

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

LUÍS FELIPE FERREIRA MOTTA BARBOSA

Drilling optimization of petroleum and natural gas wells: application of artificial
intelligence

Guaratinguetá-SP
2019

Luís Felipe Ferreira Motta Barbosa

Drilling optimization of petroleum and natural gas wells: application of artificial intelligence

Dissertação apresentada à Faculdade de Engenharia do Campus de Guaratinguetá, Universidade Estadual Paulista, para a obtenção do título de Mestre em Engenharia Mecânica na área de Energia .

Orientador: Prof Dr. João Andrade de Carvalho Júnior

Coorientador: Prof Dr. Andreas Nascimento

Guaratinguetá-SP

2019

B238d	Barbosa, Luís Felipe Ferreira Motta Drilling optimization of petroleum and natural gas wells : application of artificial intelligence / Luís Felipe Ferreira Motta Barbosa. – Guaratinguetá, 2019. 176 f : il. Bibliografia: f. 144-155 Dissertação (Mestrado) – Universidade Estadual Paulista, Faculdade de Engenharia de Guaratinguetá, 2019. Orientador: Prof. Dr. João Andrade de Carvalho Júnior Coorientador: Prof. Dr. Andreas Nascimento 1. Máquinas de perfuração 2. Poços de petróleo - Perfuração. 3. Gás natural 4. Indústria petrolífera I. Título. CDU 622.24.05 (043)
-------	---

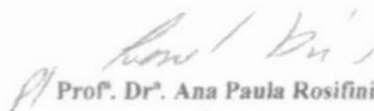
Pâmella Benevides Gonçalves
Bibliotecária/CRB-8/9203

LUIS FELIPE FERREIRA MOTTA BARBOSA

ESTA DISSERTAÇÃO FOI JULGADA ADEQUADA PARA A OBTENÇÃO DO TÍTULO DE
"MESTRE EM ENGENHARIA MECÂNICA"

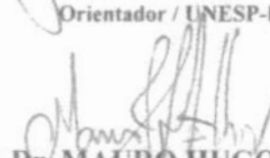
PROGRAMA: ENGENHARIA MECÂNICA
ÁREA: ENERGIA

APROVADA EM SUA FORMA FINAL PELO PROGRAMA DE PÓS-GRADUAÇÃO


Prof. Dr.ª Ana Paula Rosifini Alves Claro
Coordenadora

BANCA EXAMINADORA:


Prof. Dr. JOÃO ANDRADE DE CARVALHO JÚNIOR
Orientador / UNESP-FEG


Prof. Dr. MAURO HUGO MATHIAS
UNESP-FEG


Prof. Dr. JOSÉ LUIZ GONÇALVES
UEMG/Divinópolis

DADOS CURRICULARES

LUÍS FELIPE FERREIRA MOTTA BARBOSA

NASCIMENTO 05/06/1992 - Guaratinguetá / SP

FILIAÇÃO Fernando César Mendes Barbosa
Ana Cristina Ferreira Motta Barbosa

2010 / 2015 Curso de Graduação em Engenharia Mecânica na Faculdade de Engenharia de Guaratinguetá, Universidade Estadual Paulista

2017 / 2018 Curso de Pós-Graduação em Engenharia Mecânica, nível de Mestrado, na Faculdade de Engenharia de Guaratinguetá, Universidade Estadual Paulista

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisors Prof. Dr. João Andrade and Prof. Dr. Andreas Nascimento, for the support and guidance throughout the development of this study, to Dr. André Leibsohn from CENPES/Petrobras for inviting me for a technical visit and for conversations about drilling optimization, to Prof. Dr. Skalle Pål from Norwegian University of Science and Technology (NTNU) for his kind support in helping us to understand the drill-curves from Norway used in the current thesis, to Prof. Dra. Marília Inês Mendes Barbosa from *Universidade Federal de Uberlândia* for explaining me some relevant topics related to petroleum geology, to M.Sc. Luís César Ferreira Motta Barbosa (UNESP, Department of Production) for his important advice on relevant aspects of scientific methodology, to the colleagues, especially M.Sc. Fellipe Sartori, and all employees from UNESP, School of Engineering, Guaratinguetá, to the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) for the scholarship that I received, and, last but not least, to my parents and brothers for always supporting me.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

ABSTRACT

To meet the increasing primary energy demand, more challenging petroleum reserves started being explored, such as the reservoirs from pre-salt formation close to the Brazilian and Angolan coasts. Historically, low penetration rates in drilling the pre-salt carbonates were reported in the literature, resulting in large capital expenditure on well's construction. Since the major part of exploration cost is associated with drilling, optimizing this activity is of major importance. In this context, the main objective of the present thesis is to investigate methods for real-time drilling optimization of oil and natural gas wells. A common way to optimize drilling activities is to determine the optimum operational variables (e.g. weight-on-bit and rotational speed) that maximizes the ROP. However, this may yield a decrease in drilling efficiency. An alternative to reduce problems related to drilling inefficiency, such as excessive bit wear and vibrations, is through the selection of operational variables able to minimize the specific energy (SE) spent to excavate a volumetric unit of rock. For that, it is necessary to employ accurate predictive models able to capture how the operational variables (weight-on-bit, rotational speed, mud flow and so on) influence not only on ROP but also on SE. Therefore, the present thesis employed a well-known machine learning method, called random forest, instead of analytical equations found in drilling engineering books. Thus, it was possible to obtain accurate predictive models for ROP and SE, to be used, later, as objective functions in optimization problems to determine the optimum parameters, weight-on-bit and rotational speed. Real-time drilling data from pre-salt region and Norwegian continental shelf were employed. First, several aspects related to training process of random forests were investigated. Among them, it was confirmed the possibility of predicting the ROP with accuracy by employing only four inputs: depth, weight-on-bit, rotational speed, and mud flow. The prediction of SE was carried out by coupling the mathematical formulation with predictive models of ROP and torque (if available). Optimization problems were analyzed with one objective function, as well as with multiple objective functions through the ε -constraint technique. It was observed the sole maximization of ROP may lead to increase in the energy required to drill. However, by imposing the inequality $SE(\mathbf{x}) \leq SE_{actual} * \varepsilon$ on the maximization of ROP, it was possible to reduce significantly the amount of observations whose ROP increased due to detriment of drilling efficiency. For the minimization of SE problems, it was observed a special care to be taken when simulating low-values for weight-on-bit and rotational speed.

KEYWORDS: Multi-objective optimization. Drilling efficiency. Rate of penetration. Machine learning. Random forest.

RESUMO

Para atender à crescente demanda de energia primária, começaram a ser exploradas reservas de petróleo em áreas mais desafiadoras, tais como os reservatórios da formação do pré-sal próximos às costas brasileira e angolana. Historicamente, observa-se baixa taxa de penetração na perfuração dos carbonatos do pré-sal, resultando em altos custos na construção de poços. Como a maior parte dos custos de exploração está associado com perfuração, a otimização desta atividade é de grande importância. Neste contexto, o principal objetivo desta dissertação é investigar métodos de otimização em tempo-real de poços de petróleo e gás natural. Uma forma comum de se otimizar a perfuração é através da determinação dos parâmetros operacionais (peso na broca e rotação) que maximizem a taxa de penetração (*rate of penetration*, ROP). Contudo, isto pode acarretar na diminuição da eficiência do processo de perfuração. Assim, uma forma de diminuir problemas relacionadas a ineficiências da perfuração, tais como gasto excessivo da broca ou vibrações, é através da seleção dos parâmetros operacionais, minimizando a energia específica (*specific energy*, SE) gasta para escavar uma unidade volumétrica de rocha. Para tanto, é necessário o emprego de modelos precisos que relacionem como as variáveis operacionais (peso da broca, rotação, vazão do fluido de perfuração entre outros) influenciam, não somente o ROP, mas também a SE. Desde modo, a presente dissertação empregou um método conhecido de aprendizagem de máquinas, chamado de florestas aleatórias, em vez das equações analíticas comumente encontrados em livros de engenharia de perfuração. Assim, foi possível obter modelos de previsão precisos para ROP e SE, para, depois, serem utilizados como funções objetivas em problemas de otimização para seleção ótima dos parâmetros (peso na broca e rotação). Dados de perfuração da região do pré-sal e da plataforma continental norueguesa foram utilizados. Primeiramente, investigou-se diversos aspectos relacionados ao treinamento das florestas aleatórias. Entre eles, verificou-se a possibilidade de estimar com precisão o ROP utilizando apenas quatro parâmetros: profundidade, peso na broca, rotação e vazão de fluido. A previsão da SE, por sua vez, se fez através do acoplamento da formulação matemática com os modelos preditivos do ROP e torque (quando disponível). Foram investigados problemas de otimização contendo tanto uma função objetiva quanto problemas com múltiplas funções objetivas através da técnica ε -constraint. Verificou-se que a maximização sozinha da taxa de penetração pode acarretar em aumento da energia gasta para se perfurar. Contudo, ao impor como restrição a inequação $SE(\mathbf{x}) \leq SE_{actual} * \varepsilon$ ao problema da maximização da taxa de penetração, foi possível diminuir consideravelmente a quantidade de observações que o aumento do ROP se deu através do detrimento da eficiência da perfuração. Para o problema da minimização da SE, constatou um cuidado que se deve ter ao simular combinações de peso na broca e rotação com valores baixos.

PALAVRAS-CHAVE: Otimização multiobjetiva. Eficiência da perfuração. Taxa de penetração. Aprendizagem de máquinas. Florestas aleatórias.

LIST OF FIGURES

Figure 1.1	World primary energy demand by fuel type in million of barrels of oil equivalent per day (mboe/d)	21
Figure 1.2	Monthly prices of crude oil in USD/barrel traded in Dubai Fateh and West Texas Intermediate (WTI)	22
Figure 1.3	a) Santos basin geology with pre-salt source rocks (carbonate reservoirs) beneath the thick layer of evaporites; b) potential drilling problems when drilling the evaporites	23
Figure 1.4	Drilling optimization cycle	24
Figure 2.1	The rotary drilling process	27
Figure 2.2	Classification of drilling variables. The symbol ✓ indicates that those variables are chosen for mathematical optimization.	28
Figure 2.3	Typical drill bits: a-b) roller cone bits; c-e) fixed cutter bits	29
Figure 2.4	Drilling variables that affect the ROP.	31
Figure 2.5	Drilling mechanical parameters influences' on penetration rate (ROP)	32
Figure 2.6	Schematic description of optimum zone	36
Figure 2.7	Five levels of process control and optimization (time scales for each level) .	38
Figure 2.8	a) Typical drill-rate test data showing non-linear response below the minimum depth of cut and above the founder point; b) Bit weight (WOB) test carried during a run	39
Figure 2.9	Some possible objective functions for drilling optimization	40
Figure 2.10	Partitions and CART	44
Figure 2.11	Framework of conventional ensemble methods. The solid blue lines show the parallel flow of bagging, random subspace and Random Forest. The dashed red lines in generation and base prediction parts denote boosting ensemble framework.	47
Figure 2.12	Input space and objective space for the case of two objective functions . . .	51
Figure 3.1	Approaches for ROP modeling	57
Figure 3.2	A compilation of 58 works found on the literature considering thesis and papers (publications in journals and congresses) using other methods different than the traditional models for ROP prediction. Status: October 2018. . . .	57
Figure 3.3	Using neural networks to assess parametric analysis: a) effect of WOB, b) mud-weight (MW), and c) plastic viscosity on ROP	63
Figure 3.4	Contour plot for ROP versus normalized values of A) depth (X_1) and weight on bit (X_2), B) depth (X_1) and bit rotation speed (X_3)	64

Figure 3.5	Variable importance: <i>a</i>) automated feature selection resulted applied by Eskandarian, Bahrami and Kazemi (2017) with <i>fscaret</i> , <i>b</i>) based on random forest ROP predictor performed by Hegde et al. (2017)	65
Figure 3.6	Data Partition developed at UT Austin	67
Figure 3.7	Possible approaches to treat measurement errors in drilling data that can be applied to drilling data analysis	69
Figure 3.8	Simplifying the multicriteria optimization into a single objective functions. Response surfaces form the basis for the objective function (OBJ).	73
Figure 4.1	Raw recording of drill curves from Well H. The specific energy was calculated based on Taele's formulation.	79
Figure 4.2	A sample of real-time drilling data considering the first approx. 8 hours of available recording — Well 4 from Norway	82
Figure 4.3	Data Pre-Treatment Method	84
Figure 4.4	After a change from not-drilling to rotary drilling, the transient state is identified.	88
Figure 4.5	MAPE against the cut-off threshold, which enables to avoid the problem of small values in computing this relative metric.	91
Figure 4.6	Equidistant grid points between lower and upper limits of a variable <i>x</i>	97
Figure 4.7	Optimization workflow adopted in the current work.	98
Figure 5.1	Comparison between ROP Models	102
Figure 5.2	Histograms of R-correlation on testing dataset.	105
Figure 5.3	The initial population influenced by the random starting for the Well 8 and 10. For both cases, the cardinality of possible inputs was 25.	107
Figure 5.4	Testing R-correlation - sensitive analysis of the train ratio.	109
Figure 5.5	Cumulative out-of-bag MSE for ROP-prediction: random forests trained using the same procedure from the Experiment 01.	110
Figure 5.6	Data-driven models against coupled models for SE prediction	111
Figure 5.7	Histograms of SE calculated based on Teale's formulation, considering four levels of ROP-threshold - well 8	112
Figure 5.8	Maximum value of SE for each well against the ROP-threshold	113
Figure 5.9	SE data-driven models: actual SE against prediction intervals on test dataset.	116
Figure 5.10	Density scatter assessing the SE coupled-model based on Rabia's formulation for the Well 10 - coupling with ROP predictive-model.	118
Figure 5.11	Maximization of ROP and its influence on specific energy - Well D	123
Figure 5.12	Maximization of ROP: histograms of difference between the optimized parameters and actual values	124
Figure 5.13	Maximization of ROP: relative amount of observations with improvement in ROP versus relative amount of observations with drilling-efficiency improved as consequence of WOB and RPM optimization	126

Figure 5.14	Minimization of SE: histograms of difference between the optimized parameters and actual values - Well B	127
Figure 5.15	Minimization of SE: relative amount of observations with improvement in SE versus relative amount of observations with ROP improved as consequence of WOB and RPM optimization	129
Figure 5.16	Influence of ε on the relative amount of observations that could be optimized for the problem of minimization of $SE(\mathbf{x})$ subject to $ROP(\mathbf{x}) \geq ROP(\mathbf{x}) * \varepsilon$.130	
Figure 5.17	Minimization of $SE(\mathbf{x})$ subject to $ROP(\mathbf{x}) \geq ROP(\mathbf{x}) * \varepsilon$ - well D with not realistic optimum solutions.	132
Figure 5.18	Minimization of $SE(\mathbf{x})$ subject to $ROP(\mathbf{x}) \geq ROP(\mathbf{x}) * \varepsilon$ - optimum solutions for the well D with new lower limits.	134
Figure 5.19	Minimization of $SE(\mathbf{x})$ subject to $ROP(\mathbf{x}) \geq ROP(\mathbf{x}) * \varepsilon$ - relative amount of optimized observations.	136
Figure 5.20	Influence of ε on the relative amount of observations that could be optimized for the problem of maximization of $ROP(\mathbf{x})$ subject to $SE(\mathbf{x}) \leq SE(\mathbf{x}) * \varepsilon$ 137	
Figure 5.21	Maximization of $ROP(\mathbf{x})$ subject to $SE(\mathbf{x}) \leq SE(\mathbf{x}) * \varepsilon$ - optimum solutions for the well D	138
Figure 5.22	Maximization of $ROP(\mathbf{x})$ subject to $SE(\mathbf{x}) \leq SE(\mathbf{x}) * \varepsilon$ - relative amount of optimized observations.	140
Figure A.1	Getting the sequence of the variable names in <code>RTDD.cuve_info</code> ; showing an example for the well 6. The first line of <code>RTDD.cuve_info</code> indicates that the variable 'Time' is recorded in the first column; the variable 'DBTM' is recorded in the second column and so on	156
Figure A.2	Real-time drilling data of well 6 considering the first approx. 8 hours of available recording according to the informed sequence of variable names – state: “how it is”. It is possible to observe that the HKL is acting like the BPOS, and the TRQ like the RPMA.	159
Figure A.3	Real-time drilling data of well 6 considering the first approx. 8 hours of available recording according to our suggestion for the variable names – Our suggestion for the sequence of the variable names	160
Figure A.4	Graphical Study of the real-time drilling data – well 6 – file 1.	161
Figure B.1	Raw recording of drill curves from Well A. The specific energy was calculated based on Rabia’s formulation.	162
Figure B.2	Raw recording of drill curves from Well B. The specific energy was calculated based on Taele’s formulation.	163
Figure B.3	Raw recording of drill curves from Well D. The specific energy that was calculated based on Rabia’s formulation.	164
Figure C.1	Parametric influence of Hampel filter on detecting outliers. Some missing values (not-a-number instances) are detected as outliers.	166

Figure C.2	Relative amount of outliers for different levels of window half-width and thresholds (TH) - all drill curves from pre-salt.	167
Figure C.3	Relative amount of outliers for different levels of window half-width and thresholds (TH) - all drill curves from Norway, excluding the well 2.	167
Figure D.1	Visual approach to determine the thresholds for the automated operation recognition. For illustration purpose, drill curves from well 8 are plotted. Similar analysis for all other wells was carried out.	168
Figure D.2	An example for a transient state correctly identified - well 6 - first change identified.	169
Figure D.3	Reducing the ε_1 from 0.2m to 0.15m could improve the transient state recognition: a) the second change was not recognized when ε_1 was 0.2 m; b) the second change was recognized after reducing the tolerance to 0.15 m - well 6 - eighth change identified.	170
Figure D.4	An example for a transient states identified from not-drilling to rotary drilling, and vice-verse - well 6 - sixteenth change identified.	170
Figure D.5	The transient state detection could avoid considering some observations as rotary drilling - well 6 - twenty-first change identified.	171
Figure E.1	Two dimensional histogram of ROP against WOB for the Well 3: a) filtered data after third step with all drilling states, b) after fourth step with only rotary drilling state, c) after validation step.	172
Figure E.2	Two dimensional histogram of ROP against RPM for the Well 3: a) filtered data after third step with all drilling states, b) after fourth step with only rotary drilling state, c) after validation step.	173
Figure E.3	Two dimensional histogram of ROP against WOB for the Well 4: a) filtered data after third step with all drilling states, b) after fourth step with only rotary drilling state, c) after validation step.	173

LIST OF CODES

Code 4.1 – General use of <code>hampel</code> function, adapted from Mathworks (2018a)	86
Code 4.2 – Example of hyperparameters tuning with Bayesian optimization for the <code>TreeBagger</code> object. Adapted from Mathworks (2018e).	93
Code 5.1 – Function to initialize the population. Source: Oluleye et al. (2014a)	106

LIST OF TABLES

Table 3.1 – ROP modeling using multiple regression carried out by Moraveji and Naderi (2016)	59
Table 3.2 – Some works that used Machine Learning Methods to predict the ROP.	60
Table 3.3 – The use ROP models to optimize the drilling variables, considering a single-objective problem	70
Table 4.1 – Overview of the drilling data from pre-salt formation. The abbreviation meanings are found in Annex A	77
Table 4.2 – Amount of missing values in each variable (column). For illustration purpose, statistics only from the Well H is presented. This dataset has a total of 5 184 observations and 15 variables, therefore the total amount of values is $5\,184 \cdot 15 = 77\,760$	78
Table 4.3 – Amount of observation with k-variables as missing values – Well H	78
Table 4.4 – Relative amount of observations with k-variables as missing values	79
Table 4.5 – Real-time drilling data from Norway published by Donne (2017). The abbreviation meaning used in these dataset can be found in Annex B	81
Table 4.6 – Rules to determine drilling modes for real-time drilling data in time domain .	87
Table 4.7 – Drilling parameters required for the development of the current work.	94
Table 5.1 – Data pre-treatment setting	99
Table 5.2 – Amount of observations for each well from the original dataset to the final dataset after the data pre-treatment.	100
Table 5.3 – Simple comparison between the Bourgoyne and Young Model and a random forest regression for ROP prediction, using the drilling data of the Well B. The standard deviation is given in parentheses.	101
Table 5.4 – Inputs selected for each experiment. The inputs of the Experiment 02 consist of adding the listed variables to the group of the Experiment 01.	104
Table 5.5 – Average of evaluation metrics for ROP prediction on testing dataset.	104
Table 5.6 – Options for Feature Selection based on Binary Genetic Algorithm	106
Table 5.7 – Best subsets obtained by binary genetic algorithm.	108
Table 5.8 – Drilling variables used to calculate the SE	113
Table 5.9 – Inputs selected for the not-coupled and coupled models to predict the SE. . .	114
Table 5.10–SE prediction based on data-driven models without the core variables ROP and torque as inputs: average of evaluation metrics on train and test dataset. .	115
Table 5.11–SE prediction based on data-driven models with the core variables ROP and torque as inputs: average of evaluation metrics on test dataset.	117
Table 5.12–Evaluation metrics on test datasets for SE coupled-models based on Rabia formulation: ROP predictive model against inverse-ROP predictive model . .	120

Table 5.13–Evaluation metrics on test dataset for SE coupled-models: a comparison between Teale’s and Rabia’s formulation.	121
Table 5.14–Maximization of ROP - some statistics about the difference between optimized and actual-variable values	125
Table 5.15–Minimization of SE - some statistics about the difference between optimized and actual variable-values	128
Table 5.16–Problems identified in the third optimization problem	133
Table 5.17–Minimization of $SE(\mathbf{x})$ subject to $ROP(\mathbf{x}) \geq ROP(\mathbf{x}) * \varepsilon$ - some statistics about the difference between optimized and actual variable-values	135
Table 5.18–Maximization of $ROP(\mathbf{x})$ subject to $SE(\mathbf{x}) \leq SE(\mathbf{x}) * \varepsilon$ - some statistics about the difference between optimized and actual variable-values	139
Table 5.19–Elapsed time in seconds for training phase (including data pre-treatment) and grid-search in the optimization algorithm	141
Table A.1 – Sequence in which the variable names appear in the drilling data with 10 variables recorded (that is, the 2nd group). We highlighted the columns with the problem	157
Table A.2 – Suggestion for the variable names of the well 6. We believe that the sequence of the variable names for the well 6 should be the same as the wells 4, 5 and 7, which can be achieved by only sliding the RPM to the column # 6, and moving the yellow block downwards	158
Table F.1 – Torque prediction based on random forests: average of evaluation metrics on train and test dataset.	174
Table A.1 – Mnemonics employed in drilling data from Pre-Salt	175
Table B.1 – Mnemonics employed in drilling data from Norway	176

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Networks
CCS	Confined Compressive Strength
DSATS	Drilling Systems Automation Technical Section
ECD	Equivalent Circulating Density
ESD	Equivalent Static Density
GA	Genetic Algorithm
GBM	Gradient Boosting Machine
HMSE	Hydro-Mechanical Specific Energy
HP/HT	High Pressure and High Temperature
IEA	International Energy Agency
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MAPE _a	Adapted Mean Absolute Percentage Error
MME/EPE	<i>Ministério de Minas e Energia - Empresa de Pesquisa Energética</i>
MSE	Mean Squared Error
MWD	Measurements While Drilling
OPEC	Organization of the Petroleum Exporting Countries
PSO	Particle Swarm Optimization
RF	Random Forests
RMSE	Root Mean Squared Error
ROP	Rate of Penetration
RTDD	Real-Time Drilling Data
SE	Specific Energy

SPE	Society of Petroleum Engineers
UCS	Unconfined Compressive Strength
WOB	Weight on Bit
WTI	West Texas Intermediate

