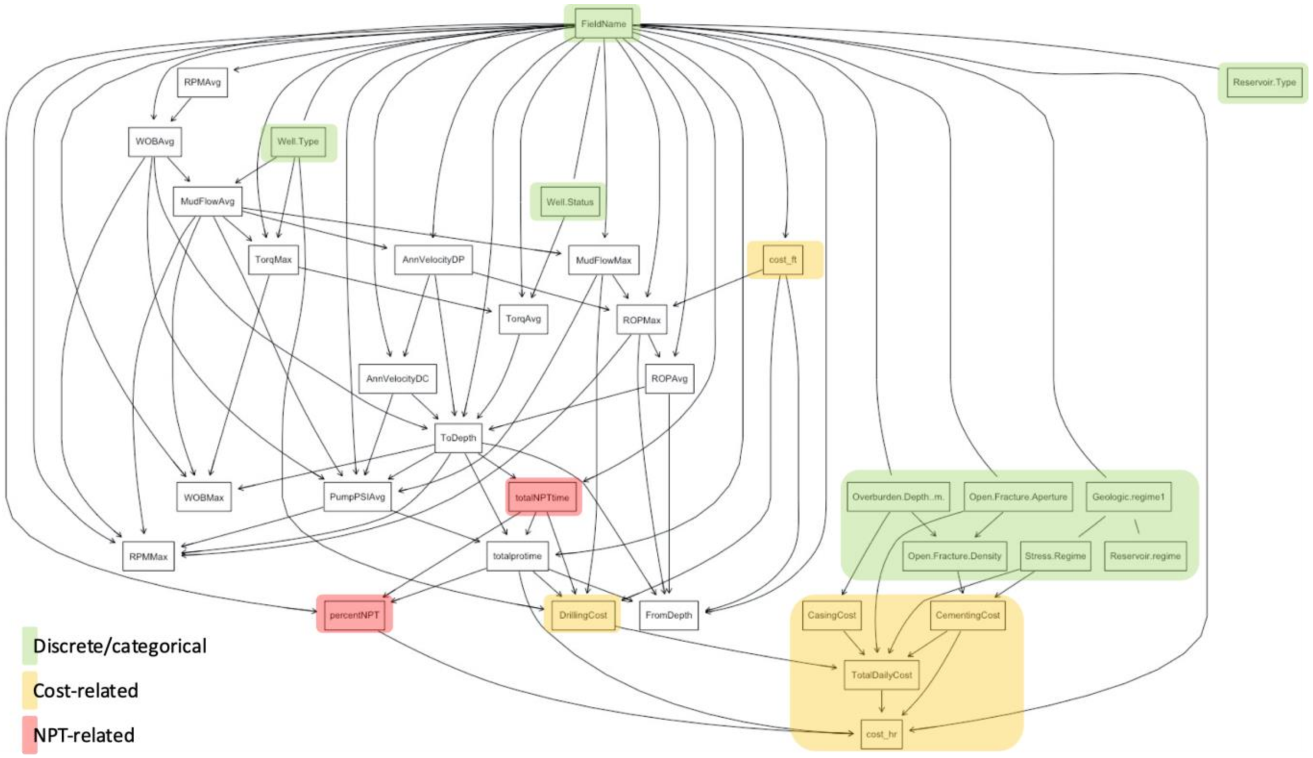
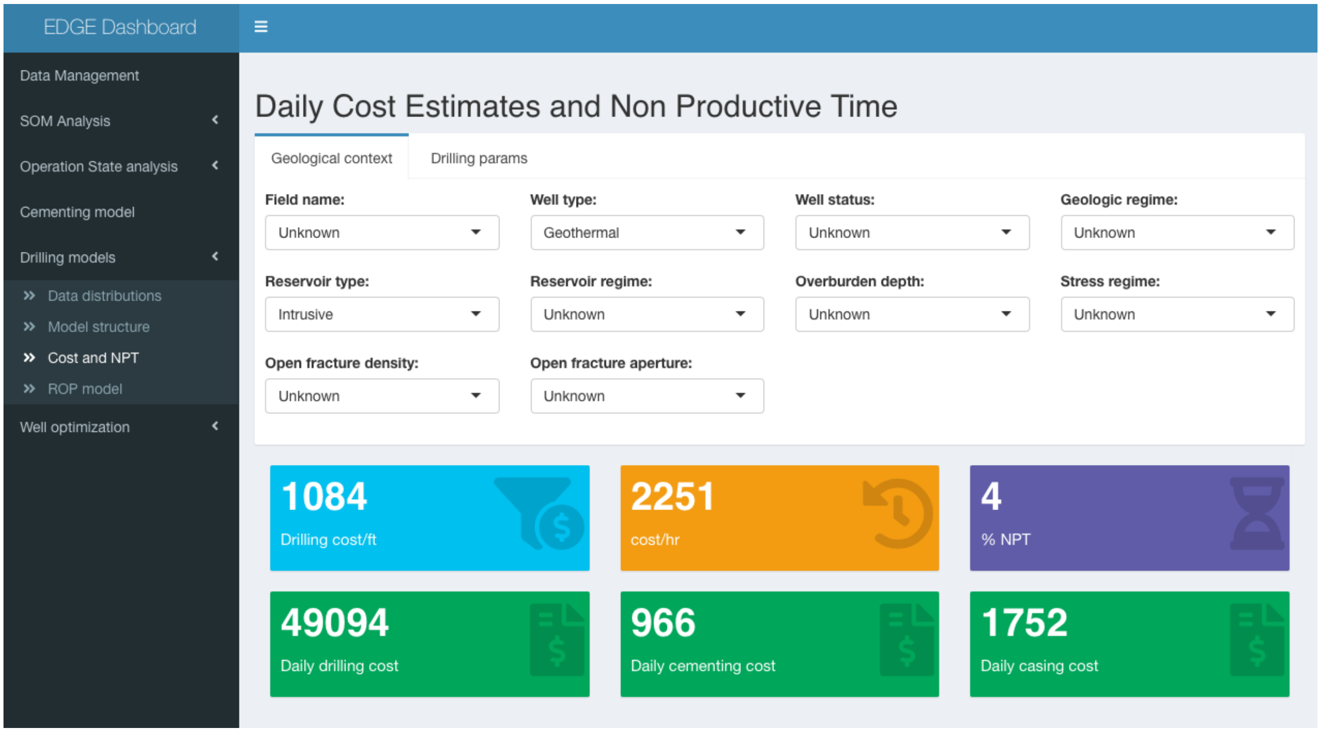