CONTENTS

1	INTRODUCTION	21
1.1	CURRENT SCENARIO	21
1.1.1	Challenges of Pre-Salt	22
1.2	DRILLING OPTIMIZATION	23
1.2.1	Drilling Optimization Based on Predictive Models and Related Works .	24
1.3	OBJECTIVES	26
1.4	WORK STRUCTURE	26
2	THEORETICAL BACKGROUND	27
2.1	OVERVIEW OF DRILLING ENGINEERING	27
2.1.1	Rotary Drilling	27
2.1.2	Drilling Variables	28
2.1.3	Performance Indicators	30
2.1.3.1	Rate of Penetration	30
2.1.3.2	Specific Energy	33
2.1.3.3	Drilling Efficiency	34
2.1.3.4	Some indicators of drilling problems	35
2.1.3.5	Economics	36
2.1.4	Optimization of Controllable Drilling Variables	37
2.2	REGRESSION MODELS BASED ON MACHINE LEARNING METHODS	40
2.2.1	Regression Analysis	41
2.2.2	Ensemble Learning	42
2.2.3	Base Learners	43
2.2.4	Parallel Learning with Bagging and Random Forests	45
2.2.5	Sequential Learning with Boosting	46
2.2.6	Other Ensemble Methods	47
2.2.7	Hyperparameter Optimization	48
2.2.8	Feature Selection	49
2.3	INTRODUCTION TO MULTI-OBJECTIVE OPTIMIZATION	49
2.3.1	The Single-Objective Optimization Problem	49
2.3.2	The Multi-Objective Optimization Problem	50
2.3.3	Multi-Objective Optimization Techniques	51
2.3.4	Decision Making Techniques	52
2.3.4.1	Global Criteria Method	52
2.3.4.2	Weighted Sum Method	53

2.3.4.3	The ε -constraint method	54
2.3.5	Optimization Techniques	54
2.3.6	Brief Remark on Multi-Objective Optimization Algorithms	55
3	MACHINE LEARNING METHODS APPLIED TO RATE OF PENE- TRATION PREDICTION AND OPTIMIZATION - A REVIEW	56
3.1	METHODS USED IN ROP PREDICTION	56
3.1.1	Traditional Models	57
3.1.2	Statistical Models	59
3.1.3	Machine Learning Applied to Predict the ROP	59
3.2	DISCUSSION ON ROP MODELS	61
3.2.1	Machine Learning Algorithms Outperforms Other Methods	61
3.2.2	Sensitive Analysis	62
3.2.3	Feature Extraction	64
3.2.4	Importance of Geological Formation in ROP Prediction	64
3.2.5	Data Partition	66
3.2.6	Drilling Data	66
3.2.7	Handling Measurement Errors	68
3.3	DRILLING OPTIMIZATION BASED ON PREDICTIVE MODELS	70
3.3.1	Single-Objective Optimization	70
3.3.2	Multi-Objective Optimization	71
3.3.3	The Need of Changing the Current Mindset	73
3.4	SUMMARY	74
4	METHODS	76
4.1	MATERIALS	76
4.1.1	Drilling Data from Pre-Salt	76
4.1.2	Drilling Data from Norway	80
4.1.3	Software and Hardware	83
4.2	PREDICTION DRILLING PERFORMANCE-INDICATORS	83
4.2.1	Data Pre-Treatment	83
4.2.2	Training the Predictive Models	88
4.2.3	Evaluation Metrics	89
4.2.4	Hyperparameters Optimization	92
4.2.5	Feature Selection	93
4.3	DRILLING PARAMETERS OPTIMIZATION	94
4.3.1	Single-Objective Optimization	94
4.3.2	Multi-Objective Optimization	95
4.3.2.1	Third optimization problem	95
4.3.2.2	Fourth optimization problem	95

4.3.3	Grid-Search Strategy	96
4.3.4	Optimization Workflow	97
5	RESULTS AND DISCUSSION	99
5.1	RATE OF PENETRATION PREDICTION	99
5.1.1	Data Pre-Treatment Setting	99
5.1.2	Simple Comparison between ROP Models	100
5.1.3	Selection of Most Important Inputs in the Prediction Task	102
5.1.3.1	Driller's Approach	103
5.1.3.2	Evolutionary Feature Selection	105
5.1.4	Ratio Influence of Training Dataset to Overall Dataset	108
5.1.5	Cumulative MSE of Random Forests	109
5.2	SPECIFIC ENERGY PREDICTION	110
5.2.1	Calculating SE	111
5.2.2	Data-driven models for the Specific Energy	114
5.2.3	Coupled Models for the Specific Energy	117
5.2.3.1	Rabia's Model	118
5.2.3.2	Teale's Model	120
5.2.4	Final Remark Regarding SE modeling	121
5.3	DRILLING PARAMETERS OPTIMIZATION	122
5.3.1	Single-Objective Optimization	122
5.3.1.1	First Optimization Problem - maximization of ROP	122
5.3.1.2	Second Optimization Problem - minimization of SE	126
5.3.2	Multi-objective optimization	129
5.3.2.1	Third Optimization Problem - minimization of SE, transforming maximiza- tion of ROP into an inequality	129
5.3.2.2	Fourth Optimization Problem - maximization of ROP, transforming minimiza- tion of SE into an inequality	136
5.3.3	On the Implementation's Feasibility of Proposed Optimization Methods	140
6	CONCLUSIONS	142
	REFERENCES	144
	APPENDIX A – NEW SUGGESTION FOR THE VARIABLE NAMES OF THE WELL 6 FROM NORWAY	156
	APPENDIX B – DRILL-CURVES FROM PRE-SALT	162

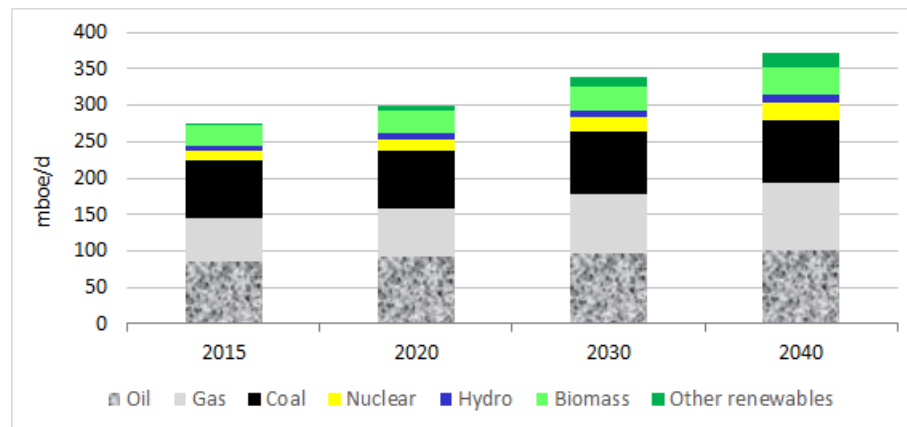
APPENDIX C – DATA PRE-TREATMENT ANALYSIS OF HAMPEL FILTER	165
APPENDIX D – DATA PRE-TREATMENT ANALYSIS OF OPERA- TION RECOGNITION	168
APPENDIX E – DATA PRE-TREATMENT ANALYSIS OF VALIDA- TION DRILL CURVES	172
APPENDIX F – TORQUE PREDICTION	174
ANNEX A – ABBREVIATIONS FOR REAL-TIME DRILLING DATA FROM PRE-SALT	175
ANNEX B – ABBREVIATIONS FOR REAL-TIME DRILLING DATA FROM NORWAY	176

1 INTRODUCTION

1.1 CURRENT SCENARIO

The energy supply is continuously increasing. From 1973 to 2015, the growth of world total primary energy supply was 124% (IEA, 2017a). Despite the increasing interest in renewable energies, the fossil fuels still have relevance as sources of primary energy. The Organization of the Petroleum Exporting Countries (OPEC) expects that the oil and natural gas demand increases on short-, medium- and long-term, as seen in Figure 1.1; for 2040, it is expected the oil and natural gas meet around 52% of world energy demand (OPEC, 2017).

Figure 1.1 – World primary energy demand by fuel type in million of barrels of oil equivalent per day (mboe/d)

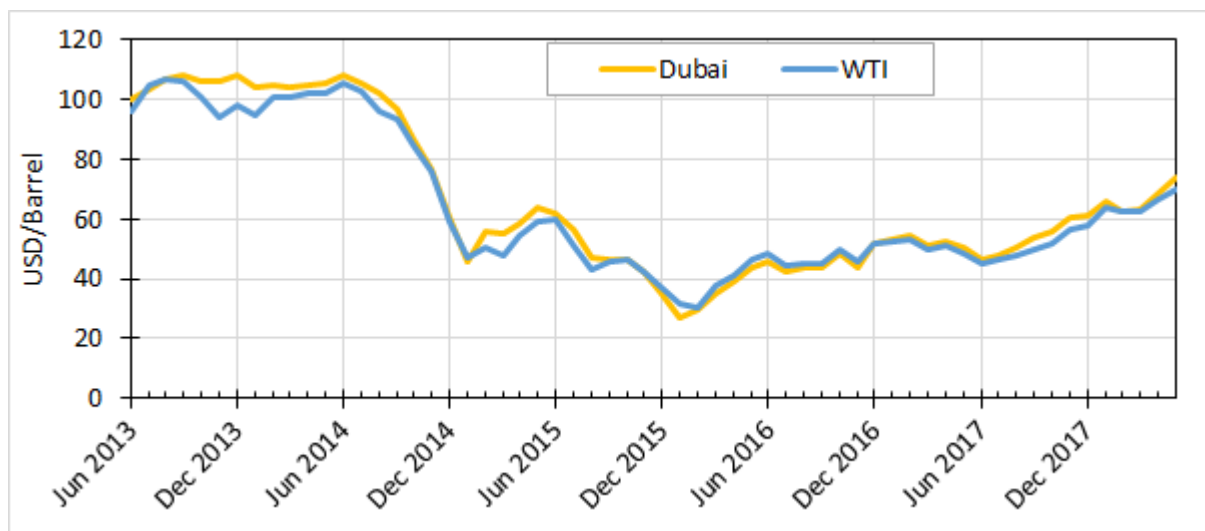


Source: OPEC (2017)

In 2014, the oil companies underwent a new scenario, which led to a considerable fall of oil prices, as can be seen in Figure 1.2. Many researchers claim that the reason for the oil plunge was the United States (US) shale oil boom (BYRNE; LORUSSO; XU, 2018; KIM, 2018; KHAN, 2017). However, another study stated that the reason behind this fall is the weakening global oil demand (PREST, 2018). Despite this controversy, it is hard to believe that a sudden change in world energy matrix will occur in next years or medium-term future. The International Energy Agency mentioned that the era of oil is not over, and, in a long-term future, energy supply expansion will be led by natural gas and renewable energies (IEA, 2017b). Since last year, the oil prices have been increasing.

To supply the future demand, investments in oil and gas exploration are required. With pre-salt discoveries, Brazil has become one of most important non-OPEC countries in terms of oil and gas production (OPEC, 2017), attracting the attention of the world. Several foreign oil companies participated in the recent bidding rounds of Brazilian pre-salt (NUNES; LIMA, 2017). For example, one of these companies was Equinor (former Statoil), that plans to invest in Brazil US 15 billions dollars until 2030 (FURLAN, 2018).

Figure 1.2 – Monthly prices of crude oil in USD/barrel traded in Dubai Fateh and West Texas Intermediate (WTI)



Source: IndexMundi (2018a) and IndexMundi (2018b).

The production of petroleum will sharply increase in Brazil, reaching 5.2 millions of barrel per day by 2026, and pre-salt fields will be responsible for 74% of the total production (MME/EPE, 2017, p. 154-155). Due to several difficulties in drilling, optimization is crucial to the future exploration and development of pre-salt areas.

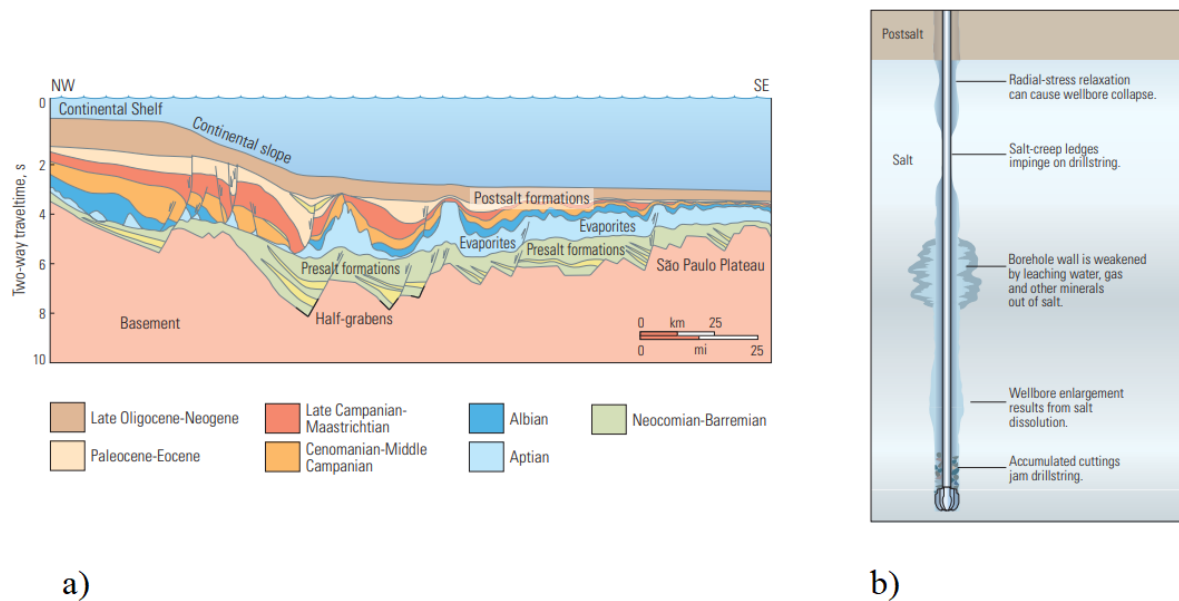
1.1.1 Challenges of Pre-Salt

The oil and gas industry has started exploring reservoirs in more challenging environments, such as the petroleum reserves from pre-salt formation close to the Brazilian and Angolan coasts (NASCIMENTO, 2016).

Wellbore construction in these areas is associated with several challenges. For example, Brazilian pre-salt reservoirs are located about 300 km from the coast in ultradeep waters. The varying composition of thick salt layer (evaporite) brings several challenges related to drilling, because each composition has different creep rates, leading to several problems (see Figure 1.3). Beneath the evaporite interval, the heterogeneous nature of layered carbonates with silica nodes affects drilling progress of pre-salt reservoirs (FRAGA et al., 2015; BEASLEY et al., 2010). Historically, low penetration rates were reported in drilling the pre-salt carbonates, resulting large capital expenditure on well's construction (NASCIMENTO et al., 2016).

Therefore, optimization of drilling operations is of major importance, and one way to achieve it is by finding the optimum combination of drilling parameters that can enhance the overall efficiency. Nascimento (2016) reported potential savings of millions of dollar by optimizing drilling activities in pre-salt regions.

Figure 1.3 – a) Santos basin geology with pre-salt source rocks (carbonate reservoirs) beneath the thick layer of evaporites; b) potential drilling problems when drilling the evaporites



Source: Beasley et al. (2010).

1.2 DRILLING OPTIMIZATION

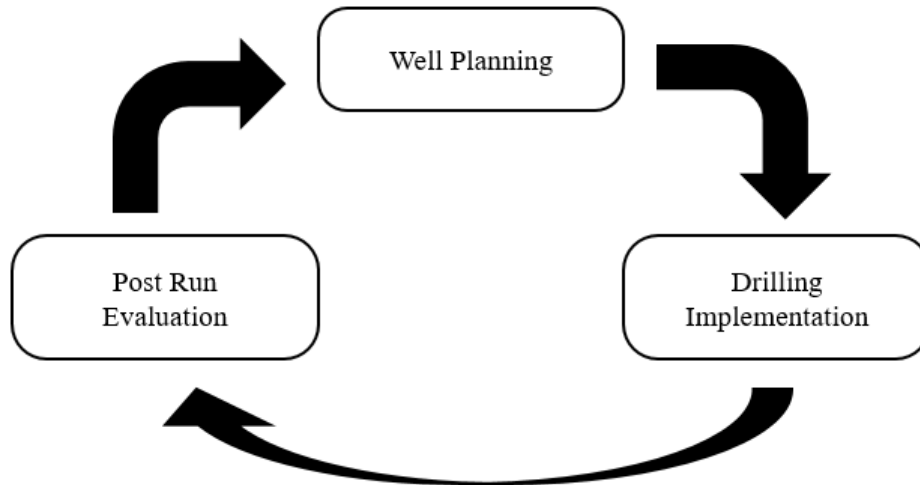
According to Lyons and Plisga (2004, p. 4.363), “the objective of optimizing drilling practices is to safely deliver a product capable of highest production capacity in a cost-effective manner”. The authors stated also that the drilling optimization must consider different aspects, namely, health, safety and environment (HSE), and production capability as well. That means, the selection of drilling actions must be taken considering the whole process.

The major part of exploration cost is from drilling, and this has direct impact on profitability of existing fields and cost of exploration (SKJERPEN et al., 2018). According to Lyons and Plisga (2004, p. 4.367), “most of well drilling cost is time dependent rather than product cost dependent”. Therefore, one of the main goals of drilling optimization is to reduce the total time, and enhance drilling efficiency.

The Figure 1.4 illustrates a drilling optimization cycle, which consists of well planning, followed by the implementation of wellbore construction. Then, post-run evaluations are responsible to identify good and bad practices, providing valuable information for the next well to be drilled, completing this cycle. Even if all phases are carefully carried out, unforeseen events may occur, so that corrections in real-time will be required. The understanding observed trends can yield positive impacts (LYONS; PLISGA, 2004).

Such trends can be obtained through drilling data analysis, which should be an integral part of well planning and operations. According to Staveley and Thow (2010), this is not the case despite the high costs of measuring and collecting drilling data. In order to provide important contributions to this industry, many researchers are studying methods to use drilling data for

Figure 1.4 – Drilling optimization cycle



Source: Lyons and Plisga (2004)

optimization purposes.

The presented thesis deals with development of real-time optimization methods of some controllable drilling variables, especially the weight on bit and rotating speed, based on data analysis, as explained in the following section.

1.2.1 Drilling Optimization Based on Predictive Models and Related Works

To formulate a optimization problem for the drilling activities, it is required to have accurate predictive models. Such models have the goal to assess how some important variables (e.g. drill bits, rotating speed, weight on bit) affect the drilling performance. Possible performance indicators are the rate of penetration and the specific energy, both called here as *drilling-performance indicators*. These predictive models has the goal to map how the decision space (containing the variables to be optimizes) influence on the objective space (e.g. rate of penetration).

The prediction of ROP as function of drilling variables paves the way to formulate the optimization problem as maximization of ROP, or minimization of total time or cost per feet drilled. The accuracy of ROP model is crucial for drilling optimization (SOARES; GRAY, 2018). However, understanding how the drilling variables really affect the ROP is an open question in drilling engineering (MITCHELL; MISKA, 2011). Despite many efforts (theoretical and experimental), modeling the ROP as a mathematical function (i.e. traditional models) of some variables is not so trivial, because this is highly non-linear problem. Soares, Daigle and Gray (2016) exposed limitations and deficiencies of traditional ROP modeling, based on analytical equations. Therefore, machine learning techniques (e.g. neural networks, support vector machine, random forests) appears to be a more reasonable approach to model the ROP (HEGDE et al., 2017).

Another important performance indicator is the specific energy, first proposed by Teale (1965). The concept of specific energy measures the amount of energy employed to destroy a unit volume of rock. As consequence, the drilling efficiency can be assessed from the energy-efficiency point of view. Many researches highlighted the use of specific energy in drilling-optimization tasks (PESSIER; FEAR, 1992; AMADI; IYALLA, 2012; BEVILACQUA; CIARAPICA; MARCHETTI, 2013; MENG et al., 2014; CHEN et al., 2016).

The following works developed predictive models of drilling-performance indicators for optimization purposes (ARABJAMALOEI; SHADIZADEH, 2011; AWOTUNDE; MUTASIEM, 2014; BATAEE; IRAWAN; KAMYAB, 2014; HEGDE; GRAY, 2017; HEGDE; GRAY, 2018; HEGDE; DAIGLE; GRAY, 2018). A common approach of these works was to employ, as objective function, the maximization of ROP. Other objective functions tested were: minimization of drilling cost (BAHARI; SEYED, 2009), minimization of total time (AWOTUNDE; MUTASIEM, 2014), and minimization of torque and specific energy (HEGDE; GRAY, 2018). In general, the drilling variables to be optimized were weight on bit and rotating speed. Some of the mentioned works added also a third variable to be optimized, which were e.g. mud flow or mud weight.

All these mentioned works proposed a single-objective optimization problem. Only few works (GANDELMAN, 2012; GURIA; GOLI; PATHAK, 2014; PAYETTE et al., 2017) tried to optimize some controllable drilling-variables (e.g. weight on bit and rotating speed), considering simultaneous multiple-objectives. We consider the thesis from Gandelman (2012) as the only work that proposed a multi-objective optimization based on predictive models trained with machine learning techniques. To be more specific, the author employed artificial neural networks to model the ROP, which was coupled with specific energy equation. In his study, the goal was to find a optimum combination of operation conditions (bit weight and rotating speed), which could achieve a ROP-desired value and, at the same time, spend as less as possible energy. The author mentioned the possibility of applying this approach in real-time optimization.

In the presented thesis, some procedures in obtaining predictive models for drilling-performance indicators are investigated, as well as how these models can be used in optimization tasks. The works from Gandelman (2012), Hegde and Gray (2018) inspired the development of optimization approach adopted in the current. Hegde and Gray (2018) showed that the common approach of trying to maximize the ROP may decrease the drilling efficiency, because it may lead to increases in specific energy and torque. As a natural extension of Hegde's work is to adopt a multi-objective optimization, where both objectives can be set: maximization of ROP and minimization of specific energy, as done by Gandelman (2012). One difference between the current and from Gandelman is how the specific energy is calculated. In the current work, the specific energy is coupled not only with the predictive model of ROP, but also with torque, as proposed by Hegde and Gray (2018).

1.3 OBJECTIVES

The present thesis aims to apply machine learning methods for real-time drilling optimization. For that, two objects of study can be formulated. The first object frames the research question of how predictive models of so-called *performance indicators*, especially rate of penetration and specific energy, can be obtained based on machine learning methods. To be more specific, the first object of study consists of:

- reviewing the state-of-art of machine learning techniques applied to rate of penetration prediction;
- proposing a systematic procedure for data pre-treatment prior to training the predictive models;
- studying the influence of most important drilling parameters;
- comparing the coupled specific-energy models against a simpler version based on exclusively data-driven models, as commonly done for the rate of penetration.

The second object of study covers the investigation of different procedures of drilling optimization based on the predictive models. To be more specific, the second object of study consists of:

- comparing different objective functions of a single-objective optimization problem, especially the maximization of ROP against the minimization of specific energy;
- investigating the feasibility of a multi-objective optimization framework, especially based on ε -constraint method;

1.4 WORK STRUCTURE

After the introduction chapter, the chapter 2 presents a briefly background about drilling engineering, machine learning methods for regression and multi-objective optimization. The review of state-of-art methods to model the ROP is presented in chapter 3. The chapter 4 presents the materials and methods employed in the current work. In chapter 5, the results are presented, followed by the conclusion chapter 6. Appendixes and Annexes were added to support the development of the current work.

2 THEORETICAL BACKGROUND

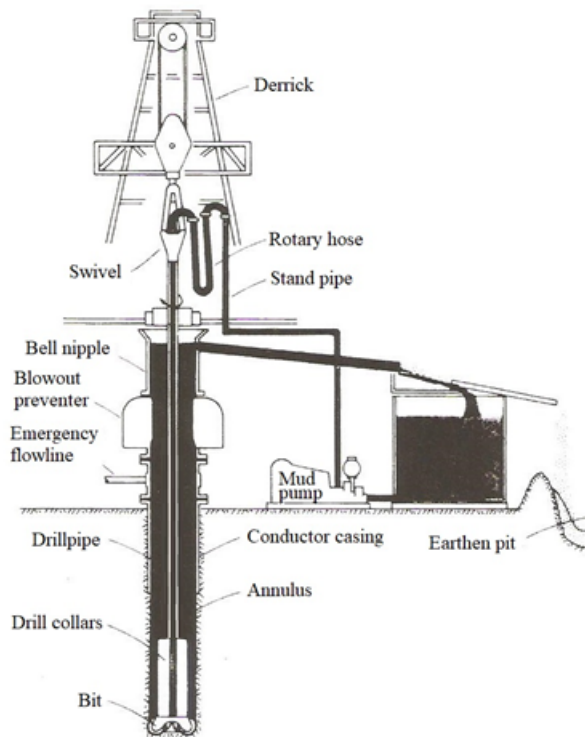
This chapter aims to briefly present the basic concepts which are necessary for the development of this work. This chapter have three main parts. The first one deals with fundamentals of drilling engineering. The second part presents some machine learning methods for regression problems. The third part gives an overview of multi-objective optimization.

2.1 OVERVIEW OF DRILLING ENGINEERING

2.1.1 Rotary Drilling

The standard oil well drilling method is rotary-drilling. The main components of rotary-drilling are shown in Figure 2.1.

Figure 2.1 – The rotary drilling process



Source: Mitchell and Miska (2011)

Mitchell and Miska (2011) provided a gentle introduction to rotary-drilling, which is summarized as following: a drill bit uses the downward force and rotation speed to break the rock into small pieces (*cuttings*). The downforce (commonly known as weight on bit) is generated by the weight of pipes (*drill collars*) above the drilling bit, and the rotation is provided normally by a surface equipment that rotates the *drillstring*, transmitting rotation to the bit. A fluid (*drilling fluid* or commonly known as *mud*) is pumped into the hollow *drillstring* to the bottom of the

hole. Then, the fluid passes through orifices of drillbit, and returns to surface transporting the *cuttings* through the annular space. At surface, the *drilling fluid* is treated, separating the *cuttings* from the fluid. Finally, the treated fluid is reused.

2.1.2 Drilling Variables

The drilling variables can be classified as alterable or unalterable as shown in Figure 2.2. This classification is not strict, because there is an interdependence among them. From experience and research, six groups of variables can be used to mathematical optimization (LUMMUS, 1970). With recent advances in directional and horizontal wells, the wellbore trajectory is an important aspect to be considered in modern drilling operations. Therefore, this parameter should be included in such lists nowadays. In the following paragraphs, the most common types of drillbits are explained briefly, as well as some properties of drilling fluids.

Figure 2.2 – Classification of drilling variables. The symbol ✓ indicates that those variables are chosen for mathematical optimization.

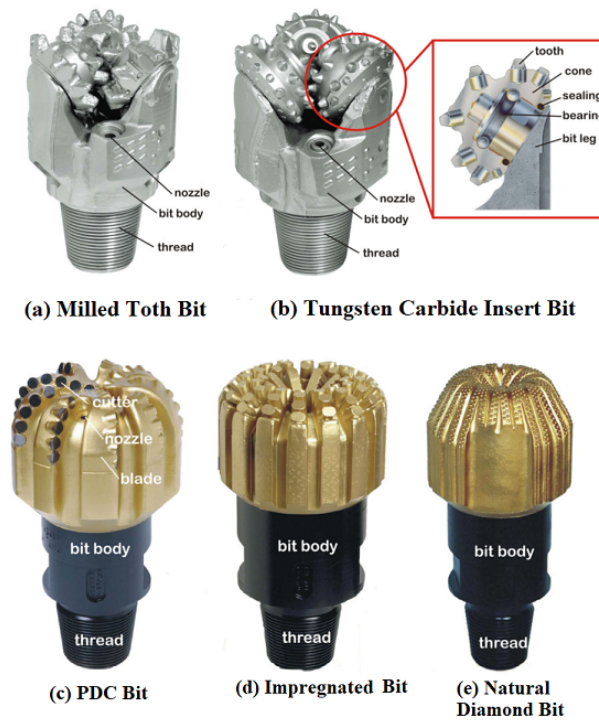
Alterable	Unalterable
✓ Mud - Type - Solids Contents - Viscosity - Fluid Loss - Density ✓ Hydraulics - Pump Pressure - Jet Velocity - Circulating Rate - Annular Velocity ✓ Bit type ✓ Weight on bit ✓ Rotary speed	• Weather • Location • Rig Conditions • Rig Flexibility • Corrosive borehole gases • Bottom-hole temperature • Round-trip time • Rock properties • Characteristic hole problems • Crew Efficiency ✓ Formation to be drilled ✓ Depth

Source: adapted from Lummus (1970)

The rotary-drilling method employs a drillbit to produce a general cylindrical hole, known as wellbore. The cutting elements (tooth or cutter) are responsible to break apart the rock. According to the cutting mechanism, the drill bits are classified into the following groups: the roller cone bits, the fixed cutter bits, and the hybrid bits (MA; CHEN; ZHAO, 2016).

The roller cone bits have normally three cones, but it is possible to see other arrangements (single, two or four cones). The cones roll as the bit rotates, yielding to a rock-failure mode of fracturing or crushing. Each cone has teeth, which are the cutting elements. The teeth can be milled directly in the cone steel (Figure 2.3-a). To improve the durability of drill bits, teeth of sintered tungsten carbide can be employed. This type of roller cone bits are known as TCI (see Figure 2.3-b) (MA; CHEN; ZHAO, 2016).

Figure 2.3 – Typical drill bits: a-b) roller cone bits; c-e) fixed cutter bits



Source: adapted from Ma, Chen and Zhao (2016)

The cutter mechanisms can be much simpler, when employing the fixed cutter elements. Depending on the manufacture, the fixed cutter bits can be classified into four classes (MA; CHEN; ZHAO, 2016):

- the drag bits were the first type used in rotary-drilling. However, this type of bit can drill only soft formations, being not often employed nowadays in the oil industry;
- the polycrystalline diamond cutter (PDC) bits are widely employed. The cutting elements consist of sintered tungsten carbide cylinder with one flat surface, which is coated with a synthetic diamond material. This type of drillbits are employed in pre-salt regions (NASCIMENTO, 2016);
- impregnated bit employs the thermal stable polycrystalline (TSP) diamond cutters.
- natural diamond bit is very similiar to impregnated bits. The diamond bits employ natural industrial-grade diamond bit in the bit body matrix. It can withstand hard abrasive formations.

The third group of drill bits is a hybrid type, which combines both rolling cutters and fixed cutter elements. The goal of this type of bit is to reduce drilling time in complex applications (e.g. hard abrasive formations) (MA; CHEN; ZHAO, 2016), also used in pre-salt regions (NASCIMENTO, 2016).

The main goals of drilling fluids are: to cool the drill bit, to maintain stable the wellbore (avoiding collapse), to clean the hole, to transport drill cuttings. Some variables are specific to drilling engineering field, which requires a brief explanation.

The mud weight (MW) is a synonym of the drilling fluid density, normally reported in *lbm/gal* (also known as pound-per-gallon *ppg*), kg/m^3 or g/cm^3 . The density controls the hydrostatic pressure (SCHLUMBERGER, 2018a).

Zamora and Roy (2000) developed a new concept of equivalent density for high pressure and high temperature (HP/HT) wells, where drilling fluids undergo changes in density. The authors have proposed the term *equivalent static density* (ESD) for static wells, and *equivalent circulating density* (ECD) for circulating wells. The ECD plays a important role in avoiding kicks and losses (SCHLUMBERGER, 2018b).

2.1.3 Performance Indicators

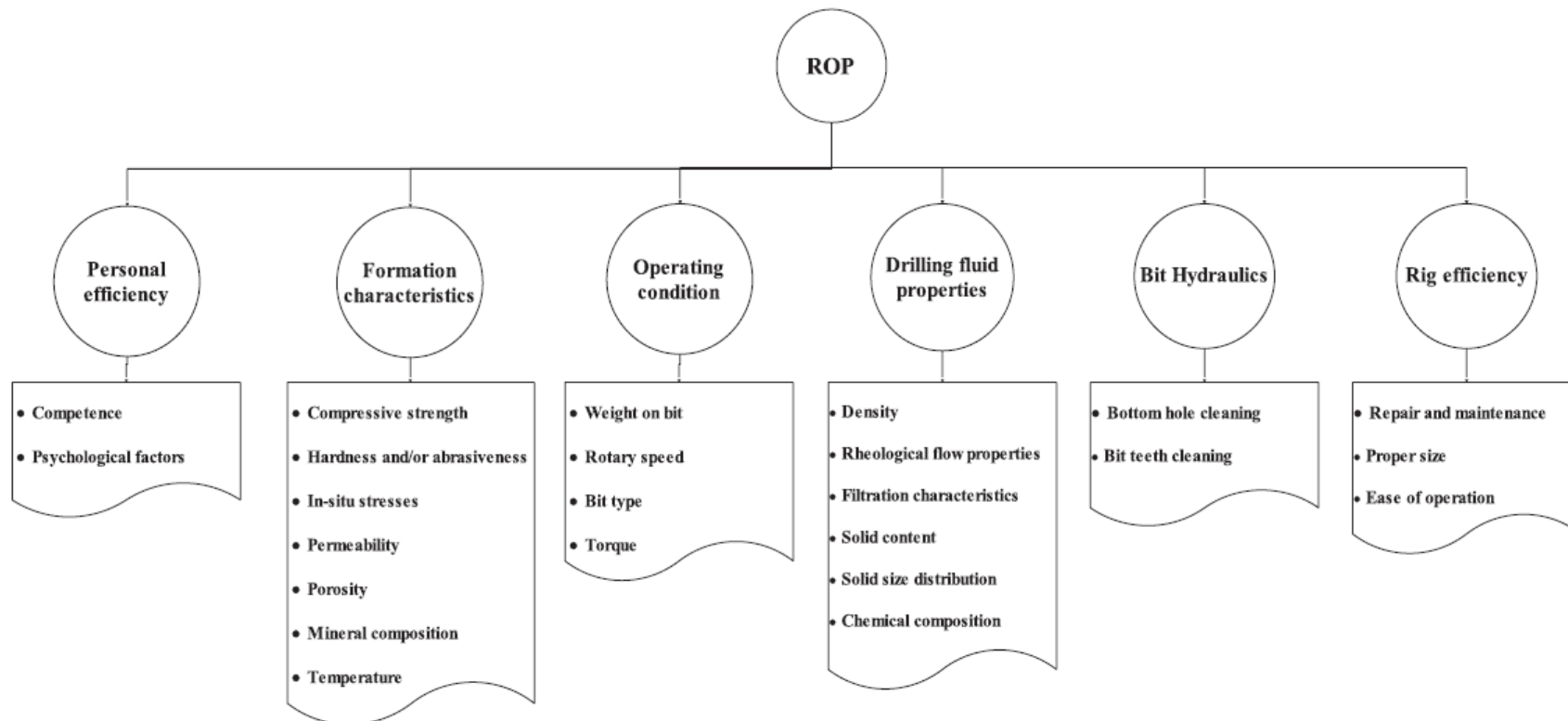
In the current work, the so-called performance indicators are those metrics that can be employed as objective function in the optimization problem of the controllable drilling parameters optimization (such as bit weight, rotary speed and mud flow). A special attention is given to two performance indicators: rate of penetration and specific energy.

2.1.3.1 Rate of Penetration

The rate of penetration (ROP) is an important drilling performance indicator. According to Mitchell and Miska (2011, p. 352), the drilling variables that affect most the ROP are: bit type, formation characteristics, drilling-fluid properties, bit operating conditions (bit weight and rotary speed), bit tooth wear, and bit hydraulics. In addition to these variables, Anemangely et al. (2018) added the personal efficiency and rig efficiency as important parameters that affect the ROP, as seen in Figure 2.4.

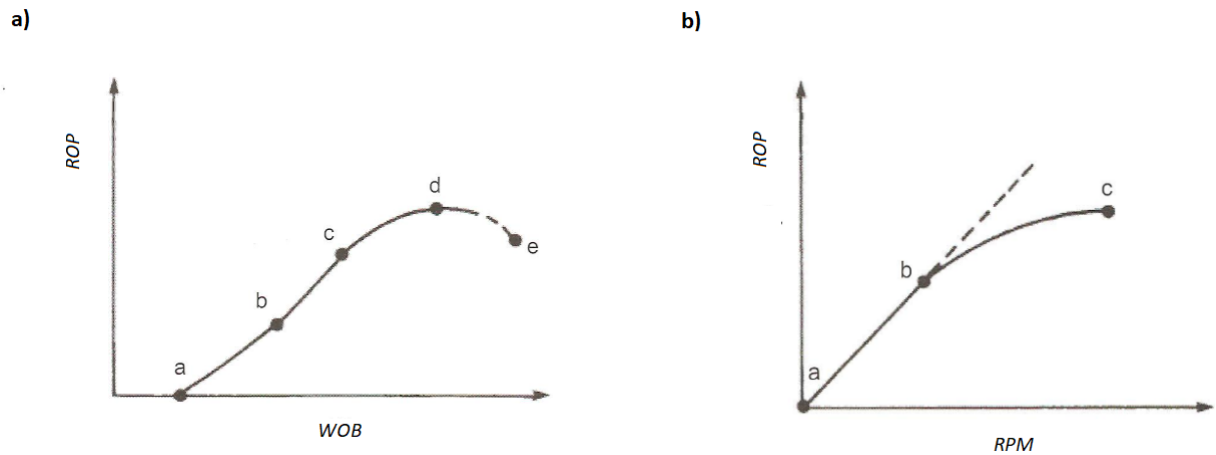
The manner in which the drilling variables affects the ROP is very complex, and not well understood (MITCHELL; MISKA, 2011). However, it is of interesting to model the ROP as function of the drilling variables, because it enables to optimize drilling process by finding the optimum controllable drilling variables. The Chapter 3 covers different methods to model the ROP as function of the most important drilling variables.

Figure 2.4 – Drilling variables that affect the ROP.



Source: Anemangely et al. (2018)

Figure 2.5 – Drilling mechanical parameters influences' on penetration rate (ROP)



Source: adapted from Mitchell and Miska (2011)

Numerous studies analyzed the effect of bit weight and rotary speed on penetration rate (DUPRIEST; KOEDERITZ, 2005; EREN; OZBAYOGLU, 2010; MITCHELL; MISKA, 2011; GANDELMAN, 2012; SOARES; DAIGLE; GRAY, 2016). The describing of the bit mechanics provided in this section is a summary from the book of Mitchell and Miska (2011).

In Figure 2.5-a, it is depicted a plot of penetration rate vs. bit weight obtained experimentally, when all other variables are held constant. Significant rate of penetration is obtained after exceeding the threshold formation stress, increasing linearly with weight-on-bit (WOB) for low-to-moderate values of bit weight (segment ab from Figure 2.5). A linear curve with steeper slope is observed at higher values of WOB (segment bc). This occurs due to a change of rock-mode failure from scraping or grinding to shearing, increasing the drilling efficiency. However, after the point C, slight improvements in ROP is obtained with increasing WOB (segment cd). It can be observed in some cases a decrease in ROP at even higher values of WOB (segment de). This behaviour is called *bit foundering*. Two facts can explain this poor response of ROP at higher WOB. One fact is a higher rate of cuttings generation, decreasing the efficiency of hole cleaning. Another reason is the lack of room for fluid bypass (MITCHELL; MISKA, 2011).

A typical plot of ROP against rotary speed with all drilling parameters held constant is illustrated in Figure 2.5-b. The ROP increases linearly with increasing rotary speed (segment ab). At higher values of rotary speed, issues of bottomhole cleaning occur (segment bc), reducing the drilling efficiency (MITCHELL; MISKA, 2011). For more details about the bit mechanisms, the readers can refer especially to the other works from Dupriest and Koederitz (2005), Gandelman (2012).

2.1.3.2 Specific Energy

The concept of specific energy measures the energy efficiency of drilling process. Teale (1965) originally defined the specific energy (known as MSE or SE)¹ as the amount of energy required to destroy a unit of volume of rock for rotary drilling. The author derived the specific energy equation, SE , by dividing the amount of axial and torsional work by the volume of rock drilled. In Teale's formulation, the work is done by thrust (i.e. weight on bit WOB in $[lbf]$) and torque, T in $[in.lbf \text{ or } ft.lbf]$. The SE is expressed by:

$$SE = \frac{WOB}{A_b} + \frac{120.\pi.N.T}{A_b.ROP} \quad [in.lbf/in^3] \quad (2.1)$$

where N denotes the rotation speed $[rev/min]$, A_b is the bit area $[in^2]$, and ROP the penetration rate $[in/hr \text{ or } ft/hr]$, according to unit of torque T . The unit of specific energy is $[in.lbf/in^3]$, i.e. the same unit of pressure $[psi]$. In the International System of Units, the SE is commonly measured in $[MJ/m^3]$, or in terms of pressure $[MPa]$.

The equation 2.1 requires torque on the bit. However, this parameters is not always available, because it requires measurement-while-drilling system (MITCHELL; MISKA, 2011). For this reason, Pessier and Fear (1992) introduced the sliding friction μ to calculate the torque as function of weight on bit, as following:

$$T = \mu \left(\frac{d_b.WOB}{36} \right) \quad (2.2)$$

where d_b is the drill bit diameter. Substituting the equation 2.1 by equation 2.2 yields:

$$SE = WOB \left(\frac{1}{A_b} + \frac{13.33\mu N}{d_b.ROP} \right) \quad (2.3)$$

In order to use the model of Pessier and Fear (1992), the sliding friction μ needs to be estimated. An alternative was proposed by Rabia (1982), who presented a specific energy model without torque as input. The simplified model is:

$$SE = 2.35 \left(\frac{WOB.N}{d_b.ROP} \right) \quad [MJ/m^3] \quad (2.4)$$

when the unit of variables is: N $[rev/min]$, WOB $[kg]$, d_b $[mm]$, and ROP $[m/hr]$. Rabia (1982) presented also this equation in Imperial Units:

$$SE = 20 \left(\frac{WOB.N}{d_b.ROP} \right) \quad [lbf/in^2] \quad (2.5)$$

when N is given in $[rev/min]$, WOB in $[lbf]$, d_b in $[in]$, and ROP in $[ft/hr]$.

¹ MSE denotes mechanical specific energy, as well as mean squared error, which is a common statistic metric for assessing the predictive models. To avoid any misunderstanding, we prefer to name the concept of specific energy as SE. In additional, Teale called this concept as specific energy in his paper from 1965.

The values obtained from the previous formulations may be substantially higher than the material's strength. Dupriest and Koederitz (2005) mentioned the importance to adjust the SE to the same magnitude of material's strength. The authors adapted the original formulation from Taele in order to improve the usefulness of SE in field operations. Their model can be expressed by:

$$SE_{adj} = EFF_M \cdot SE = EFF_M \left(\frac{WOB}{A_b} + \frac{120\pi \cdot N \cdot T}{A_b \cdot ROP} \right) \quad (2.6)$$

where SE_{adj} is the adjusted specific energy, the EFF_M is the mechanical efficiency factor. Dupriest and Koederitz (2005) mentioned the value of EFF_M can be from 0.3 to 0.4.

All the previous models consider that the work is done by axial force and torque applied to the drill bit. Therefore, they are also known as mechanical specific equations. The hydraulic power is not taken into account despite the importance of drilling fluid for process. Mohan, Adil and Samuel (2014) extended the original Taele's formulation by adding the hydraulic work done by the bit. The authors named this new concept as hydromechanical specific energy (HMSE):

$$HMSE = \frac{WOB_e \cdot ROP + 120\pi \cdot N \cdot T + \eta \cdot \Delta P_b Q}{A_b \cdot ROP} \quad (2.7)$$

where all parameters are the actually values on drill bit, i.e., WOB_e is the effective weight on bit. N is the rotation speed, and T is torque on the bit. The term ΔP_b denotes the pressure drop across the bit, η is a dummy factor for energy reduction and Q is the flow rate.

Recently, some new specific mechanical energy models were developed. Chen et al. (2016) proposed a new formulation for SE in slide and horizontal drilling, adapting the original formulation for positive displacement motor. Al-Sudani (2017) presented new approach based on control engineering to simulate the bit behavior and to predict the drilling efficiency by estimating the transferred and wasted mechanical energy.

2.1.3.3 Drilling Efficiency

The amount of energy to destroy a rock is correlated to material's strength. Teale's experiment has showed that the values of SE were roughly equal to compressive strength of rock. Two common properties are used to measure the compressive strength: unconfined compressive strength (UCS) and confined compressive strength (CCS). Both measure the maximum axial compressive stress that a sample of material can withstand. When the confining stress is zero, the condition of this compression test is unconfined, therefore the obtained value is UCS. When a confining stress is applied, the rock is said to be confined, therefore the respective compressive strength is CCS.

Knowing this rock property is useful, because it provides a reference point for drilling efficiency. When drilling at maximum efficiency, it is expected all energy applied to bit is transferred to destroy the rock. In this case, the specific energy is equal to confined compressive strength. When drilling dysfunction occurs, part of energy is lost. Dupriest and Koederitz (2005)

showed the monitoring SE in real-time boosted the drilling efficiency of several wells drilled, because drilling inefficiencies could be detected in real-time. Amadi and Iyalla (2012) achieved cost reduction in deepwater development by applying mechanical specific energy optimization techniques in real-time.

A study on drilling parameters optimization from Chen et al. (2016) proposed that the minimum specific energy is equal to confined compressive strength (CCS):

$$\min(SE) = CCS \quad (2.8)$$

In this case, the drilling efficiency, η , is at maximum. In a general case, the drilling efficiency is given by:

$$\eta = \frac{CCS}{SE} \quad (2.9)$$

Dupriest and Koederitz (2005) showed that, in real drilling operations, the η ranges from 0.3 to 0.4. For that reason, the authors proposed the mechanical efficiency factor EFF_M (equation 2.6).

In the case of hydromechanical formulation (MOHAN; ADIL; SAMUEL, 2014), the drilling efficiency is expressed by:

$$\eta = \frac{CCS}{HMSE} \quad (2.10)$$

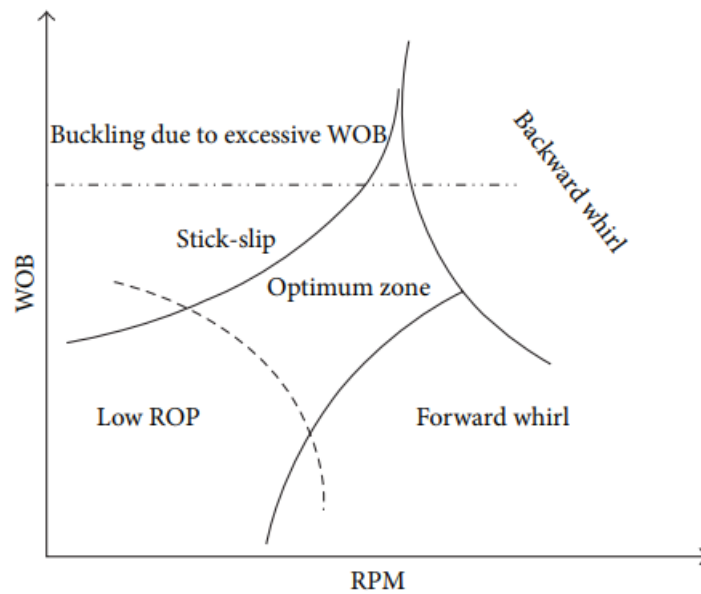
To assess the rock strength properties (UCS or CCS), uni-axial or tri-axial compressive tests can be carried. The downside of these tests is the need to have rock samples, so that any attempt to monitor the rock strength while drilling is not possible. An alternative is to estimate the rock properties from p-wave velocity with empirical equations. Since P-wave velocity is closely related to physico-mechanical properties of rock and is non-destructive test, this measurement enables to estimate the rock properties in situ conditions (e.g. while drilling) (SARKAR; VISHAL; SINGH, 2012; SHARMA; SINGH, 2008). Empirical equations for rock strength from p-wave velocity was investigated also under different degrees of saturation (KARAKUL; ULUSAY, 2013).

Hamada et al. (2018) proposed a new method to determine the rock strength based only on drilling parameters such as drill string rotational torque, bit depth, and string rotational speed. The authors called this new concept as “equivalent strength” (EST).