Figure 10: Learned Bayesian network from geothermal operation data. The set of features used for adjusting the model include a combination of continuous and discrete variables.

Figure 10 depicts the structure of the Bayesian model learned from the well data in the EDGE data repository. In addition to setting the basis for probabilistic predictive modeling, the structure of the Bayesian network provides unvaluable information regarding the most influential drilling parameters that drive operation cost and performance. Although the relationships that stem from a Bayesian network cannot be interpreted directly as causal relations, it is interesting to note that based on the data, casing and cementing cost are mostly dependent on features related to the geological context. Similarly, the drilling costs are mainly driven by the well type (geothermal, injection, etc), non-productive time, and depth. Other variables that are also indirectly related to drilling cost include ROP and WOB. Non-productive time (NPT) is mostly influenced by the depth (which also accounts for the effects of other operation variables) and the geothermal field (which integrates information on the geological context).

Many of the costs attributed to drilling are time-dependent, therefore anything that speeds up the hole advance without compromising safety, hole stability, or directional path is beneficial. Increasing ROP contributes to reduce drilling time, however increased ROP may result in more trips and shorter equipment life. Bit performance is an important factor to increase drilling speed and extend equipment life. However, optimizing drilling performance is complex and requires historical data for similar geological formations and well operation regimes.

The Bayesian models have been integrated in the EDGE dashboard, which is designed as a web-based application that provides integrated access to all the tools and models developed in the project. Figure 11 shows details of the current prototype that integrates predictive models developed in R and Python.