2.1.3.4 Some indicators of drilling problems

Vibration is a challenge for drillers. Several problems are attributed to the drillstring vibration, such as lower penetration rate, drillstring components' failure, excessive bit and stabilizer wear, bit failure and so on. Depending on type of vibration, a specific term is employed to describe the phenomena. According to Ghasemlooia, Geoff Rideout and Butt (2015, p. 151),

Figure 2.6 – Schematic description of optimum zone



Source: Zhu, Tang and Yang (2014)

“bit bouncing, stick-slip and whirling are extreme examples of coupled vibration dominated by axial, torsional and lateral motions, respectively”.

Ghasemloonia, Geoff Rideout and Butt (2015) identified several strategies for vibration isolation, which are:

- proper design of drillstring configuration (Bottom-Hole-Assembly length and stabilizer location) to stay far away from resonance state;
- passive isolators;
- active controllers;
- real-time drilling input parameters optimization based on “Measurement While Drilling” tools.

With “Measurements-while-drilling” (MWD) tools, it is possible monitor the controllable drilling variables in order to avoid or reduce vibration. Ghasemloonia, Geoff Rideout and Butt (2015) appointed this solution as one of the best way to mitigate the problem of vibration. In Figure 2.6, it is shown that certain combinations of bit weight and rotary speed can mitigate some drilling vibrations problems.

2.1.3.5 Economics

One of the main goal of drilling optimization is to minimize the drilling costs. For that, the selection of a suitable bit is of major importance. According to Mitchell and Miska (2011), criteria for bit selection are:

- cost-per-foot for a single bit run;
- run-cycle speed (RCS) which measures the effective ROP by including the effect of trip time and nonrotating time;
- break-even-analysis to determine if it is economically feasible replacing a current bit by a more expensive new bit;
- to estimate the termination of a bit run.

2.1.4 Optimization of Controllable Drilling Variables

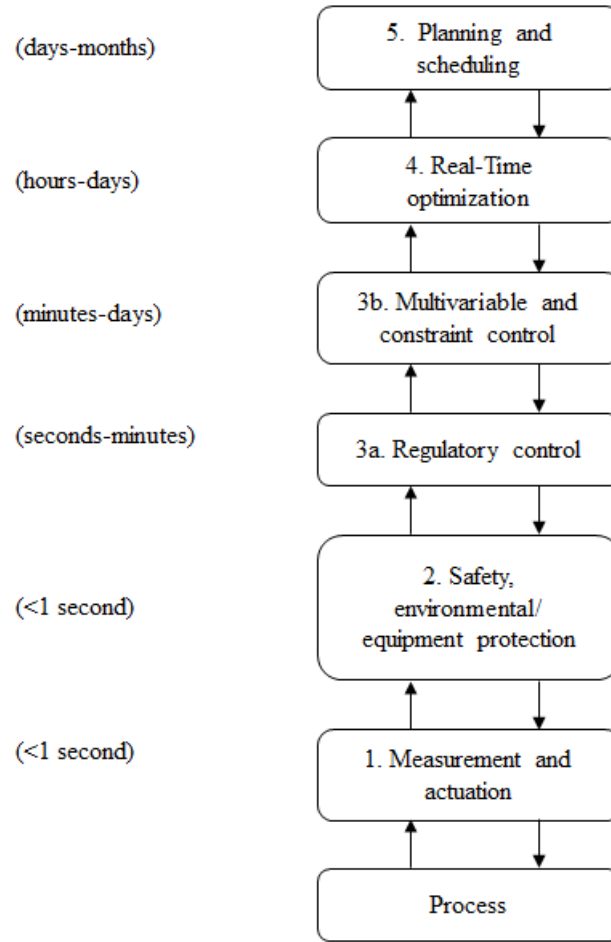
The Figure 2.7 depicts a process control hierarchy, normally applied in industrial process, with several features (optimization, control, monitoring and data acquisition). In Level 5, planning and scheduling operate normally over long time periods. In Level 4, the Real-Time Optimization provides the optimal set points (called supervisory control). Changes in set-points for multivariable process control are made in Level 3b. One technique to implement it is the model predictive control. The regulatory control acts in Level 3a for loop control. The Level 2 covers safety environmental/equipment protection, such as alarm and shutdown systems. In the Level 1, data acquisition and on-line analysis occur (SEBORG et al., 2011).

Gandelman (2012) used the definition from Seborg et al. (2011) for real-time optimization to identify what would be real-time optimization for drilling activities. Here we extended Gandelman's comparison for other common optimization approaches employed in oil and gas industry, which are: historical drilling data; drilling tests; and (near) real-time approaches; closed-loop control.

By analogy with the five levels of process control from Seborg et al. (2011), the Level 5 covers the historical drilling data analysis. This approach enables to capture trends between the variables and drilling performance indicators. This knowledge obtained from previous wellbore construction can be employed in planning and scheduling phase for new nearby wells to be drilled, or to comparable formations. This optimization process is commonly employed by drillers (LYONS; PLISGA, 2004). The cumulative experience gained by rig crew and drilling companies in a specific region yields a better understanding of this region. As consequence, best practices are employed, for example to avoid problems which may result in large non-productive time. The report from Vesconte, Tinkhof and Hardman (2014) is an example of such optimization procedure.

In industrial process, the optimum set point can vary very often within time range from hours to days, requiring real-time optimization (SEBORG et al., 2011). Gandelman (2012) mentioned that the same occurs in drilling activities of oil and gas wells. In this case, real-time optimization aims to determine the optimum drilling variables while drilling. This approach enables to adapt in (near) real-time to the changes of down-hole conditions. According to Gandelman (2012), the following issues require actions to set the drilling process again to its

Figure 2.7 – Five levels of process control and optimization (time scales for each level)



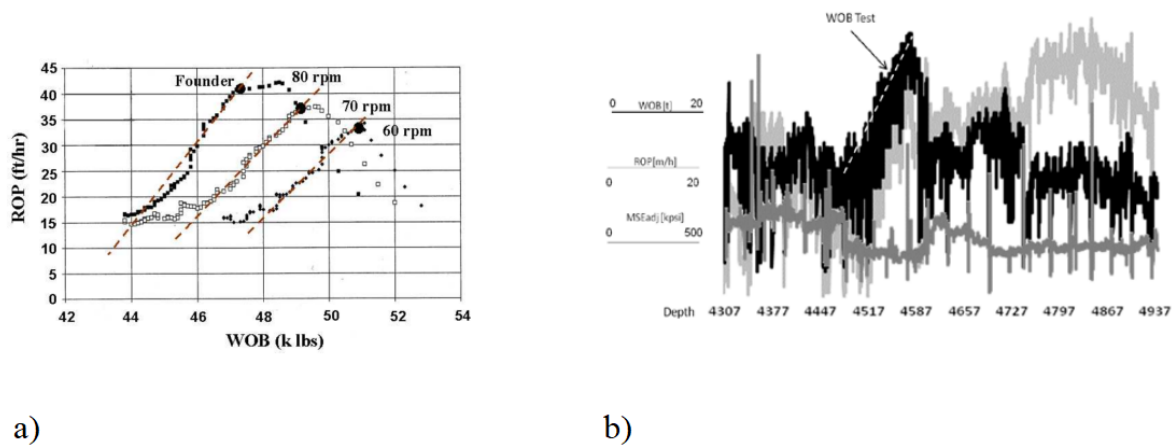
Source: Seborg et al. (2011).

optimum point: lithology changes, hole cleaning problems, weather changes (especially for offshore operations), operating problems with pumps and topdrive.

One way to implement real-time drilling optimization is through pre-operational tests. In this case, the ROP response is monitored when varying mainly the bit weight and rotary drilling, while other parameters are held constant. The so-called *drilloff* test enables to carry out such analysis in a fast way, by applying a large weight to the bit and then decreasing the bit weight, while the rotary speed is held constant (DUPRIEST; KOEDERITZ, 2005). The value of WOB which maximizes the ROP is selected as optimum value (MITCHELL; MISKA, 2011). A dynamic approach for pre-operation tests was later proposed (NASCIMENTO et al., 2016; NASCIMENTO et al., 2015a), and further developed (DUTRA, 2016). Some examples found in the literature of drilling tests are shown in Figure 2.8.

In order to implement control techniques in closed-loop, sophisticated models for drilling performance indicators are required. One key aspect for the control technique is a reliable and accurate ROP modeling (SOARES; GRAY, 2018). In Chapter 3, it is shown the machine learning techniques have the ability to do so.

Figure 2.8 – a) Typical drill-rate test data showing non-linear response below the minimum depth of cut and above the founder point; b) Bit weight (WOB) test carried during a run



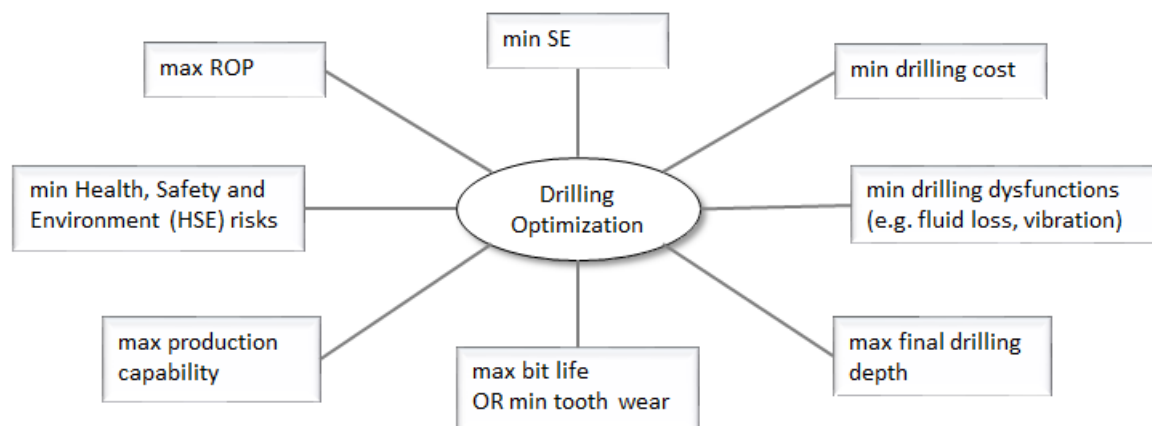
Source: a) Dupriest and Koederitz (2005); b) Bevilacqua, Ciarapica and Marchetti (2013)

Some studies reported control techniques applied for the following drilling activities: for the problem of vibration and shock (DONG; CHEN, 2016), and managed pressure drilling (GODHAVN et al., 2011). To further develop the automation in drilling, the Society of Petroleum Engineers (SPE) is organizing an international student competition as part of Drilling Systems Automation Technical Section (DSATS), called Drillingbotics². Student teams must design and build a small drill rig, which has to drill autonomously a rock sample by applying control engineering techniques. Bilgesu et al. (2017) presented the winner solution from the 2016 competition, where real-time interactive drill-off tests were combined with artificial intelligence.

Several drilling performance indicators can be employed as objective functions. The Figure 2.9 illustrates eight possible objective functions for drilling optimization. These functions can be employed solely, or in combination. These functions were obtained mainly from the following works (HEGDE; GRAY, 2018; HEGDE; GRAY, 2017; GURIA; GOLI; PATHAK, 2014; GANDELMAN, 2012; MITCHELL; MISKA, 2011; MITCHELL; MISKA, 2007; DUPRIEST; KOEDERITZ, 2005; LYONS; PLISGA, 2004; LUMMUS, 1970).

² Homepage on: <<https://drillingbotics.com/>>

Figure 2.9 – Some possible objective functions for drilling optimization



Source: own authorship.

2.2 REGRESSION MODELS BASED ON MACHINE LEARNING METHODS

One field of Artificial Intelligence that is widely known as machine learning consists basically of algorithms that are able to learn from previous examples, enabling prediction for novel inputs. The initial phase of artificial intelligence is very close to development of artificial neural networks (PONTES et al., 2010). For example, McCulloch and Pitts (1943) described mathematically the neuron. Later, Rosenblatt (1958) developed a theory, called *perceptron* for a hypothetical nervous system, or machine.

However, it was only in the early 1990s that the field began to have widespread practical impact. Over the last decade in particular, there has been a rapid increase in the number of successful applications, ranging from web search to autonomous vehicles, and from medical imaging to speech recognition. This has been driven by the increased availability of inexpensive computers (BISHOP, 2013, p. 1–2).

In order to have an idea about the amount of machine learning algorithms, Fernández-Delgado et al. (2014) compared a total of 179 classifiers from 17 different families. The authors applied some well-known methods from artificial intelligence (e.g. neural networks, support vector machines, decision trees, boosting, bagging and random forests), as well as from statistics field (e.g. generalized linear models, partial least squares, logistic and multinomial regression).

As seen, the amount of machine learning techniques is vast. This can hinder the application of these methods by those who have expertise in other fields (e.g. physicists, engineers or biological scientists). Many softwares (Matlab, Statistica, Weka, Mathematica) and open sources packages in programming languages (e.g. R or Python) have already implemented machine learning algorithms, so that they are easily accessible. However, the lack of background knowledge around the learning algorithms can lead to unsuccessful results. A work developed in this School of Engineering by Pontes et al. (2010) analyzed several publications that applied

neural networks to model machining processes surface roughness. The authors showed that those studies were carried out mainly by engineers, and some of them did not follow the good practices recommended by acknowledged scholars in neurocomputing and statistics. It is not hard to imagine that the same may happen in other areas. Therefore it is important to understand the basics of the machine learning methods.

The basics of regression models is presented based on the works from Hastie, Tibshirani and Friedman (2009), Friedman (2006). The machine learning paradigm is presented from the perspective of ensemble models; the review articles from Sagi and Rokach (2018), Ren, Zhang and Suganthan (2016) are good source of information about ensemble models.

2.2.1 Regression Analysis

Let y denote “output” or “response” variable, and $\mathbf{x}$ the “input” or “predictor” variables. The goal of a predictive or machine learning problem is to predict (estimate) $\hat{y}$ given an input vector $\mathbf{x}$. If the output is a continuous variable, i.e. $y \in \mathbb{R}$, the prediction task is a regression problem. If y is a set of classes (labels or categorical values), then estimating the output is a classification problem. In the current work, we deal with regression problem when trying to estimate the rate of penetration, specific energy or any other continuous drilling variable.

The goal of regression is to map a point $\mathbf{x}$ in the space of all input variables (called also as feature space) to a point $\hat{y}$ in the space of response values (FRIEDMAN, 2006). The predicted value $\hat{y}$ can be expressed by a general function

$$\hat{y} = f(\mathbf{x}, \Theta), \quad \hat{y} \in \mathbb{R} \quad (2.11)$$

where $f(\cdot)$ is the regression function and Θ is the regression function’s parameter set (REN; ZHANG; SUGANTHAN, 2016).

The main goal is to produce a prediction model $f(\cdot)$ with high accuracy. For that, many methods can be applied to obtain a good predicting function, such as neural networks, decision trees, kernels methods, linear/logistic regression and so on (FRIEDMAN, 2006). In machine learning, the use of previous examples from a specific problem enables to teach a predictive model how to map the input space $\mathcal{X}$ to output space $\mathcal{Y}$, i.e. $f : \mathcal{X} \rightarrow \mathcal{Y}$. In the learning phase, the parameters set of regression function Θ is obtained. After learning, the model can predict the output for new observations. The data base employed in learning phase is called as “training” data set

$$\mathcal{D}_{train} = \{y_i, \mathbf{x}_i\}_1^N \quad (2.12)$$

and has N observations where the output y has been jointly measured with the inputs variables $\mathbf{x}$ (FRIEDMAN, 2006).

In order to know whether the model $f(\cdot)$ is good or not, evaluation metrics are employed to assess the predictive model accuracy or its lack of accuracy. According to Friedman (2006,

p. 176), “the ‘loss’ criterion reflects the cost of mistakes: $L(y, \hat{y})$ is the loss or cost of predicting a value $\hat{y}$ for the response when its true value is the output”. The goal of learning is to define the set of parameters of the predicting function Θ by minimizing the loss functions (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). For regression-type problems, the absolute error $L(y, \hat{y}) = |y - \hat{y}|$ is a common cost function, as well as the squared-error $L(y, \hat{y}) = (y - \hat{y})^2$, which yields to much simpler algorithms of minimization (FRIEDMAN, 2006).

2.2.2 Ensemble Learning

Ensemble is a general term for methods that combine multiple predictors (i.e. multiple learners). The idea behind of ensemble methods is that the error of a single model can be compensated by other models. As consequence, the overall prediction performance of the ensemble is better than the estimate of a single model. That is why ensemble models are considered state-of-art techniques for machine learning challenges (SAGI; ROKACH, 2018). For example, Chen and Guestrin (2016) reported that their package (XGBoost) was employed in many winning solutions of machine learning challenges hosted at the homepage Kaggle in 2015.

The theory of ensemble methods is bias-variance-covariance decomposition. A key element of ensembles is the diversity in several forms: data diversity, parameter diversity and structural diversity (REN; ZHANG; SUGANTHAN, 2016).

Using the notation of Sagi and Rokach (2018), an ensemble learning model uses an aggregation function $G(\cdot)$ that aggregates K inducers (or base learners), $\{f_1, f_2, \dots, f_k\}$, towards predicting a single output

$$\hat{y} = G(f_1, f_2, \dots, f_k) \quad (2.13)$$

Building an ensemble model consists of selecting a training method for the base learners and choosing an aggregation process (SAGI; ROKACH, 2018).

The ensemble methods can be classified by how the base learners are obtained: in a sequential approach or in parallel. Sequential ensemble methods generate sequentially the base learners (e.g. gradient boosting). In this case, the learning process of k -th base learner f_k depends on the performance of previous base learner f_{k-1} . In parallel ensemble methods (e.g. bagging), the base learners are trained independently from each other. Therefore, this approach yields parallel ensemble models (REN; ZHANG; SUGANTHAN, 2016).

When an ensemble has all learners of the same type, this ensemble is said to be homogeneous. When different learning techniques (e.g. neural networks and decision trees) are employed to obtain the base learners, it yields to heterogeneous ensembles (REN; ZHANG; SUGANTHAN, 2016; MENDES-MOREIRA et al., 2012).

The following section presents a classical method to obtain the base learners for an ensemble models, namely the classification and regression tree. Then, we present two classical

approaches (bagging and boosting) to aggregate the base learners in an ensemble are presented, showing briefly some recent advances in ensemble studies.

2.2.3 Base Learners

The decision tree “are conceptually simple yet powerful” (HASTIE; TIBSHIRANI; FRIEDMAN, 2009, p. 305). Trees are at the heart of some classical ensemble methods, such as random forest and gradient boosting machines. Therefore, an overview about tree-based models is given, because they are the base learners of aforementioned ensemble methods.

The regression models based on decision trees have mainly of two steps in the learning phase (HASTIE; TIBSHIRANI; FRIEDMAN, 2009): partition of input space into a set of rectangles; in the second step, a suitable model is fitted in each region (e.g. a constant value). Extensive details about the tree-based models are given by Hastie, Tibshirani and Friedman (2009), whose work serves the basis to present a classical method to grow a tree, known as Classification And Regression Trees (CART). The Figure 2.10 illustrates the process of recursive binary partition adopted by CART. A fake data was used to illustration the idea of decision tree based on CART. A general partition can have some complex regions to be described. The feature space partition can be simplified by the recursive binary splitting. The variable and split-point are determined when the best fit from all possible combinations is achieved. The corresponding prediction surface and decision tree are presented. Each region has a constant value. The terminal nodes in the tree represents the leafs.

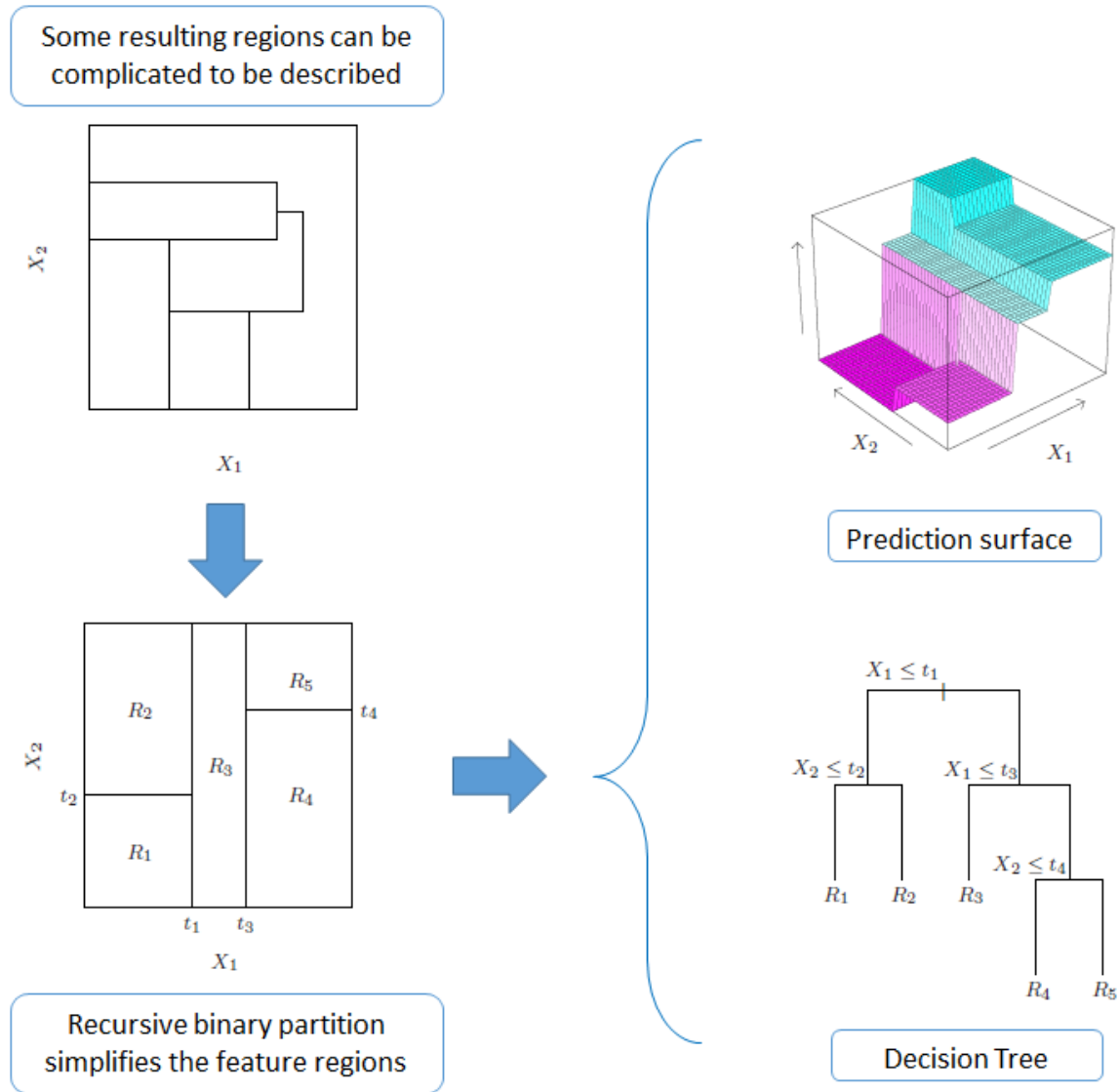
Now, let us focus on how a tree can be mathematically formulated, following the comprehensive introduction and formulation given by Hastie, Tibshirani and Friedman (2009). The authors mentioned that the training algorithm has to automatically determine how the tree should grow by determining the splitting variables and split values, and also tree’s topology (shape). Let $\mathbf{x}_i$ denote an i th vector input, such as $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ with p inputs. If the feature space is partitioned into M regions $R_1, R_2, \dots, R_M$, and the response is modeled as a constant c_m in each region. According to the authors, this tree can expressed by:

$$f(\mathbf{x}) = \sum_{m=1}^M c_m I(\mathbf{x} \in R_m) \quad (2.14)$$

here $I(\cdot)$ is the indicator function, having the value 1 if the argument ($\mathbf{x} \in \mathbb{R}$) is true or 0 otherwise. The best $\hat{c}_m$ is the average of y_i in the region R_m , if the criterion of minimization of the sum of squares $\sum (y_i - f(\mathbf{x}_i))^2$ is adopted (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

As finding the best binary partition by minimizing the sum of squares is generally computationally unfeasible, a greedy algorithm is proceeded (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). For that, the binary partition occurs by obtaining a pair of half-planes as follows (HASTIE;

Figure 2.10 – Partitions and CART



Source: adapted from Hastie, Tibshirani and Friedman (2009).