Daily Cost Estimates and Non Productive Time

Field name: Unknown

Well type: Geothermal

Well status: Unknown

Geologic regime: Unknown

Reservoir type: Intrusive

Reservoir regime: Unknown

Overburden depth: Unknown

Stress regime: Unknown

Open fracture density: Unknown

Open fracture aperture: Unknown

1084 Drilling cost/ft

2251 cost/hr

4 % NPT

49094 Daily drilling cost

966 Daily cementing cost

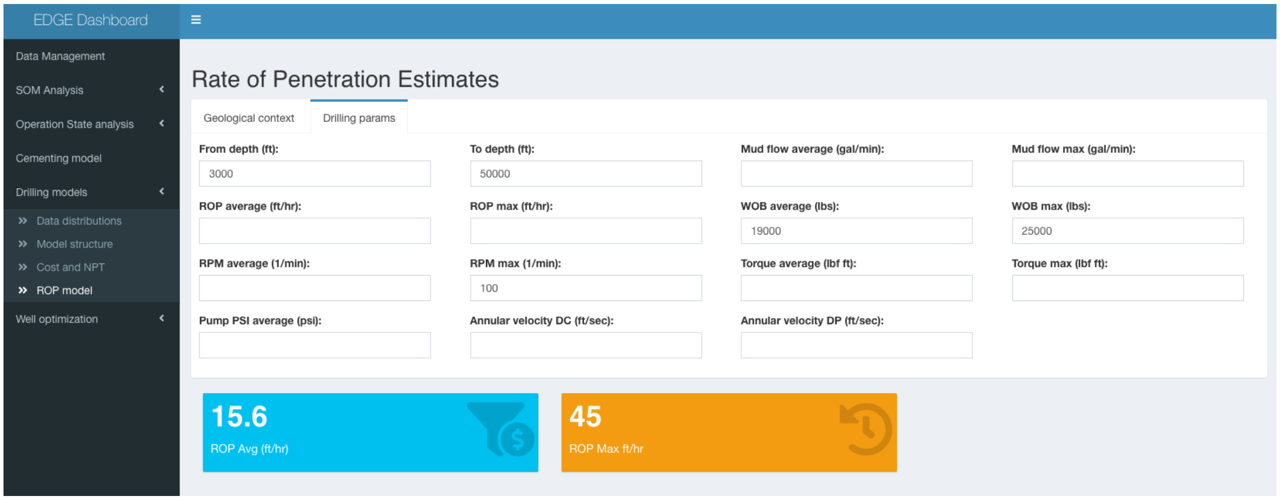
1752 Daily casing cost

Figure 11: Web interface of the Bayesian model to estimate drilling cost and non-productive time

Input information for the Bayesian network includes information on the geological context and the drilling operation parameters. The main advantage of the Bayesian approach is that provides implicit support for uncertainty on the input data, providing predictions when some of the parameters are missing or in the presence of uncertainties (e.g., 15000≤WOB Average≤19000). Bayesian models can also be used for more complex queries, including sampling from the distributions encoded in the Bayesian model to provide uncertainty estimates.

Figure 11 shows the web interface of a Bayesian network to estimate cost and non-productive time. Operators can provide the incomplete information and the Bayesian network will provide cost estimates based on the provided evidence and the drilling data used to parameterize the network. In this example, the only information provided is the well and reservoir types. With only this information the model estimates a drilling cost/ft of $1,084. As we add more evidence, for instance a WOB average of 20,000 lbs. and a WOB maximum of 30,000 lbs., the cost estimate is recomputed resulting in a lower value of $806/ft.

The same Bayesian network structure can be used to develop ROP models. Using the same initial evidence as before, the model provides an initial estimate for ROP average of 12.5 ft/hr and ROP maximum of 26 ft/hr. After adding information on the WOB, the new estimates for ROP average and maximum are of 15 ft/hr and 36 ft/hr, respectively. Figure 12 displays the form used in the ROP model to provide information on the drilling parameters to be used to generate the estimates. As with the geological context, model users are not required to provide information for all the parameters in the model. This flexibility of Bayesian models facilitates the exploration of scenarios in the presence of uncertain information.



Rate of Penetration Estimates

From depth (ft): 3000

To depth (ft): 50000

Mud flow average (gal/min):

Mud flow max (gal/min):

ROP average (ft/hr):

ROP max (ft/hr):

WOB average (lbs): 19000

WOB max (lbs): 25000

RPM average (1/min):

RPM max (1/min): 100

Torque average (lbf ft):

Torque max (lbf ft):

Pump PSI average (psi):

Annular velocity DC (ft/sec):

Annular velocity DP (ft/sec):

15.6
ROP Avg (ft/hr)

45
ROP Max ft/hr

Figure 12: ROP model for different geological contexts and drilling operation parameters

Models included in the EDGE dashboard can be used by the optimization modules to find a set of optimal parameters that minimize or maximize a specific cost function. For instance, we can use the optimization included in the EDGE Dashboard to fine tune the values of WOB average and RPM average to maximize ROP for a geothermal well. After running the optimization, and providing as sole evidence the well type, we obtain an optimal WOB of 19928 lbs. and RPM of 60.42 to yield an ROP of 12.5 ft/hr. It is important to note that the optimization process requires a good initial estimate that is then fine-tuned based on the estimates provided by the Bayesian network.