TIBSHIRANI; FRIEDMAN, 2009):

$$R_1(j, s) = \{X | X_j \leq s\} \quad \text{and} \quad R_2(j, s) = \{X | X_j > s\} \quad (2.15)$$

where j denotes a splitting variable j and s a split point. The authors explained that the goal is to determine a splitting variable j and split point s that solve

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right] \quad (2.16)$$

For any j and s , the inner minimization is solved by (HASTIE; TIBSHIRANI; FRIED-

MAN, 2009):

$$\hat{c}_1 = \text{ave}(y_i | x_i \in R_1(j, s)) \quad \text{and} \quad \hat{c}_2 = \text{ave}(y_i | x_i \in R_2(j, s)) \quad (2.17)$$

The determination of the best pair (j, s) is feasible by scanning through all of the inputs, because the determination of the split point s can be done quickly for each splitting variable. The splitting process can be repeated on all resulting regions (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Hastie, Tibshirani and Friedman (2009) discussed how large a tree should be grown. If the resulting tree is too large, it might overfit the training data, so that the obtained tree may not predict accurately new inputs. However, if the resulting tree is too small, important structures may not be captured. One approach to tune the tree size is to test iteratively several tree sizes, and then check which one best fits the data. Another approach is to grow a large tree and then prune it. Hastie, Tibshirani and Friedman (2009) detailed a method of pruning based on cost-complexity.

Other methods of training tree-based models are detailed in a review from Kotsiantis (2013). Although it is not the intention of the current work to provide deep details about tree-models, it is interesting to mention some features of these models, reported by Breiman (1996), Hastie, Tibshirani and Friedman (2009).

- the instability of trees is one major concern resulted from their high variance. This high variance occurs due to hierarchical nature of trees' growths. Even small changes in the data may result in a very different tree (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). However, this same instability improves the accuracy of ensembles models obtained by a method called "bagging", as reported by Breiman (1996);
- tree-based models can handle missing values. One approach is to create a new category for "missing". Another approach is more general. It is based on the construction of surrogate variables, by trying to evaluate correlations between missing predictor and other observed predictors (HASTIE; TIBSHIRANI; FRIEDMAN, 2009);

2.2.4 Parallel Learning with Bagging and Random Forests

Bagging is acronym for "**boosting aggregating**" originally presented by Breiman (1996). The idea behind of bagging is to generate multiple predictors based on different learning sets obtained by bootstraps replicates. The bootstrapping forms replicate data-sets drawn at random with replacement from the original data set $\mathcal{D}_{train}$. This means, each $\{y_i, \mathbf{x}_i\}$ may appear repeated times or not at all in a particular learning set $\mathcal{D}_{train}^{(m)}$ obtained by bootstrapping from the original $\mathcal{D}_{train}$. After training the predictors, the aggregation averages the response over all predictors when predicting a numerical.

Breiman (1996) stated that "for unstable procedures bagging works well", improving the prediction accuracy. This is because a small change in the training data set yields to a

significant change in the prediction function $f(\cdot)$. Examples of unstable procedures are trees, neural networks, and their variance decrease with bagging strategy (SAGI; ROKACH, 2018). However, Breiman (1996) mentioned that bagging decreases the accuracy when the base learners are stable procedures (e.g. k-nearest neighbor methods).

The Figure 2.11 illustrated two conventional ensemble methods. The solid blue lines show the flow of bagging method, and the dashed red lines show the flow of boosting. In the illustration, $\mathbf{X}$ is the original dataset, $\mathbf{X}^{(i)}, i \in \{1, 2, \dots, M\}$ are the generated data set, $f^{(i)}$ are the base predictors and f_{en} is the aggregation function.

Breiman (2001a) exploited even more the randomness in developing the random forests. The author combined the bagging procedure with random selection of splitting variable in the growth process of the tree. This method is known as random forests, and is “probably the most popular ensemble method developed” (SAGI; ROKACH, 2018, p. 9). Random forests are obtained by following the three main steps (SAGI; ROKACH, 2018; BREIMAN, 2001a):

- each new training set is drawn randomly from the original training set with replacement (bootstrap samples);
- a tree is grown on the new training set using random feature selection, and is not pruned. The CART methodology without pruning can be employed or any other training algorithm;
- in regression problems, the random forest predictor is formed by taking average over all trees.

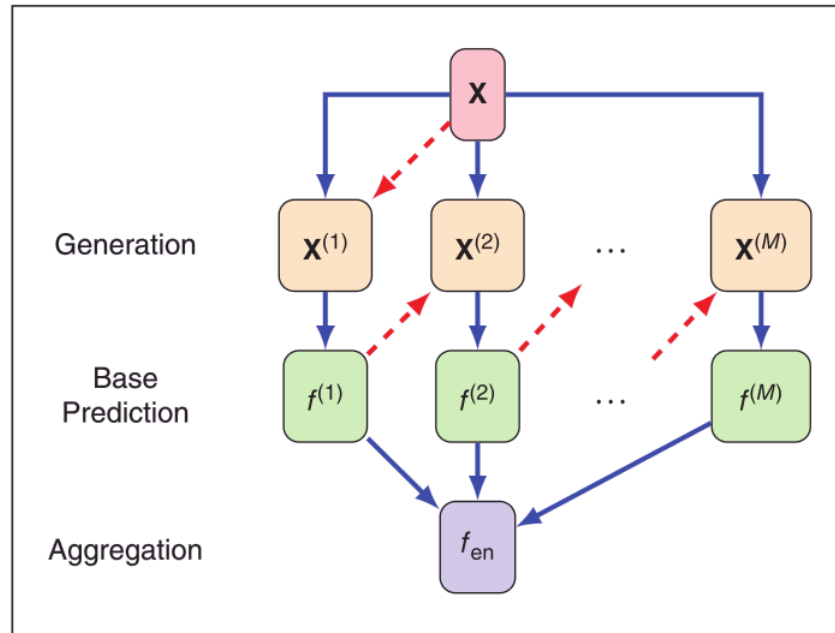
The bagging process enables to estimate the generalization capability of an ensemble without a testing dataset. About one-third of the observations is left out in each bootstrap training set (BREIMAN, 2001a). Estimating the generalization error of an ensemble based on out-of-bag observations provides a good estimation of the model accuracy. This approach provides an alternative to cross-validation process in assessing the model accuracy. Wolpert and Macready (1999) developed methods to estimate the generalization error for bagging predictors.

2.2.5 Sequential Learning with Boosting

Boosting methods can convert the prediction of many “weak learners” (slightly better than random guessing) into strong learners (SAGI; ROKACH, 2018; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). The process of boosting is more complex than bagging. One classical method is the AdaBoost algorithm (FREUND; SCHAPIRE, 1997), which stands for Adaptive Boosting, and is a well-known method for classification problems. The idea behind of AdaBoost is to fit a sequence of learners to weighted version of the training data. In the fitting process of k th weak learner $f^{(k)}$, more weight is given to observations miss-classified by the previous weak learner $f^{(k-1)}$. In this case, the k th weak learner focus on the deficits from the previous learner.

Friedman (2001) presented Gradient Boosting Machines as an optimization problem for the boosting method. In these machines, the new models are consecutively fitted, so that the

Figure 2.11 – Framework of conventional ensemble methods. The solid blue lines show the parallel flow of bagging, random subspace and Random Forest. The dashed red lines in generation and base prediction parts denote boosting ensemble framework.



Source: adapted from Ren, Zhang and Suganthan (2016).

response accuracy of the ensemble improves. “The principle idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble” (NATEKIN; KNOLL, 2013).

Chen and Guestrin (2016) presented their package for machine learning problems called XGBoost, which is a scalable tree boosting system. One of the most important factors behind the success of this method is its scalability to problems with billions of examples. This was achieved due modifications in tree growing process and the introduction of a novel sparsity-aware algorithm for parallel tree learning.

2.2.6 Other Ensemble Methods

Although it is not the main objective of this current work to develop new machine learning algorithms, it is worth of mentioning some recent advances in ensemble methods. Besides the conventional ensemble methods (bagging and boosting), there are other ensemble techniques, as reviewed by Ren, Zhang and Suganthan (2016):

- The decomposition based ensemble methods - applied to time series forecasting;
- The negative correlation learning based ensemble methods - ability to introduce strong diversity among base learners without having to change the dataset for the base learners;
- The multi-objective optimization based ensemble methods - use of state-of-art techniques of multi-objective optimization in the learning phase of an ensemble. For example, it can

be used to tune some parameters related to base learners;

- The fuzzy ensemble methods - they combine the high accuracy of ensemble methods with the ability of fuzzy logic and fuzzy sets for imperfect data management;
- The multiple kernel learning based ensemble methods - it brings together different kernels (e.g. support vector machines) into an strong predictor;
- The deep learning based ensemble methods can improve the accuracy. Deep structures have multiple layers of non-linear functions, and can learn high-level abstraction for challenging tasks (e.g. vision recognition).

2.2.7 Hyperparameter Optimization

Most machine learning algorithms have parameters to be tuned in order to provide a good generalization capability. “Unfortunately, this tuning is often a ‘black art’ that requires expert experience, unwritten rules of thumb, or sometimes brute-force search” (SNOEK; LAROCHELLE; ADAMS, 2012). For example, when training a random forest, some hyperparameters to be tuned are: the amount of trees, the amount of variables to be selected at random for each decision split, the deep of tree base learners, the method to grow the tree base learners (BREIMAN, 2001a; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). For gradient boosting trees, the learning rate is of major importance (FRIEDMAN, 2001).

An common approach to optimize the hyperparameters is to perform grid search. However, this approach suffers from spending too much effort in exploring unimportant dimensions. By grid searching, the global optima may not be achieved. To overcome these drawbacks from grid search, Bergstra and Bengio (2012) investigated the random search to optimize the hyperparameters of machine learning methods. The authors concluded that randomly chose trials are more efficient for hyperparameter optimization, especially in multi-dimension searching space.

Snoek, Larochelle and Adams (2012) presented an automatic way to hyperparameters optimization problem with Bayesian Optimization. This method considers that an algorithm’s generalization performance is a sample from a Gaussian Process, enabling to efficient use the information from previous experiments in order to determine the next try. The authors showed that this procedure could even outperform expert-level performance in optimizing the machine learning algorithms.

Other methods for global optimization can be employed to determine the optimum hyperparameters. Biology inspired algorithms, such as genetic algorithm and particle swarm, exploit the information from previous results in order to determine the next possible trials. Open sources packages enables to apply these techniques. For example, the “Optunity” (CLAESEN et al., 2014) implements several optimization methods, including grid search, random search, particle swarm optimization and several others. The package “GA” implements a wide range of genetic

algorithm methods in programming language R (SCRUCCA, 2013). Another example is the package “DEAP” (Distributed Evolutionary Algorithms in Python) (FORTIN et al., 2012). We just mentioned three packages, but it is possible to find many other packages and libraries.

2.2.8 Feature Selection

Feature selection is a process of selecting an subset of features (input variables) to feed a machine learning algorithm. By selecting a small subset of relevant features, the following benefits can be achieved: faster learning process, simpler model and better accuracy (XUE et al., 2016). However, the main drawback is that the search space can be very huge.

Instead of “brute force” (extensive search), other approaches can be employed to overcome the challenge of finding the best subset of features. Guyon and Elisseeff (2003) stated that greedy search strategies (forward selection and backward elimination) can be computationally advantageous and robust against overfitting. The forward selection incorporates progressively the variables that improve the accuracy of model. In backward elimination, the model starts with all inputs, and then eliminates progressively the least promising inputs.

Very similar to hyperparameter optimization, it is possible to employ evolutionary computation approaches (e.g. genetic algorithm or particle swarm optimization) to obtain an optimum subset of features (XUE et al., 2016). These approaches have the ability to reach the global optima, or to obtain a result that is near to the optimum point.

2.3 INTRODUCTION TO MULTI-OBJECTIVE OPTIMIZATION

If it is desired to determine the optimum values of the controllable drilling parameters (e.g. weight on bit and rotary speed), then several combinations of these parameters needs to be evaluated, assessing the respective impact on drilling performance indicators (i.e. objective functions).

The common approach is to determine the optimum combination of e.g. WOB and rotary speed that maximizes the ROP. In this case, maximization of ROP is the only objective function, leading to a single-objective optimization problem. If multiple drilling performance indicators are taken into account, e.g. maximizing ROP and maximizing bit-life, a multi-objective optimization needs to be formulated. This is because more than one objective is considered, and they may be contradictory to each other.

Here, we present the general frameworks of the single-objective optimization problem and the multi-objective optimization problem.

2.3.1 The Single-Objective Optimization Problem

Depending on the problem, we seek to either maximize or minimize an objective function $f(\mathbf{x})$. Cui et al. (2017) showed that the general problem of single-objective optimization can be

defined as a minimization problem, because with the transform

$$\max f(\mathbf{x}) \Leftrightarrow \min (-f(\mathbf{x})) \quad (2.18)$$

it is possible to transform a maximization problem into a minimization problem and vice versa.

According to Chiandussi et al. (2012), a general single-objective optimization can be defined as the minimization objective function $f(\mathbf{x})$, subject to inequality constraints $g_i(\mathbf{x}) \leq 0, i = \{1, 2, \dots, p\}$ and equality constraints $h_j(\mathbf{x}) = 0, j = \{1, 2, \dots, q\}$.

Let us suppose that we wish to determine the optimum solution $\mathbf{x}^*$ that minimizes a given objective function f . This function may have several local minima, so that the employed optimization algorithm may achieve a local minimum point of f instead of converging to the global optima. The Global Optimization methods try to find the global optimum solution, avoiding the problem of being trapped in local minima (CHIANDUSSI et al., 2012).

2.3.2 The Multi-Objective Optimization Problem

A multi-objective problem aims to optimize simultaneously multiple objective functions, and can be formulated by the equation (ZHOU et al., 2011):

$$\begin{aligned} &\text{minimize} \quad F(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x})]^T \\ &\text{s.t.} \quad \mathbf{x} \in \Omega \end{aligned} \quad (2.19)$$

where the decision variable $\mathbf{x}$ belongs to the decision space Ω . In the case of m -objective functions, the objective space belongs to a m -dimensional vector space $\mathbb{R}^m$.

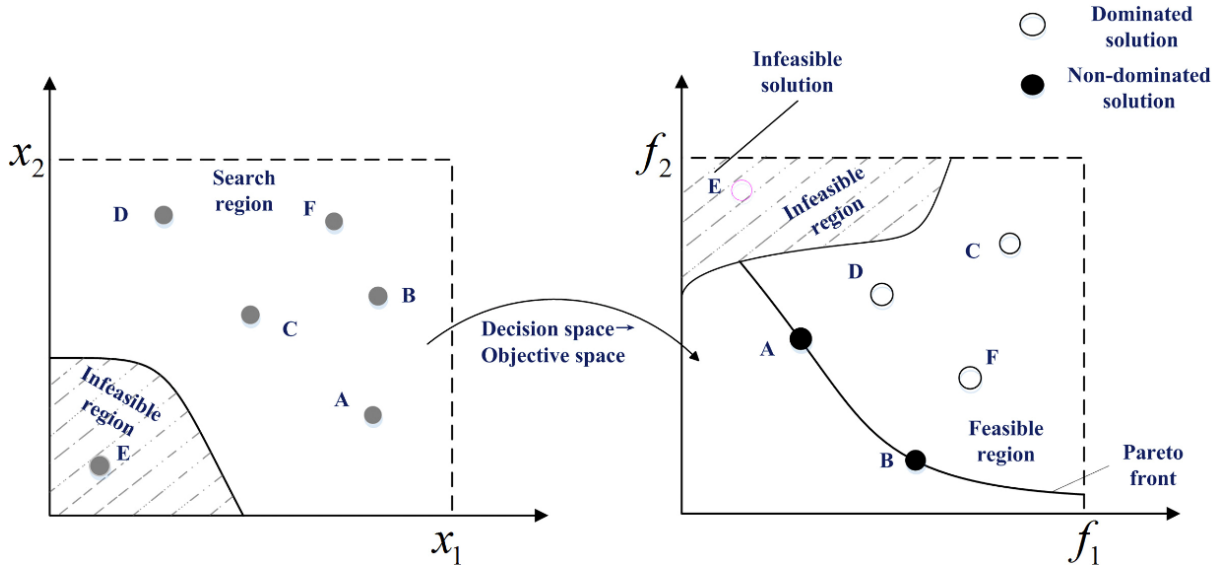
The objectives in equation 2.19 are often contradictory to each other, i.e. an improvement of one objective may lead to deterioration of another. As consequence, there is no single optimum solution able to optimize all objectives simultaneously. In multi-objective problems, a set of optimal solutions are obtained instead of a single one solution. This set is called *Pareto optimal solutions* (ZHOU et al., 2011).

Several works explain the concept of Pareto optimality (ANTONIO; COELLO, 2017; CUI et al., 2017; CHIANDUSSI et al., 2012; ZHOU et al., 2011). Here, we present the definition of Pareto optimal solutions based on Zhou et al. (2011); extensive details about the definitions regarding multi-objective optimization can be found in Chiandussi et al. (2012).

Definition 1 (Zhou et al. (2011)) A vector $\mathbf{u} = (u_1, u_2, \dots, u_m)^T$ is said to dominate another vector $\mathbf{v} = (v_1, v_2, \dots, v_m)^T$, denoted as $\mathbf{u} \prec \mathbf{v}$, if $\forall i \in \{1, 2, \dots, m\}, u_i \leq v_i$ and $\mathbf{u} \neq \mathbf{v}$.

Definition 2 (Zhou et al. (2011)) A feasible solution $\mathbf{x}^* \in \Omega$ of the problem 2.19 is called *Pareto Optimal Solution*, if $\nexists \mathbf{y} \in \Omega$ such that $F(\mathbf{y}) \prec F(\mathbf{x}^*)$. The set of all the Pareto Optimal Solutions is called *Pareto Set (PS)*, denoted as $PS = \{\mathbf{x} \in \Omega \mid \nexists \mathbf{y} \in \Omega, F(\mathbf{y}) \prec F(\mathbf{x})\}$. The image of the PS in the objective space is called the *Pareto Front (PF)*, denoted as $PF = \{F(\mathbf{x}) \mid \mathbf{x} \in PS\}$.

Figure 2.12 – Input space and objective space for the case of two objective functions



Source: Cui et al. (2017).

A general multi-objective optimization problem can be subject to inequalities and equalities (ANTONIO; COELLO, 2017; CUI et al., 2017; ZHOU et al., 2011). In this case, the search space Ω can be formulated as follows (ZHOU et al., 2011):

$$\Omega = \begin{cases} g_i(\mathbf{x}) \leq 0, & i = \{1, 2, \dots, p\} \\ h_j(\mathbf{x}) = 0, & j = \{1, 2, \dots, q\} \\ x_i^{\min} \leq x_i \leq x_i^{\max}, & i = \{1, 2, \dots, n\} \end{cases} \quad (2.20)$$

where Ω is a n -dimensional search space for the decision variable $\mathbf{x}$, determined by the upper bound $\mathbf{x}^{\max} = [x_1^{\max}, x_2^{\max}, \dots, x_n^{\max}]^T$ and the lower bound $\mathbf{x}^{\min} = [x_1^{\min}, x_2^{\min}, \dots, x_n^{\min}]^T$, p inequalities $g_i(\mathbf{x}) \leq 0, i = \{1, 2, \dots, p\}$, and q equalities $h_j(\mathbf{x}) = 0, j = \{1, 2, \dots, q\}$. These constraints lead to two regions: a feasible region and infeasible region. For the special case $p = q = 0$, the multi-objective optimization problem is said to be unconstrained (CUI et al., 2017).

The Figure 2.12 illustrates the input space and objective space for the case of two objective functions. We can observe that the feasible solution A is a good solution for a objective function f_1 , but the solutions B and F are better than A with respect to the objective f_2 . The solutions A and B dominate all other feasible solutions (C, D and F). However, the solution A does not dominate B and vice versa. In this example, the Pareto Set is formed by the non-dominated solutions A and B (CUI et al., 2017).

2.3.3 Multi-Objective Optimization Techniques

The aim of multi-objective problem is to find good compromise solutions (or trade-offs). The solution of a multi-objective optimization problem consists basically of two main steps

(CHIANDUSSI et al., 2012): one step is to generate the Pareto Set; another step is how the *decision maker* selects one or more optimum solutions from the Pareto Set. The search process and decision making can be combined in several ways. The works (CHIANDUSSI et al., 2012; MARLER; ARORA, 2004; HWANG; MASUD, 1979; COHON; MARKS, 1975) classified the multi-objective decision making into four groups according to the stage at which the information is needed for *decision maker*:

- No Articulation of Preference Information;
- *A Priori* Articulation of Preference Information - the *decision maker* takes decision prior searching, e.g. by pre-ordering objectives;
- *A Posteriori* Articulation of Preference Information - first a search is carried out without any prior preference of the *decision maker*;
- Progressive Articulation of Preference Information - interactive methods which integrate search and decision making.

In the following section, we present some decision making techniques that aggregate multiple objectives into a single objective function. By having a single objective function, it is possible to apply standard optimization engines. The main advantage of this approach is its simplicity (MARLER; ARORA, 2004). This approach is employed in the current work.

2.3.4 Decision Making Techniques

To access several methods for decision making with multiple objective functions, the readers can refer to Marler and Arora (2004). In this current work, we present three decision making techniques:

- the Global Criteria Method (no articulation of preference), what can be extended for Weighted Global Criteria Methods (*a priori* articulation of preference);
- Weighted Sum Method (*a prior* articulation of preference articulation);
- The ε -constraint method (*a posteriori* articulation of preference).

2.3.4.1 Global Criteria Method

The Global Criteria Method seeks to minimize a global criterion which measures how close a solution is to the ideal vector $\mathbf{F}^0$ (CHIANDUSSI et al., 2012). The ideal vector, which is also known as *utopia point* (MARLER; ARORA, 2004), can be defined as follows.

Definition 3 (Chiandussi et al. (2012)) *The optimum value for the i -th objective function is denoted as f_i^0 . The ideal vector, denoted as $\mathbf{F}^0 = [f_1^0, f_2^0, \dots, f_m^0]^T$, is the ideal solution where all m objective functions achieve separately their optimum values.*

Chiandussi et al. (2012) formulated the function of global criterion as follows:

$$f(\mathbf{x}) = \sum_{i=1}^m \left(\frac{f_i^0 - f_i(\mathbf{x})}{f_i^0} \right)^p \quad (2.21)$$

where m is the amount of objective functions. Common values for the exponent are $p = 1$ or $p = 2$, but this parameter can assume any other value. The value of p has a great impact on optimum solutions (CHIANDUSSI et al., 2012).

Another global criteria method is based on L_p -metrics, which measure the relative distance to the ideal vector (CHIANDUSSI et al., 2012):

$$L_p(f) = \left[\sum_{i=1}^m \left\| \frac{f_i^0 - f_i(\mathbf{x})}{f_i^0} \right\|^p \right]^{1/p}, \quad 1 \leq p \leq \infty \quad (2.22)$$

The previous equations 2.21 2.22 are mathematical functions with no articulation of preference information (MARLER; ARORA, 2004), but they can be written as Weighted Global Criteria. That is, the *decision maker* can set different weights for each objective function. In this case, the so-called Weighted Global Criteria is a method with *a priori* articulation of preference information. For more details see Marler and Arora (2004).

The main advantages of the global criteria methods are their simplicity and effectiveness. The downsides of them are computational effort to define the desired goal, and the utopia point must be in the feasible region (CHIANDUSSI et al., 2012).