Table 2: Elements of the Pareto Set during the optimization of ROP and Cost

WOB (lbs.)	**RPM** (1/min)	**ROP** (ft/hr.)	**Cost/ft**
30000	101.42	12.84	$1038.1
30000	198.21	12.85	$1038.3
30000	197.52	12.83	$1034.2

For multi-objective optimization problems, the solution is not unique, and we obtain a collection of optimal values known as the Pareto set. In the context of geothermal drilling, multi-objective optimization can be used to select the optimal combination of values to optimize a set of cost functions. For instance, Table 2 shows the set of optimal WOB and RPM values that maximize ROP and minimize cost/ft in a generic geothermal well.

4. CONCLUSIONS

Data-driven methods constitute a promising approach to optimize geothermal drilling and for the early identification and mitigation of well failures. In the past two years, the EDGE project has assembled a data repository that contains operation and performance data for more than 100 geothermal wells located in different geological contexts. After a preliminary data analysis, the project has developed several data-driven approaches to predict relevant drilling operation parameters including ROP, non-productive time, and drilling cost. The models and analysis developed in the project are being integrated in a web-based platform (EDGE Dashboard) implemented in R using the Shiny framework. Models developed in Python are integrated using the R package *reticulate*.

In this paper we have presented several of the elements that compose the EDGE Dashboard and provided examples of use of these techniques in the context of analysis, modeling, and optimization of geothermal drilling. The components of the dashboard include: 1) self-organizing maps to visualize well operation data and process trajectories, 2) process mining to analyze drilling event logs and discover process models, 3) Bayesian models to predict relevant operation and performance metrics from incomplete and uncertain drilling data, and 4) algorithms to optimize well drilling under specific operational constraints.

Ongoing work is focused on using drilling data and machine learning techniques to identify main causes of well failure and to develop mitigation strategies aimed at reducing the impact of such failures. Drill bit failure was identified as a reason for frequent BHA trips – which in turn led to higher drilling costs. In a previous study (Witt-Doerring et al., 2021), it was found that when drilling hard formations with PDC bits, if a bit was damaged beyond repair (DBR) such as being rung out, subsequent bit runs in the same well performed worse than in cases where the previous bit was pulled before substantial damage. A similar study is being conducted here on the geothermal well dataset, the main difference being that roller cone bits (as opposed to PDC bits) were used in the wells being investigated. Additionally, the dataset is also being used to develop a methodology to determine bit pull criterion, such that drilling and tripping decisions can be optimized. Tool failure due to the higher-than-normal temperature is also an issue when drilling geothermal wells. Such failures occur when the mud temperature exceeds the rated temperature limits of the tools. Towards that end, we are developing models that can be used to predict downhole temperatures, such that the downhole heat can be properly managed during drilling operations.

ACKNOWLEDGEMENT

This research has been conducted in the framework of the EDGE project supported by the Department of Energy under Award Number DE-EE0008793 – 0000.

REFERENCES

Carbonari, R., Ton, D. Bonneville, A., Bour, D. Cladouhos, T., Garrison, G., Horne, R., Petty, S., Rallo, R., Schultz, A., Sorlie, C., Thorbjornsson, I., Uddemberg, M., Weydt, L.: First year report of EDGE project: An International Research Coordination Network for Geothermal Drilling Optimization Supported by Deep Machine Learning and Cloud Based Data Aggregation. Proceedings, 46th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, February 15-17, 2021.

Witt-Doerring, Y., Pastusek, P., Ashok, P., and van Oort, E.: Quantifying PDC Bit Wear in Real-Time and Establishing an Effective Bit Pull Criterion Using Surface Sensors, SPE Annual Technical Conference and Exhibition, 21-23 Sept 2021, SPE-205844-MS

Kohonen, T.: The Self-Organizing Map. *Proc. IEEE*, **78**, (1990), 1464-1480.

Aalst, W. van der, Weijters, A., Maruster, L.: Workflow Mining: Discovering Process Models from Event Logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9), (2004), 1128–1142.

W. van der Aalst: Process Mining - Data Science in Action, Second Edition. Springer, (2016).

Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, (1988).

Scutari, M: Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3), (2010), 1–22.

van Buuren, S., Groothuis-Oudshoorn, K.: Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), (2011), 1–67.

Binois M., Picheny, V.: GPareto: An R Package for Gaussian-Process-Based Multi-Objective Optimization and Analysis, *Journal of Statistical Software*, 89(8), (2019), 1-30.