2.3.4.2 Weighted Sum Method

Another way to transform the vector $\mathbf{F}$ into a scalar objective function is to perform a weighted sum of the objective functions (CUI et al., 2017; MARLER; ARORA, 2004). The scalar objective function is expressed as following (CHIANDUSSI et al., 2012):

$$\min \sum_{i=1}^m \alpha_i f_i(\mathbf{x}), \quad \text{s.t. } \mathbf{x} \in \Omega \quad (2.23)$$

where $\alpha_i \geq 0$ for all i . By varying these weights, α_i , it is possible to obtain the Pareto Set (CHIANDUSSI et al., 2012). This was demonstrated by Zadeh (1963).

The objective function 2.23 is a linear combination of the objectives functions. Therefore, it is possible to find some authors (CHIANDUSSI et al., 2012) calling this method as Linear Combination of Weights.

The simplicity inherent in solving a scalar objective function is the main advantage of linear combination weights. However, the magnitude of weights plays a key role in this method. Therefore, the *decision maker* must carefully determine them. Another drawback of this approach is the inability of equation 2.23 to generate optimum solutions in a concave Pareto Front (CHIANDUSSI et al., 2012).

2.3.4.3 The ε -constraint method

Haimes, Lasdon and Wismer (1971) formulated the optimization problem of two objective functions as a constraint optimization problem, calling this method as ε -constraint formulation. In addition to sum approach, this technique “is probably the best known technique to solve multi-criteria optimization problems” (CHIANDUSSI et al., 2012, p. 919).

In multiple objective functions, the so-called ε -constraint method consists of optimizing only one objective function, while all other objective functions are transformed into inequalities fixed by a threshold ε . We can substitute the general problem of multi-objective optimization (equation 2.19) by the ε -constraint problem (CHIANDUSSI et al., 2012):

$$\begin{aligned} \min f_j(\mathbf{x}) \\ \text{s.t. } f_k(\mathbf{x}) \leq \varepsilon_k, \quad k = \{1, 2, \dots, m\}, k \neq j \end{aligned} \quad (2.24)$$

According to Chiandussi et al. (2012), the main disadvantages of the ε -constraint method are: its possible high computational cost and the need of setting the preliminary individual values ε_i . On the other hand, the main advantage of this techniques is the relative simplicity of implementing it, being therefore popular particularly in the engineering field.

The proposed optimization method in the current work employed the ε -constraint technique. However, other decision making techniques could also have been employed. For future works, it is worth of investigating other strategies in the solution of multi-objective problems applied to the drilling engineering.

2.3.5 Optimization Techniques

According to Cui et al. (2017), two kinds of methods can be employed to solve optimization problems: (a) analytical method and (b) numerical method. The analytical method reaches the exact solution, but it may not solve real problems. Numerical methods employ iterative calculation procedures in order to reach an approximate solution. With numerical methods, the objective functions can be black box models, such as those any machine learning methods seen in Section 2.2.

The numerical optimization methods can be split into two main groups (CUI et al., 2017): classical methods and intelligent methods. The classical methods (e.g. Newton iteration, simplex) have high searching efficiency and fast convergence, but they normally require gradient information. On the other hand, the Intelligent Methods employ heuristic search algorithms based on phenomena from the nature. It is possible to classify the intelligent optimization methods into four groups (BEHERA; SAHOO; PATI, 2015; CUI et al., 2017): biology inspired algorithms, physics inspired algorithms, geography inspired algorithms, and social-culture inspired algorithms.

In the drilling optimization problem, it is possible to perform the same optimization approaches seen for Hyperparameter Optimization (Section 2.2.7) and Feature Selection (Section

2.2.8). For this present study, we selected the grid search strategy. However, any other techniques could be employed, such as random search, and intelligent optimization methods (e.g. genetic algorithm, particle swarm optimization, differential evolution algorithm). The main characteristics of these techniques are the following:

- the main advantage of the grid-search is its simplicity in implementation. However, the computational cost may be high. The global optima is not guaranteed to be reached. Another drawback is the inefficient search strategy, because too much effort is employed in not so important variable (BERGSTRA; BENGIO, 2012);
- the main advantage of random search is the computational efficiency (BERGSTRA; BENGIO, 2012). With this approach is possible to carry out an extensive in the decision space Ω . However, it is not guaranteed to find a optimum solution.
- the so-called Biology inspired algorithms for optimization problems are based on evolutionary theorem from Darwin and swarm behaviour of particles and animals (CUI et al., 2017; BEHERA; SAHOO; PATI, 2015). For more details about genetic algorithms, the readers can refer to a comprehensive overview given by Roberts et al. (2017). For more details about particle swarm optimization, Zhang, Wang and Ji (2015) provided a comprehensive survey, showing also its application.

2.3.6 Brief Remark on Multi-Objective Optimization Algorithms

The intelligent optimization methods for multiple objectives are state-of-art solutions to approximate the Pareto Front in several complex problems. Some algorithms for multi-objective problems are Non-dominated Sorting Genetic Algorithm (NSGA-II), Multiple Objective Particle Swarm Optimization (MOPSO), Nondominated Neighbor Immune Algorithm (NNIA) (ANTONIO; COELLO, 2017; CUI et al., 2017; CHIANDUSSI et al., 2012; ZHOU et al., 2011).

Such algorithms are global optimization methods. However, in multi-objective optimization, the term of global solution is unclear, because there is usually a set of optimum solutions (Pareto Set).

The possibility of having several optimum solutions paves the way to implement a more sophisticate selection of optimum drilling variables. For the case of drilling, it is normally desired to optimize the drilling variables, without applying too oft changes. Abrupt changes in drilling controllable variables are neither possible nor desired (GANDELMAN, 2012). A more sophisticate selection of optimum drilling variables can avoid oft changes in magnitude of drilling variables. However, avoiding this problem goes beyond the scope of the current work.

3 MACHINE LEARNING METHODS APPLIED TO RATE OF PENETRATION PREDICTION AND OPTIMIZATION - A REVIEW

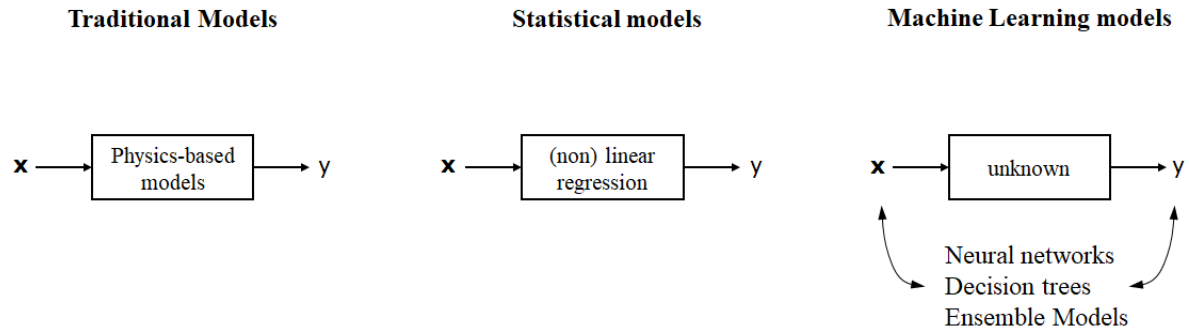
This chapter reviewed the current progress on using machine learning methods and statistics methods to estimate the Rate of Penetration (ROP). A preliminary version of this study was presented in a Conference (BARBOSA et al., 2018). That paper reviewed partially the use non-traditional models to predict the ROP; 11 studies from a total of 45 works were analyzed. In addition, that work was restricted to ROP modeling. The current chapter extended the previous work: (i) by increasing the compilation of works from 45 to 58, and (ii) by showing how the ROP-models can be employed to optimize the controllable drilling variables.

Some reviews have recently covered the applicability of artificial intelligence methods in drilling engineering. However, none of them focused exclusively on current progress of applying machine learning techniques to predict the ROP. Bello et al. (2016) reviewed the application of artificial intelligence in several fields related to oil and gas industry, for example: reservoir simulation, seismic pattern recognition, reservoir characterization, permeability and porosity prediction, drill bit diagnosis, well production optimization and so on. The review from Agwu et al. (2018) covered the application of artificial intelligence techniques applied to drilling fluid engineering. Rahmanifard and Plaksina (2018) reviewed several techniques of artificial intelligence, giving a special attention to heuristic optimization methods, such as genetic algorithm, particle swarm and differential evolution. Rahmanifard and Plaksina (2018) showed also how these techniques were applied in themes related to oil and gas industry in general (e.g. minimum miscibility pressure, oil production rate, and volume of CO₂ sequestration). To the best author knowledge, no trial has been already attempted to review the current progress on machine learning techniques applied to rate of penetration prediction and optimization.

3.1 METHODS USED IN ROP PREDICTION

Hegde et al. (2017) classified the ROP models into two groups: traditional (physics-based) models and data-driven models (regression models and machine learning methods). In the current work, the data-driven models were split into statistical models and machine learning models. The resulting classification is shown in Figure 3.1, and is the same general classification adopted by Breiman (2001b).

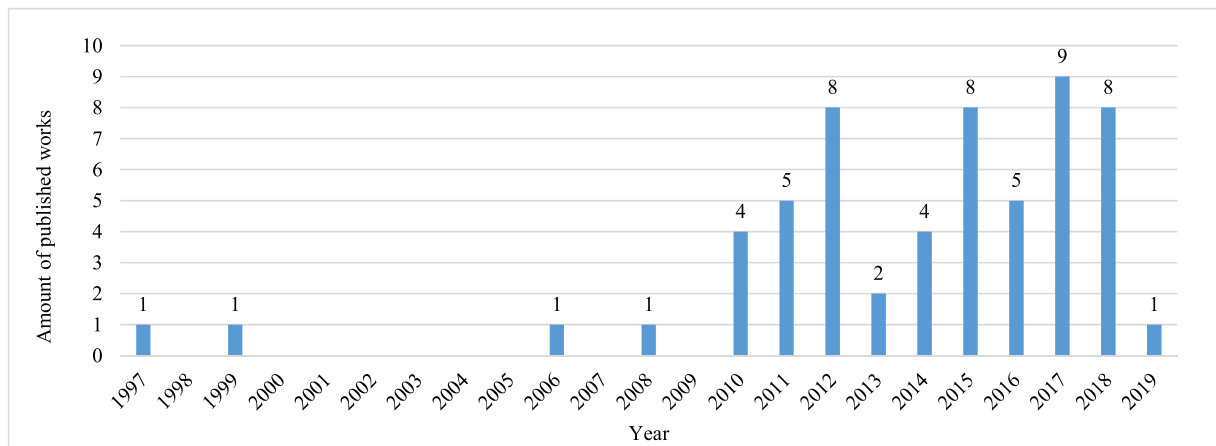
Figure 3.1 – Approaches for ROP modeling



Source: adapted from Breiman (2001b)

This chapter focus on those works that used non-traditional models to predict the ROP. The use of machine learning to predict drilling parameters started in 1990s. Arehart (1990) employed neural networks to predict an important drill bit parameter (bit wear). Laboratory data for the training process was employed. Later, Bilgesu et al. (1997) published the first work that applied artificial neural networks to predict the ROP. After a long period without further development in this area, researchers have been publishing several works since 2010, as shown in Figure 3.2.

Figure 3.2 – A compilation of 58 works found on the literature considering thesis and papers (publications in journals and congresses) using other methods different than the traditional models for ROP prediction. Status: October 2018.



Source: an update from Barbosa et al. (2018).

3.1.1 Traditional Models

Here, those ROP models which try to establish mathematical equations among the drilling variables are called as traditional models. This is because most of these models appeared in the initial phase of the scientific research of drilling optimization (EREN; OZBAYOGLU, 2010). In addition, those models do not rely solely on the drilling data, as the machine learning models do. As consequence, it is possible to find out other researchers (HEGDE et al., 2017) calling

them as physics-based models, or as simple as drilling models (HARELAND; HOBEROCK, 1993). Recently, Soares and Gray (2018) called these models as analytical models.

There are many models describing the effects of several parameters on ROP. Eren and Ozbayoglu (2010) provided a literature survey, explaining some traditional models. The readers can also refer to other works that provided an extensive review about the ROP models (SOARES; GRAY, 2018; SOARES; DAIGLE; GRAY, 2016; NASCIMENTO, 2016; MITCHELL; MISKA, 2011; HARELAND; HOBEROCK, 1993).

Some of traditional models which are worth of citing, because they are used in some papers for comparison purpose with data-driven models, are the following: Graham and Muench (1959), Maurer (1962), Bingham (1965), Young (1969), Bourgoyne and Young (1974), Warren (1987), Hareland and Rampersad (1994). One of the most important studies on drilling optimization was developed by Bourgoyne and Young (1974), where a multiple regression analysis of the drilling data was conducted to achieve minimum drilling costs by maximizing the ROP in a typical operation.

As already mentioned, the actual relationship among the drilling variables is not very well understood (MITCHELL; MISKA, 2011). Therefore, some efforts (DENG et al., 2016; MOTAHHARI; HARELAND; JAMES, 2010) have been made to better understand the relationship between the drilling variables and how they affect the ROP. Deng et al. (2016) proposed a theoretical model for determining the ROP for roller cone bit, and this model was validated with lab drilling results. The authors used the rock dynamic compressive strength instead of static compressive strength, what increased the accuracy of the theoretical model.

Some researchers have recently tried to obtain ROP-models for specific wells drilled based on traditional models. Such studies consisted of fitting the empirical coefficients from the ROP-models, tailored to drill curves available. One example of application of an traditional ROP model was carried out by Nascimento et al. (2015b). The authors employed the Bourgoyne and Young ROP model (BYM) for presalt layers. This study showed that normalizing the factor of BYM formulation and allowing a wider range of applicable drillability coefficients could decrease the relative error of the ROP prediction from 46% to 27%, respectively. Bahari et al. (2008) employed genetic algorithm to determine the coefficients of Bourgoyne and Young model to predict the ROP. Formighieri and Filho (2016) estimated these coefficients with Markov Chain Monte Carlo simulation. In additional, other works employed the traditional models to predict the ROP (SOARES; DAIGLE; GRAY, 2016; KUTAS et al., 2015; GANDELMAN, 2012; EREN; OZBAYOGLU, 2010).

ROP modeling is not restricted to drilling of oil and gas wells. Basarir, Tutluoglu and Karpuz (2014) mentioned other ROP models, applied to rock excavation and tunnel boring, showing the correlation between the rock properties and the ROP.

3.1.2 Statistical Models

Multiple regression can be applied to model the ROP as functions of drilling parameters. In this case, the dependent variable (ROP) is a function of independent variables $(X_1, X_2, \dots, X_K)$, where K represents the amount of inputs into the ROP model. When modeling the ROP with this approach, a specific regression method must be selected. Moraveji and Naderi (2016) applied the full quadratic form of multiple regression with linear, quadratic and interaction coefficients. The authors modeled the ROP as function of six drilling parameters, as listed in the Table 3.1.

Table 3.1 – ROP modeling using multiple regression carried out by Moraveji and Naderi (2016)

Drilling Data Source	Amount of Inputs	Model	Accuracy %
1732 observations from one southern Iranian gas field	6	Full quadratic form of multiple regression	$R^2 = 71.5$ $R_{adj}^2 = 71.1$

Source: adapted from Moraveji and Naderi (2016), Barbosa et al. (2018)

There are also other works (ESKANDARIAN; BAHRAMI; KAZEMI, 2017; HEGDE et al., 2017; ARABJAMALOEI; SHADIZADEH, 2011) that employed multiple regression, with either linear coefficients or both linear and non-linear coefficients. Their purpose, however, was to compare different techniques of modeling the ROP (e.g. multiple regression with machine learning techniques), or to select the most important features in ROP prediction.

In this review, the work from Moraveji and Naderi (2016) is the only one found that focused on estimating the ROP with statistical regression models. The downside of pre-selecting a specific regression method for any real-world problem was discussed by Breiman (2001b). Breiman compared two different cultures in the use of statistical modeling: one called “data models”, and another called “algorithmic models”. The first assumes the data is generated by a given stochastic data model (as done by Moraveji and Naderi (2016)), and the last approach is based on algorithms (such as neural networks, decision trees) that are able to learn from data, adapting themselves to the problem. Breiman (2001b) mentioned if the goal is to use data to solve real-world problems, then it is necessary “to move away exclusive dependence on data models and adopt a more diverse set of tool”, such as machine learning techniques. Perhaps, this is the reason why most of works preferred to use machine learning methods to obtain ROP-models, instead of multiple-regression methods.

3.1.3 Machine Learning Applied to Predict the ROP

Here the Table 3.2 is reproduced from Barbosa et al. (2018). A summary of 10 works was given, focusing mainly on those studies that compared the ROP prediction with machine learning techniques with other methods, which serves as the basis for the upcoming discussions.

Table 3.2 – Some works that used Machine Learning Methods to predict the ROP.

(continued)

Authors	Drilling Data Source (amount of points used)	Machine Learning Method ^(a)	Amounts of Inputs	Compared with other methods?	ML was the best?	Testing Accuracy of ML %
Bilgesu et al. (1997)	Laboratory Data from rig floor simulator (8000)	Neural Networks SLFN	(a) – 10 (b) – 6	✗	-	(a) $R = 98.2$ (b) $R = 95.5$
	USA - Field Data from several wells (500)	Neural Networks SLFN	(a) – 10 (b) – 8	✗	-	(a) $R = 96.5$ (b) $R = 98.0$
Arabjamaloei and Shadizadeh (2011)	Iran - 35 wells drilled in Ahwaz oilfield (330)	Neural Networks SLFN	7	Yes, with multiple regression and BYM	✓	$R^2 = 74.0$
Amar and Ibrahim (2012)	Mediterranean Sea - Offshore, provided by Eren and Ozbayoglu	Neural Networks RBF and ELM	7	Yes, with BYM	✓	$APRE = 9.6$
Basarir, Tutluoglu and Karpuz (2014)	Turkey – 7 boreholes drilled in 6 different regions	Adaptive neuro-fuzzy inference system (ANFIS)	4	Yes, with linear and non-linear multiple regression	✓	$RMSE = 0.33\#$ (validation)
Bataee, Irawan and Kamyab (2014)	15 wells (1810 points for training)	Neural Networks MLP	5	Yes, with Bingham, BYM, Warren	✓	$R^2 = 85.7$ (validation)
Ansari, Hosseini and Amirpour (2017)	Persian Gulf - 19 wells (248)	Support vector regression	8 after feature selection	✗	-	$R = 90.6$
Bezminabadi et al. (2017)	Iran – Azadegan Oilfield	Neural Networks SFLN	(a) – 5 (b) – 9	Yes, with multivariate non-linear regression	✓	(a) $R = 75$ (b) $R = 86$

Table 3.2 Some works that used Machine Learning Methods to predict the ROP.

(conclusion)

Authors	Drilling Data Source (amount of points used)	Machine Learning Method ^(a)	Amounts of Inputs	Compared with other methods?	ML was the best?	Testing Accuracy of ML %
Eskandarian, Bahrami and Kazemi (2017)	Iran - 5 wells in South West (226)	3 methods - Cubist, RF and MON-MLP	4 and 6 after feature selection of 13 non-constant variables	Yes, with linear multiple regression	✓ overall MON-MLP	$R^2 = 80.1$ (10 fold cross-validation)
Hegde et al. (2017)	USA – 1 vertical well	RF	4	Yes, with (Bingham, Motahhari, Hareland) and linear regression	✓	~ 13 Normalized Error (Median for all formation)
Diaz et al. (2018)	South Korea - 1 well (7034)	Neural Net-works MLP	7 after feature selection	Yes, with BYM	✓	$R^2 = (90, 99)$

^(a) APRE: absolute percent relative error; ELM: extreme learning machines; MLP: multi-layer perceptron (with more than one hidden-layer); MON-MLP: monotone multi-layer perceptron; SFLN: single hidden-layer networks (a MLP with one hidden-layer); RBF: radial basis function networks

Source: Barbosa et al. (2018)

3.2 DISCUSSION ON ROP MODELS

Here, the original work (BARBOSA et al., 2018) was extended, by detailing especially the Section 3.2.4 (The importance of geological formation), the Section 3.2.5 (Data Partition), and the Section 3.2.6 (Drilling Data).

3.2.1 Machine Learning Algorithms Outperforms Other Methods

In Table 3.2, we can observe that 8 out of 10 works compared machine learning with other techniques (traditional models and/or regression models). All these comparative works showed that the use of learning algorithms provided a better ROP prediction. The reason for that is the capability of those models to capture non-linear relationship among the variables.

In this selected sample of works (Table 3.2), most of them applied neural networks to predict the ROP. However, it is not possible yet to affirm which method is preferred or more employed in this type of study due to the small amount of works that were deeply analyzed. Other methods, such as support vector regression and Random Forest (RF), were employed also in the ROP prediction tasks.

The use of Artificial Neural Networks (ANN) resulted in predictive models with good generalization capability as seen in Table 3.2. However, these models are complex. For that reason, ANNs are commonly referred as black-boxes, because it is not easy to understand them. On the other hand, it is possible to employ some techniques from which understandable rules can be extracted. A fake example of these rules are: if WOB is low, then ROP is low; if WOB is high, then ROP is high. Eskandarian, Bahrami and Kazemi (2017) employed Random Forest to model the ROP. The authors could split the ROP values into three levels: low, medium and high. In this case, some rules of thumb could be extracted for each ROP level. The work from Basarir, Tutluoglu and Karpuz (2014) employed the Adaptive Neuro-Fuzzy Inference System (ANFIS). With this technique, the authors could understand some simple rules about how the drilling variables affected the ROP.

3.2.2 Sensitive Analysis

The rule extraction is one way to assess the influence of the drilling variables on the ROP. Another way is to perform sensitive analysis¹, i.e. varying some drilling variables, while others remain unchanged. Eskandarian, Bahrami and Kazemi (2017) performed this analysis, helping the authors to understand the relationship among the variables, as seen in Figure 3.3. By analyzing the plots, the authors determined the range of controllable drilling parameters at which ROP is near to the maximum point. Another study carried out similar analyses as well (ARABJAMALOEI; SHADIZADEH, 2011).

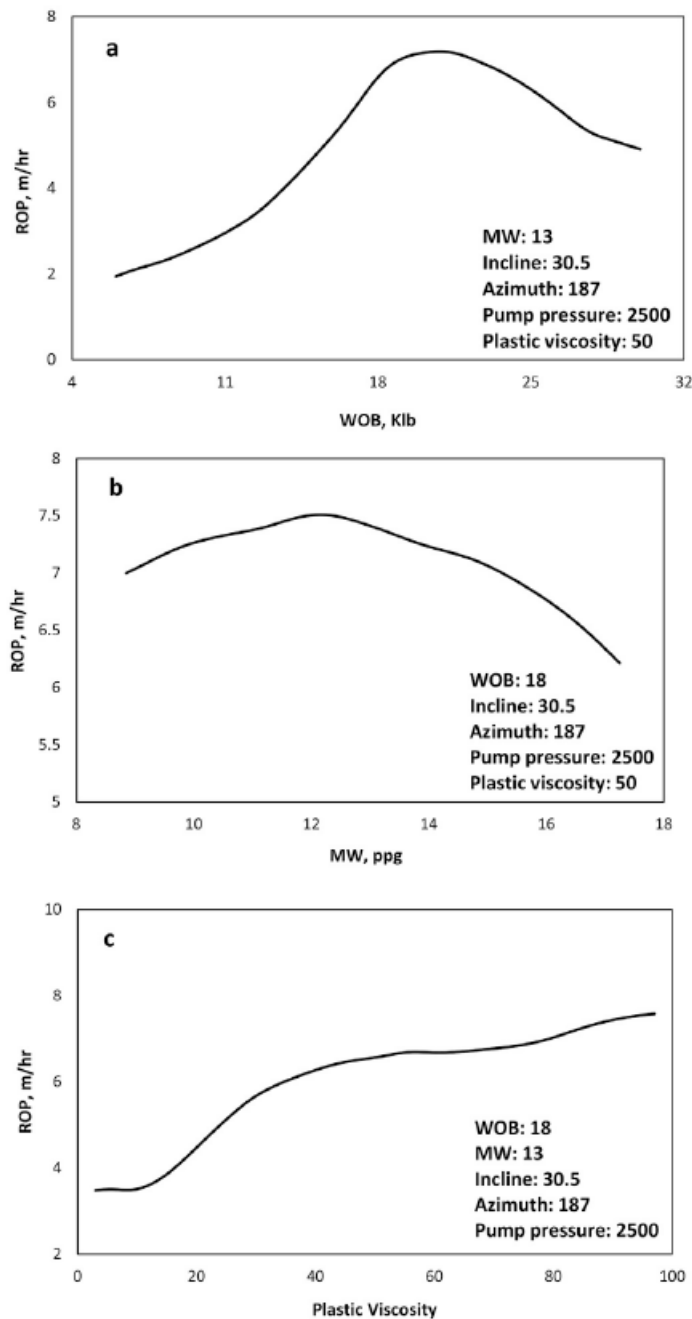
Moraveji and Naderi (2016) performed an extensive sensitive analysis, after obtaining a ROP-model based on multiple regression methods (see Table 3.1). Just to illustrate the sensitive analysis, we added the Figure 3.4. This figure shows the interaction effects of Depth (D) against other two other variables: a) depth vs. weight on bit; b) depth vs. rotation speed. In let figure, we can observe that:

the reduction of penetration rate by increasing weight on bit (bit floundering effect) occurs because of less efficient hole cleaning and cutting transport at higher rates of cuttings generation or excessive bit tooth wear as a result of complete penetration into the formation being drilled (MORAVEJI; NADERI, 2016, p. 835–838).

The Figure 3.4 B illustrates the interaction D x drill bit rotation speed. “The reduction of penetration rate by increasing bit rotation speed occurs because of reduced stability and increased size of the wellbore” (MORAVEJI; NADERI, 2016, p. 838). The readers can refer to the original work to assess the complete parametric analysis.

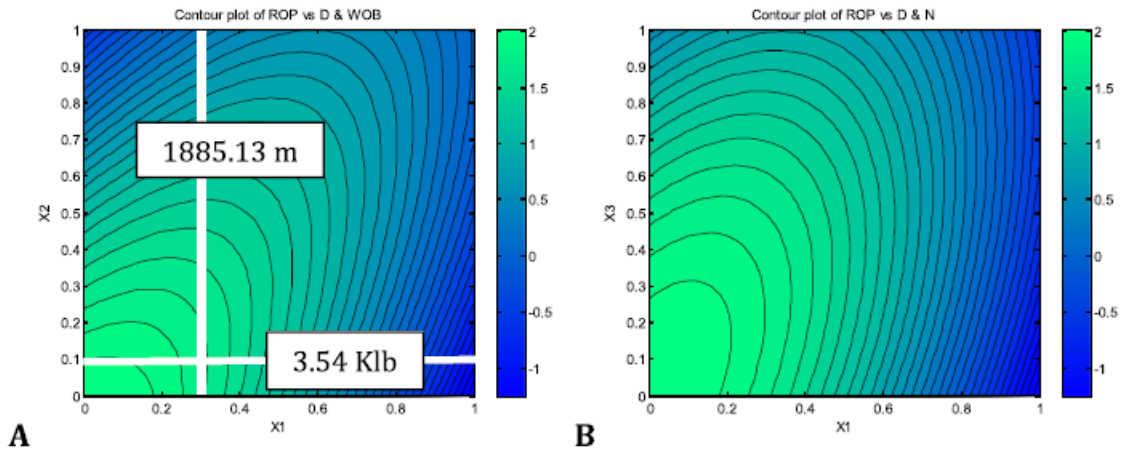
¹ We changed the original name of this Section from “Parametric Influence” to “Sensitive Analysis”. The reason for this change is to avoid any confusion to the reader, who may associate the term parametric with hyperparameters of machine learning methods. However, the objective of this section is clearly different; it aims to present some methods to assess how the some inputs (e.g. WOB and RPM) affect the output (ROP).

Figure 3.3 – Using neural networks to assess parametric analysis: a) effect of WOB, b) mud-weight (MW), and c) plastic viscosity on ROP



Source: Eskandarian, Bahrami and Kazemi (2017)

Figure 3.4 – Contour plot for ROP versus normalized values of *A*) depth (X_1) and weight on bit (X_2), *B*) depth (X_1) and bit rotation speed (X_3)



Source: Moraveji and Naderi (2016)

3.2.3 Feature Extraction

Some studies (ANSARI; HOSSEINI; AMIRPOUR, 2017; BEZMINABADI et al., 2017; ESKANDARIAN; BAHRAMI; KAZEMI, 2017; BILGESU et al., 1997) analyzed which combination of inputs provided a better ROP prediction. This type of analysis is commonly known as feature selection, where, through some method, only those variables with significant impact on the prediction are selected. Eskandarian, Bahrami and Kazemi (2017) showed a comprehensive way of assessing the most important parameters for modeling the ROP, as shown in Figure 3.5. In that case, the authors used an open-source package called *fscaret* (SZLEK; MENDYK, 2018), which is written in R-programming language and provided an automated way to assess the most important drilling parameters on ROP prediction. Hegde et al. (2017) showed also the possibility to perform a similar analysis with random forests, by using the variable importance.

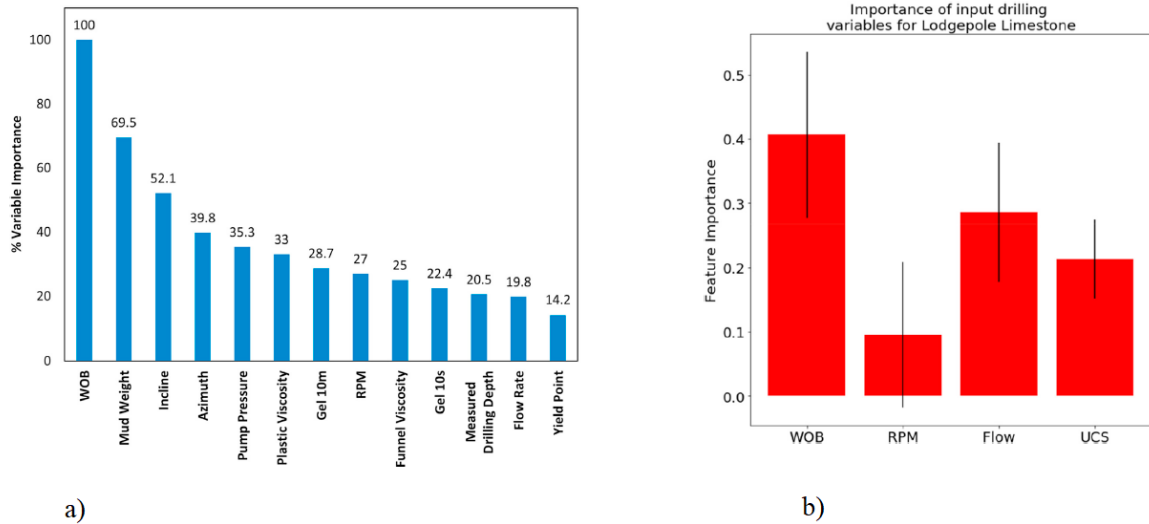
However, it is possible to find some studies that only mentioned the use of the best combination of inputs (DIAZ et al., 2018; ARABJAMALOEI; SHADIZADEH, 2011). Other used a set of inputs, based drilling engineering knowledge, not providing any further discussion on the selection of the variables (AMAR; IBRAHIM, 2012).

3.2.4 Importance of Geological Formation in ROP Prediction

It was found four possible approaches regarding the use of geological formation on ROP prediction. The first approach does not take into account any parameter related to geological formation or lithology. In the second approach only part of the inputs are related somehow to the drilled formation. The third approach emphasizes the importance of the formation on ROP prediction. In the fourth approach, drilling data is split for each geological formation or lithology.

Bilgesu et al. (1997) reported the possibility to obtain reliable predictive ROP models by only using surface parameters (weight on bit, drill bit rotation, mud flow rate, and so on). The

Figure 3.5 – Variable importance: *a)* automated feature selection resulted applied by Eskandarian, Bahrami and Kazemi (2017) with *fscaret*, *b)* based on random forest ROP predictor performed by Hegde et al. (2017)



Source: a) Eskandarian, Bahrami and Kazemi (2017), b) Hegde et al. (2017)

authors trained two networks on the data from rig floor simulator. The case (*a*) had 10 inputs, and (*b*) 6 inputs after excluding bit tooth, bearing wear, formation abrasiveness and drillability. Even without any parameter related to the drilled formation or to bit status, the use of neural networks resulted in a good prediction capability for the ROP values (see Table 3.2). Other studies predicted the ROP also without any parameter related to formation (ESKANDARIAN; BAHRAMI; KAZEMI, 2017; BATAEE; IRAWAN; KAMYAB, 2014)

Many examples of the second approach can be found (DIAZ et al., 2018; ANSARI; HOSSEINI; AMIRPOUR, 2017; BEZMINABADI et al., 2017; HEGDE et al., 2017; AMAR; IBRAHIM, 2012; ARABJAMALOEI; SHADIZADEH, 2011). In this approach, some parameters related to formation properties or lithology are used as inputs to predict ROP values.

Some studies emphasized the importance of the formation on ROP prediction (ANEMANGELY et al., 2018; BASARIR; TUTLUOGLU; KARPUZ, 2014)². In this approach, a combination of surface drilling data and petrophysical logs provides accurate ROP-models.

The fourth approach is based on splitting the drilling data for each formation. Gandelman (2012) built neural networks to predict ROP with high accuracy by having artificial neural networks for each lithology. In this case, there is a unique ROP-model for each lithology. A similar approach was employed by Hegde et al. (2017), but the authors employed random forests to predict the ROP instead of neural networks. Another difference between both studies is the data partition. While Gandelman (2012) randomly partitioned the drilling data for each formation into training and testing sets, Hegde et al. (2017) used the initial phase of drilling to train the

² This article was added in this review after submitting the preliminary conference paper. Therefore this is not include in Table 3.2.

ROP model. The rest of drilling data of each section was used to validate the ROP-models; this process is detailed in following Section 3.2.5.

3.2.5 Data Partition

A common approach of those works from the Table 3.2 was to split the drilling data into two data set: one for training (when required, part of this training set was used in the validation process during the training) and another for testing, which assesses the generalization capability of the predictive model. This data partition is randomly carried out, and can be applied as post-analysis or prior to drilling a well with similar conditions of those used to build the ROP models.

Hegde et al. (2017) proposed another way to partition the data into training and testing data set, so that it can be used in a real-time environment. The trained model can be employed to optimize the drilling activities by finding out the optimum values of the controllable drilling variables, such as weight on bit and bit rotation speed, which maximize the ROP (HEGDE; GRAY, 2017).

The proposed method by Hegde et al. (2017) is based on splitting the drilling data into several data sets for each lithology. In the initial phase of a formation or lithology section, the drilling data are used to build the ROP model, i.e., the training data set is the initial data of this section. After training an expert model based on Random Forests Trees for this respective formation, this model is able to predict the ROP for the rest of the length of this section with better accuracy than the traditional ROP-models. When a new formation is encountered, then a new model is trained until it reaches a good prediction capability for the section ahead, and this process goes on. When using Random Forests, it is possible to predict the generalization error of the model by using out-of-bag error prediction.

Recently, a detailed study about this approach was published (SOARES; GRAY, 2018). Except for these trials, originated in the same university, we found no works that tried to build the ROP model while the drilling data are generated and sent to the rig crew. An illustration of this process is given in the Figure 3.6. In the current work, this process was named after the university where the studies were developed.

3.2.6 Drilling Data

In preliminary analysis, it was possible to observe that most of works used drilling data from two sources: drilling daily reports (ANSARI; HOSSEINI; AMIRPOUR, 2017; ESKANDARIAN; BAHRAMI; KAZEMI, 2017) or real-time drilling data (DIAZ et al., 2018; HEGDE et al., 2017; GANDELMAN, 2012).

The drilling data from daily reports are important source of information. However, the downside when using daily reports is that only few observations are available to build predictive models based on machine learning. For example, if it is required a data set with more than

Figure 3.6 – Data Partition developed at UT Austin

	Formation	ROP	WOB RPM	FLOW Pump Pres.	Data Partition after Hegde	
Depth	A				Training	Expert model for formation A
					Testing	
	B				Training	Expert model for formation B
					Testing	
	C				Training	Expert model for formation C
					Testing	

Source: after Hegde et al. (2017)

hundred points, it will not always be possible to analyze drilling daily reports from an individual well. While Diaz et al. (2018) used real-time drilling data from a drilled well with length of 4.6 km in South Korea with a total of 7043 observations, Ansari, Hosseini and Amirpour (2017) used drilling daily reports and needed to gather information from 19 wells drilled in Persian Gulf in order to have available 248 points. The difference in the amount of observations has a direct impact on training the predictive ROP model.

The real-time drilling data (RTDD) can be stored in time domain or in depth domain. When using drilling data in time domain, the raw recording can also have data during non-productive time. As consequence, manipulations are required to identify when the hole was being drilled prior to assessing how the drilling variables affect the drilling rate, what is the main goal of obtaining a ROP model. To properly identify the main drilling activities (e.g., rotary drilling, sliding, tripping connection), some works (MATHIS et al., 2007; TAVARES, 2006) reported so-called automated operation recognition systems.

One of the few works that used RTDD in time domain to predict drilling parameters was carried out by Fruhwirth, Thonhauser and Mathis (2006). However, the authors estimated the pump pressure instead ROP. In their work, there was available only information about data partition regarding the procedures for training, validation and testing; there was no mention about how different drilling operations were recognized. Gandelman (2012) reported also the use of RTDD in time domain, in his study of ROP prediction and optimization. The author detailed the data pre-treatment process employed in order to validate the drill-curves, but he did not mention about the existence of non-productive time in the original data set. Some sampling rates of drill curves employed in the previous works were 1s (FRUHWIRTH; THONHAUSER; MATHIS, 2006), 5s (DONNE, 2017), or 15s – 30s (GANDELMAN, 2012).

The depth domain is the common way of geologists to plot the logs against the depth. Using the depth as index simplifies the data preparation, since only observations while drilling

are used in modeling. Some sampling rates found of RTDD when stored in depth domain were 0.25 ft (HEGDE et al., 2017; HEGDE; GRAY, 2017) or 0.5 ft (NASCIMENTO et al., 2015b).

3.2.7 Handling Measurement Errors

Some researches (OTALVORA et al., 2016; ARNAOUT et al., 2013) reported some common measurement errors in drilling data, and proposed methods to determine quality indexes in real-time of streaming data. For our purpose, the best description of the real-time measurement problems was given by Arnaout et al. (2013). The authors classified the measurement errors in the three categories:

- Time problems: missing timestamp, invalid time format, wrong time zone, incorrect or no time synchronization;
- Depth Problems: bit depth/hole depth resets, heave compensation (floating rigs);
- Data Channel Problems: wrong channel description, wrong units, calibration, gaps (missing values and null values), different frequencies, outliers and drifting values.

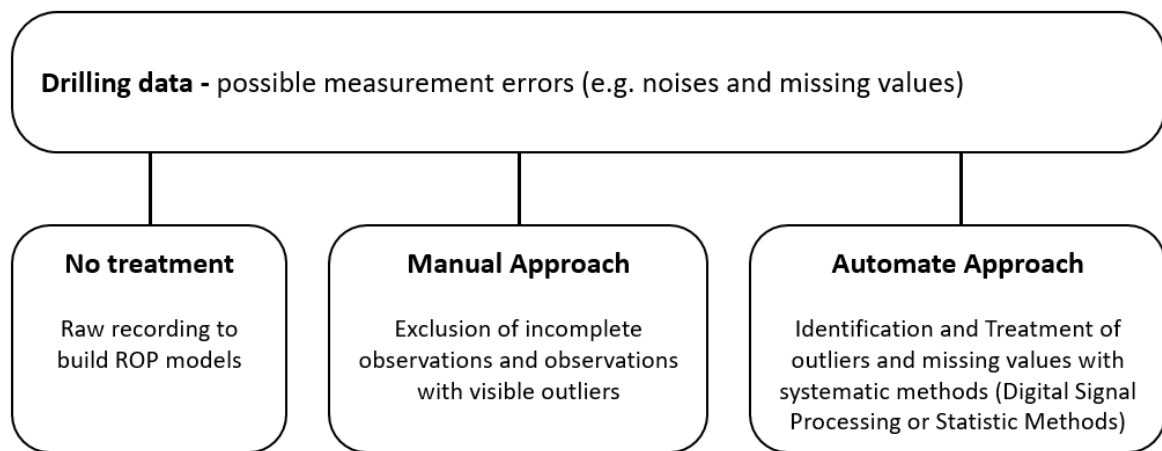
Most of works drew their attention to two types of problems from the category “Data Channel Problems” which are: outliers³ and gaps. To treat the outliers, it is required first to identify them, and then to apply a suitable treatment for them (i.e. replace the identified outliers by a suitable value or not consider the whole observation with an outlier). The problem of the gaps (missing values or null values) can be very challenging to be treated, because the original values is unknown and any attempt to treat or impute the data has high risk (LITTLE; RUBIN, 2002).

Noises (i.e. error in data), or considering a broader concept of outliers (i.e. discordant), make difficult the task of obtaining machine learning models, and increases the training time (GARCÍA; CARVALHO; LORENA, 2013; QUINLAN, 1986). Therefore, the reduction of noises in data brings together benefits to the learning process of data-driven models (GARCÍA; CARVALHO; LORENA, 2013). There are many methods to identify and treat both outliers and noises (SALGADO et al., 2016; AGGARWAL, 2013). Some different approaches to treat measurement errors in drilling datasets were found in the literature, and a summary of them is illustrated in Figure 3.7.

At this point of this review, it was found that only some works mentioned how outliers were treated. This type of error can be handled: with a manual approach (HEGDE; GRAY, 2017; GANDELMAN, 2012; ARABJAMALOEI; SHADIZADEH, 2011), or with filter to smooth

³ Sometimes, the terms noises and outliers can be confusing. The readers can refer to Salgado et al. (2016) in order to understand the difference between them. To sum up, noises are mislabeled examples or errors in the values of attributes. Outlier is a broader concept, because this term includes errors and discordant data (also called *abnormalities*, *discordants*, *deviants* and *anomalies*). Such discordant data is not necessarily an error in the measurement, but a deviation from a population.

Figure 3.7 – Possible approaches to treat measurement errors in drilling data that can be applied to drilling data analysis



Source: own authorship.

the recording, eliminating noises (DIAZ et al., 2018; ANEMANGELY et al., 2018). A manual approach is subject to human interpretation, and can be very time-consuming job, especially when analyzing many drill curves. However, it can be suitable when analyzing a small dataset, because those observations which are visible outliers can be simply removed from the data set, as done by Hegde and Gray (2017).

Instead of the manual approach, the automate approach employs robust techniques to identify whether a observation is indeed an outlier or not. Few works applied filter to smooth drill recording, reducing noises in data, as consequence some outliers. Some filters employed were low-pass parabolic filter (DIAZ et al., 2018), and Savitzky-Golay filter (ANEMANGELY et al., 2018). Robust techniques to identify mainly outliers, for example median absolute deviation, were not used in the ROP studies.

Another important issue is the completeness of drilling data. It is not rare to encounter drilling data with “gaps” in their recordings. A gap is when a failure in data transmission occurs, so that measurements of one or more drilling variables are not transmitted for a period of time (ARNAOUT et al., 2013). Having this in mind, it is not difficulty to suppose that some of 58 studies had faced this type of problem when training a ROP model. However the common approach was to omit this problem.

One exception was the Gandelman’s study, who mentioned the problem of missing data. His approach was to use complete-cases. The downside of this approach is to cause a substantial loss of information (LITTLE; RUBIN, 2002). Gandelman (2012) detailed the data pre-treatment process employed to validate the drill recordings, eliminating possible error measurements. One rule employed by the author to validate the drill curves was the following: ROP could only be above zero, if, and only if, the WOB and RPM were both above zero. Otherwise, this observation would not be considered as a valid one. The author mentioned the reduction of dataset from 43 524 observations (200 hours of drilling activities) to 23 949 valid observations (160 hours).

3.3 DRILLING OPTIMIZATION BASED ON PREDICTIVE MODELS

ROP models can be employed in drilling optimization. Two approaches were employed in the previous works. The first has the goal to determine the values of the controllable drilling parameters that resulted, usually, in the maximum value of the ROP. This approach led to a single objective optimization problem. A more sophisticated method takes in account that maximizing the ROP may induce drilling dysfunctions. To avoid or mitigate the drilling problems, some works formulated a multi-objective optimization problem. The following section covers both approaches.

3.3.1 Single-Objective Optimization

Attempts to formal optimize drilling variables (WOB, RPM and bit hydraulics) are not new. One example is the work from Tansev (1975), where multiple logarithmic regression of field data were employed to model the ROP and bit life. With both prediction models, the controllable variables could be optimized in order to minimize the cost per foot, subject to controllable variables' bounds. In the last decade, some works employed predictive models to formulate optimization problems. A summary about these works is given in the Table 3.3.

Table 3.3 – The use ROP models to optimize the drilling variables, considering a single-objective problem

Author	Obj. Function	Constrained?	Opt. Method	Opt. Parameters
Bahari and Seyed (2009) ^(a)	min. Cost	✓	trust region	WOB, RPM
Arabjamaloei and Shadizadeh (2011)	max. ROP	✓	genetic algorithm	^(b) WOB, RPM, bit hydraulic
Awotunde and Mutasiem (2014) ^(c)	max ROP min. Total Time	✓	differential evolution	WOB, RPM, mud flow rate
Bataee, Irawan and Kamyab (2014)	max. ROP	✓	not mentioned	WOB, RPM, mud weight
Hegde and Gray (2017)	max. ROP	✓	brute force	WOB, RPM, mud flow rate
Hegde and Gray (2018) ^(c)	max ROP min. Torque min. SE	✓	particle swarm (PSO)	WOB, RPM, mud flow rate

^(a) considering the optimization problem of operating conditions WOB and RPM. The authors studied also the optimization of hydraulics and bit tooth wear. ^(b) Only added those variables whose optimization were detailed. ^(c) These studies investigated different objective function for the single-objective optimization problem.
Source: own authorship

Bahari and Seyed (2009) studied the optimization the controllable drilling variables (WOB

and RPM) by minimizing the cost per foot drilled, subject to operating ranges recommended by manufacturing companies and limited to the maximum applicable mechanical energy. The basis of his method is the Bourgoyne and Young model to predict the ROP. The single-objective problem was solved by constrained optimization algorithm called trust region (COLEMAN; LI, 1996).

Arabjamaloei and Shadizadeh (2011) employed genetic algorithm in order to find which values of the controllable variables maximize the ROP. Bataee, Irawan and Kamyab (2014) studied also the optimization of ROP, but the authors did not mention which optimization method was employed.

Awotunde and Mutasiem (2014) employed the Warren model to model the ROP. The authors compared two objectives functions in single-optimization framework: maximization of ROP and minimization of total-time (drilling time, tripping and bit-change time). At shallow depths, the maximization of ROP yields to lowest total time, but, at deeper depths, the minimization of total time yields to the lowest overall time. As global optimizer, the differential evolution (STORN; PRICE, 1997) was employed.

Hegde and Gray (2017) performed the so-called “brute-force” algorithm to optimize the drilling performance indicators (e.g. rate of penetration). In this study, a single objective optimization was formulated, whose goal was to maximize the ROP subject to constraints from two sources. According to the authors, the magnitude of drilling variables are limited due to design restrictions. The other limitation comes from the data. That is, it is not reliable to predict ROP values when the inputs are out of the range from the data set used in training phase. This extrapolation leads to high uncertainty around the ROP estimation. As already mentioned, this method could be employed in a real-time environment, because the ROP model could be trained while the drilling takes place.

Later, Hegde and Gray (2018) analyzed other objective functions: minimization of torque, minimization of the specific energy, or maximization of ROP. The authors concluded that the minimizing the specific energy was the better approach. In this study, a particle swarm optimization (PSO), able to deal with constraints, was employed.

3.3.2 Multi-Objective Optimization

Not many works proposed a multi-objective optimization problem for the drilling activities, expect for the following works Gandelman (2012), Guria, Goli and Pathak (2014), Payette et al. (2017).

Gandelman (2012) developed neural networks for different Brazilian rock formations (12 in total), to be used latter in real-time optimization of mechanical drilling variables (WOB and RPM). His optimization approach considered two objective functions at the same time: $\min SE$ and $\min E$, where SE represented the mechanical specific energy, and E a function error between the predicted $ROP_{predict} = f(WOB, RPM)$ and a desired ROP_{set} , which the driller set. The goal of this optimization was determine the optimum WOB and RPM that achieved a

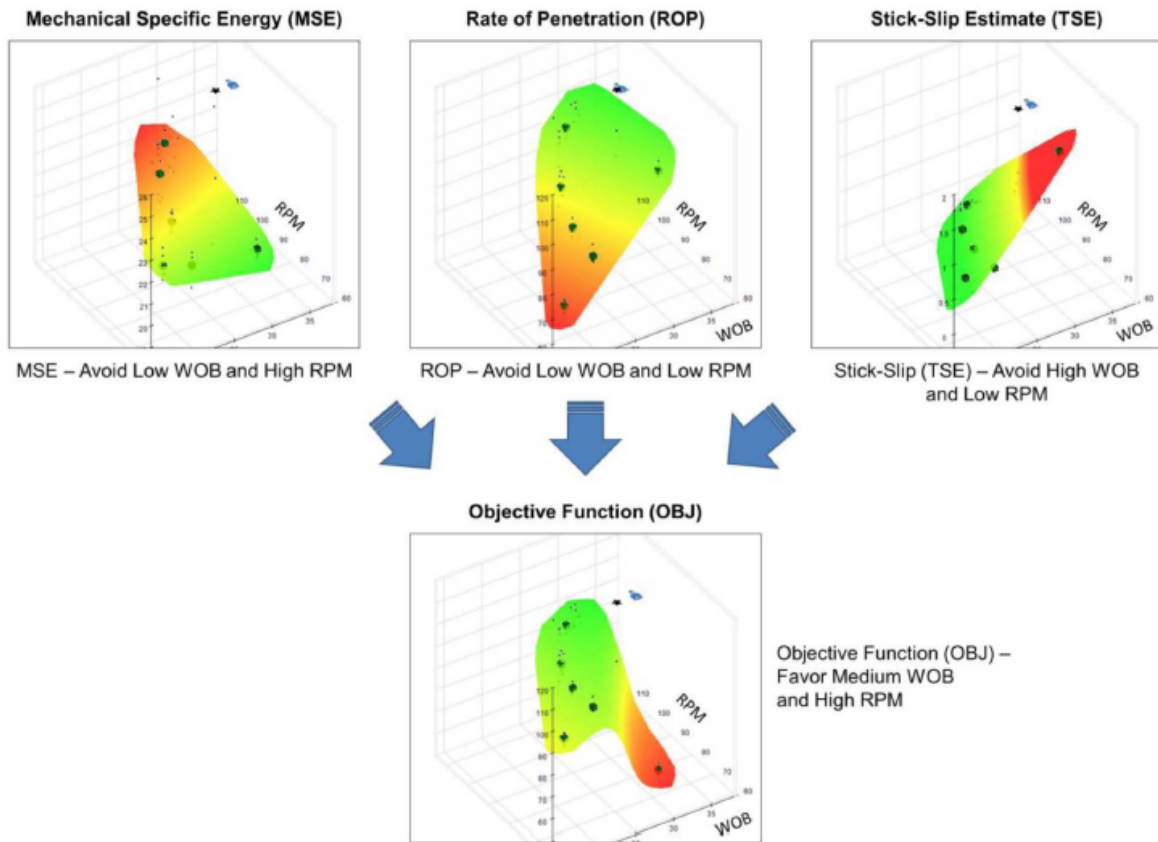
desired ROP, trying, at the same time, to spend as less as possible energy in the process. This method could be classified as ε -constraint optimization, where the goal was to minimize the SE subject to a constraint $(ROP_{predict} - ROP_{set}) \leq \varepsilon$, where the decision variables $x \in \Omega$ were within the design limits of drill equipment.

Among other things, Gandelman (2012) tested two optimization methods: particle swarm optimization, and an own developed exhaustive search. This last approach combined *if-then* rules with the grid search in the feasible region Ω of the decision variables. The author selected the bit weight and rotary speed as the variables to be optimized. For each iteration, 195 000 combinations of WOB and RPM were tested. The author concluded that this method was the only one able to determine the optimum combination of WOB and RPM.

Guria, Goli and Pathak (2014) reported another work of drilling optimization with multiple objectives. The authors employed the Bourgoyne and Young formulation for ROP prediction and tooth wear. With these predictive models, a multi-objective optimization involving conflicting objectives was developed. The objective functions were: (i) maximization of the drilling depth, (ii) minimization of the drilling time, (iii) minimization of the drilling cost. The controllable drilling-variables used in this optimization study were four: equivalent circulation mud density, drill bit rotation speed, weight on bit, Reynolds number of circulating mud through drill bit nozzles. The authors employed the non-dominated sorting genetic algorithm (NSGA-II), as optimization technique.

In the drilling advisory presented by Payette et al. (2017), no ROP model was actually obtained, so that this work actually goes beyond the scope of this review. However, it is worth of mentioning the adopted strategy to simplify the multi-objective optimization problem. Up to three different objectives (ROP, SE and stick-slip risk) are aggregated into a single scalar function. It facilitates the optimization task. The Figure 3.8 illustrates the aggregation process of different objective functions into a single one.

Figure 3.8 – Simplifying the multicriteria optimization into a single objective functions. Response surfaces form the basis for the objective function (OBJ).



$$OBJ(MSE, SS, ROP) = \frac{\delta + ROP / ROP_o}{\delta + MSE / MSE_o + SS / SS_o},$$

(δ factor to be determined),

Source: adapted from Payette et al. (2017)

3.3.3 The Need of Changing the Current Mindset

Wang and Salehi (2015), Hegde and Gray (2017) mentioned the need of changing mindset to optimize the drilling process. The term *set-it-and-forget-it* indicates a typical approach to determine the values of controllable parameters while drilling. However, if controllable drilling variables are not changed during the drilling process, it is not possible to determine whether the process is at its optimum point or not.

Determining the optimum parameters can be facilitated by more often implementation of tests, such as *drill-off* or *drill-rate* tests (DUPRIEST; KOEDERITZ, 2005; NASCIMENTO et al., 2016; PAYETTE et al., 2017). Nascimento et al. (2016) proposed a method to plot ROP against WOB curves while not only performing pre-operational test, but also during normal drilling activities.

Payette et al. (2017) showed that through drilling advisory systems it is possible to optimize

the drilling process. These systems encourage the rig crew to change the controllable parameters within an operational range. These systems provide information about how each variable affects the ROP, as well as nonlinear effects on ROP.

Despite the efforts employed in collecting huge amount of drilling data, comprehensive analysis does not take part as the best practices of drilling engineering, even knowing that simple plan of activities can boost the results (STAVELEY; THOW, 2010). It is possible to find discussion that reports wastes of 60% in the drilling engineering, what is unacceptable in most of industries (BOND et al., 1998). One way of boosting the efficiency is to avoid or, at least, mitigate drilling dysfunctions that lead to non-productive time.

One technique from artificial intelligence, called case-based reasoning, can have a great impact in detecting drilling problems, and proposing the respective solutions by combining data analysis of historical and real-time data. This approach enables the identification and mitigation of events that lead to non-productive time (SKALLE; AAMODT; ERIKGUNDERSEN, 2013). For that, knowledge is extracted from past events provided from data fusion from historical data and rig crew experience. So it is possible to identify the most likely causes of the drilling problems as well as the respective solutions. This process resembles the method employed by rig crew, facilitating the knowledge transfer. A detailed review of this method for drilling optimization is given by Shokouhi, Skalle and Aamodt (2014). Yuan et al. (2009) showed also the possibility to use case-based reasoning to optimize the drilling parameters, such as the optimum drill bit and what values of drilling variables maximize the ROP.

3.4 SUMMARY

Regarding the ROP models, we observed the following issues:

- As seen in the Table 3.2, the works that compared different methods to predict the ROP concluded that the machine learning methods outperformed other methods (regression or traditional models) in this type of task;
- Machine learning algorithms are normally black-boxes models, but, depending on the algorithm employed, it is possible to extract rules that represents the relationship among the drilling variables. These rules can be help the drill crew to select the optimum drilling variables;
- Valuable information can be obtained with sensitive analyses as well. With such analysis, partial plots (e.g. ROP vs. WOB) can provide graphs that would help identifying the drilling conditions, as done by Eskandarian, Bahrami and Kazemi (2017) (See Figure 3.3);
- Most of works carried out historical analysis for wells drilled in a similar region, what can be employed as post-analyses or prior to drilling a similar well. However, few tries obtained ROP-models while the drilling (SOARES; GRAY, 2018; HEGDE et al., 2017; HEGDE; GRAY, 2017);

- It was also seen that not many works tried to formally treat possible measurement errors in the dataset.

Regarding the optimization based on data-driven models, it was possible to observe:

- maximizing ROP is a common approach adopted. However, this is not always the best approach, because drilling dysfunctions can occur at higher drilling rate. Therefore, other metrics need to be taken into account, yielding a multi-objective optimization problem;
- A multi-objective optimization seems to be a more reasonable approach in the drilling optimization, due to the complexity of drilling process;
- Applying decision making techniques for multiple-objective functions (e.g. global criteria method, ε -constraint method) enables a easier implementation of a multi-objective optimization problem.

4 METHODS

This chapter covers three parts. It is presented, first, the material used in the development of this current work. Then, we describe the method employed to develop predicting models for drilling performance indicators, such as rate of penetration and specific energy, covering also the data pre-treatment employed in the current work. The third part covers how the obtained models can be used to optimize the drilling controllable variable such as weight on bit and rotating drillbit speed.

4.1 MATERIALS

Drilling data from pre-salt region used by Nascimento (2016) were employed in the current work. In order to have the results reproducible by other researchers, we employed an open dataset from Norwegian wells, which was published by Donne (2017). The describing of both datasets are given in next sections.

4.1.1 Drilling Data from Pre-Salt

Drilling data from pre-salt layer were used for the development and test of the methods to be developed in this current work. These data were studied previously by Nascimento (2016), Dutra (2016). The readers can refer to these works in order to obtain details about the geological characteristics of drilling data. In this present work, we used drilling data of four wells drilled in the carbonate formation of pre-salt (a total of 1 212 *km*). These drill-curves have as index the depth, recorded at the sampling rate of 0.1524 *m* (0.5 *ft*). In Table 4.1, a summary of those drilling data is given.

In the original recording of the drilling data from pre-salt formation, missing values (or null values) were observed, especially for those parameters measured by downhole equipment, such as collar rotating speed or vibration. Prior to performing drilling data analysis, a driller engineering can select those variables that he believes to be the most important in ROP analysis, and then exclude the other parameters. This approach can reduce the amount of missing values, helping the further data cleansing process. However, it hinders more elaborate analyses such as feature selection in order to identify the most important parameters in ROP predictive model, as explained in the following paragraphs.

Table 4.1 – Overview of the drilling data from pre-salt formation. The abbreviation meanings are found in Annex A

Well	Drilling Variables		Hole Depth (m)		Points
	#	description	initial	final	#
A	14	DEPT, DHAP, DHAT, ECD, GR_CAL, ROP5, CRPM, TVDE, BLKP, HKLA, RPM, SPPA, SWOB, TFLO	3 999	4 194	1 326
B	24	DEPT, BLKP, PVEL, ROP, DWOB, SWOB, HKLA, SHKR, SHKRSK, SHOCKS_ACC, DTOR, CRPM, VIB_LAT, RPM, STICK, TRPM, TFLO, SPPA, VIB_X, SHKL, ROP5, STICKRATIO, STUCK, ECD	3 422	3 507	509
D	26	DEPT, BLKA, BLKP, CDEPTH, HKLA, ROP5, SPPA, SWOB, DHAP, DHAT, ECD, GR, SHKL, AJAM, CRPM, SHKRSK, SHOCKS_ACC, STICK, STUCK, TRPM, VIB_LAT, VIB_X,	4 403	4 546	927
H	15	DEPT, DHAP, DHAT, ECD, GR_CAL, SHKL, ROP5, RPM, SPPA, STOR, SWOB, TFLO, CRPM, SHKRSK, TRPM	5 050	5 840	5 184

Source: dataset from Nascimento (2016).

Some statistics about the missing values in the drilling data from Well H are given in the Tables 4.2-4.3. Many variables have up to 20% missing values, and trying to use only complete cases would result in data set with roughly 30% of the original information. This approach of analysis which considers only complete-case data would causes substantially loss of information, especially if we observe that about the half of observations have only 1 to 3 missing values (see Tables 4.3- 4.4).

As explained by Little and Rubin (2002), the “complete-case analysis may be justified in terms of simplicity when the loss of precision and the bias is minimal, so that the pay-off of exploiting the information in the incomplete cases will be minimal”. Therefore, this simple approach of discarding incomplete observations is not appropriate in the analysis of this dataset from pre-salt formation. Sophisticated techniques from statistics field, such as likelihood-based approached to the analysis of incomplete data Little and Rubin (2002), or from Digital Signal Processing (DSP) science can be applied to treat the stored data before building the ROP predictive models. In the current work, the strategy to treat measurement errors is presented in the Section 4.2.1.

Table 4.2 – Amount of missing values in each variable (column). For illustration purpose, statistics only from the Well H is presented. This dataset has a total of 5 184 observations and 15 variables, therefore the total amount of values is $5\,184 \cdot 15 = 77\,760$.

Variable	Amount of missing values	
	#	%
DEPT	0	0
DHAP	782	15.08
DHAT	1 327	25.60
ECD	813	15.68
GR_CAL	571	11.01
SHKL	1 297	25.02
ROP5	58	1.19
RPM	31	0.60
SPPA	21	0.41
STOR	31	0.60
SWOB	31	0.60
TFLO	44	0.85
CRPM	1 741	33.58
SHKRSK	1 746	33.68
TRPM	1 721	33.20
Total	10 214	13.14

Source: own authorship.

Table 4.3 – Amount of observation with k-variables as missing values – Well H

k	#	%
0	1 597	30.81
1	999	19.27
2	723	13.95
3	831	16.03
4	456	8.80
5	278	5.36
6	137	2.64
7	87	1.68
8	56	1.08
9	17	0.33
10	3	0.06

Source: own authorship.

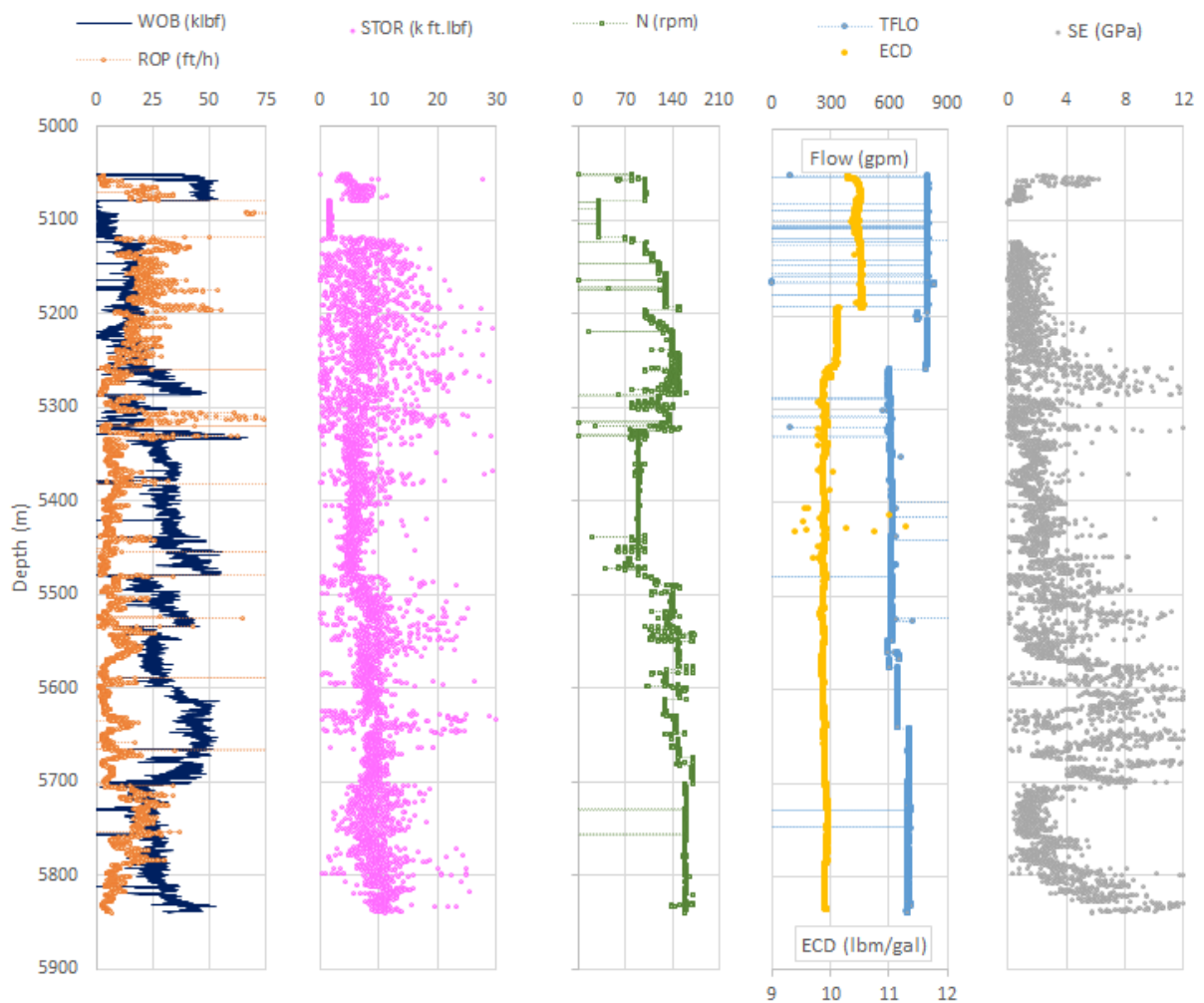
Table 4.4 – Relative amount of observations with k-variables as missing values

k	Well A	Well B	Well D	Well H
	%	%	%	%
0	37.86	23.73	80.69	30.81
1 – 3	51.66	47.25	13.38	49.25
≥ 4	10.48	29.02	5.93	19.95

Source: own authorship.

The Figure 4.1 illustrates one of four wells drilled in pre-salt region, with raw recordings. The drill crew experienced low rates of penetration due to the hardness and abrasiveness of carbonate formation, requiring recurrent drill bit changes (NASCIMENTO, 2016; DUTRA, 2016). To calculate the specific energy, the original formulation from Taele was employed (equation 2.1). We added other drill-curves from pre-salt in Appendix B.

Figure 4.1 – Raw recording of drill curves from Well H. The specific energy was calculated based on Taele's formulation.



Source: own authorship.

4.1.2 Drilling Data from Norway

Donne (2017) published drill curves from Norway that he used in his Master's Thesis. The drilling data are anonymous, i.e. the file names were randomly given by Donne, preserving the confidentiality of the data. The use of this dataset enables the readers to reproduce the proposed methods in this current work.

This dataset contains real-time drilling data (RTDD) of 10 wells drilled in an Norwegian continental shelf field, called Statfjord (EQUINOR, 2018). Donne selected 10 wells from more 100 wells drilled available in their data-bank. The criterion for this selection was the data quality. These data are in time domain base, recorded at the sampling rate of 5 s. An overview of this dataset is given in the Table 4.5. The selected drill curves for the development of the current work are, in general, free of the missing data problem.

Considering the amount of variables recorded for each well drilled in the RTDD in the original files, we can classify them into three different groups, as shown in the Table 4.5. In present work, we used drilling data of all wells, except the drill-curves from the well 2 which has long period of missing values (over 50 minutes), as well as calibration problems. Those types of errors go beyond the scope of this work and therefore the well 2 was not considered in this work. We found wrong channel descriptions in the data of the well 6, however it was possible to recover the data, as explained in Appendix A. The drill curves from the third group have actually 36 variables, however nine of them contain only not-a-number values, being therefore not taken into account.

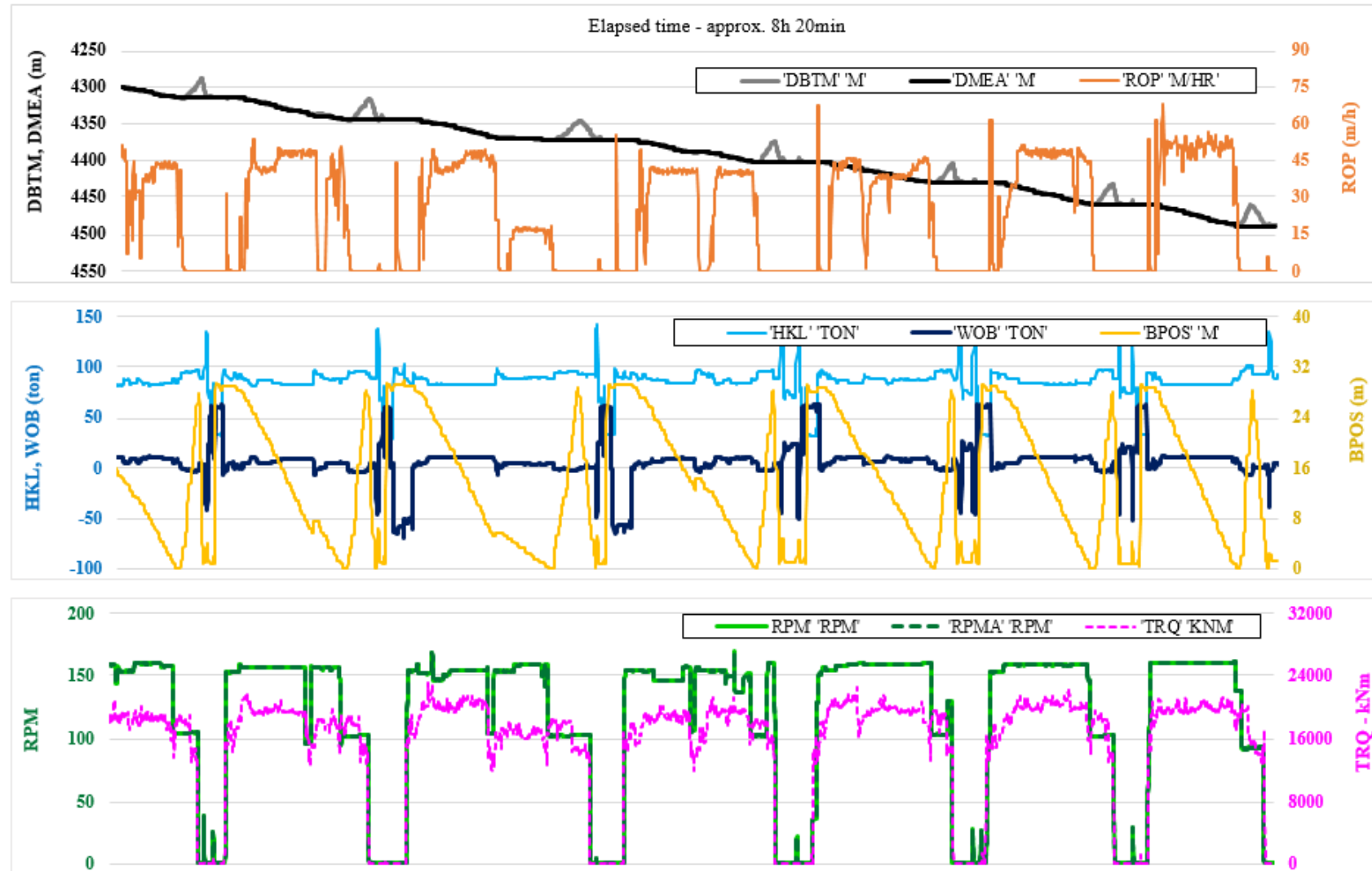
Unlike pre-salt drill curves, the drilling data from Norway were recorded in time domain base. A sample of those drilling data is given in the Figure 4.2. We observed an continuous measurement of all variables for all wells. That is, all drill curves from Norway have in their recordings all possible drilling operations, including tripping connection, rotary drilling, reaming and so on (see Figure 4.2). For this reason, it is necessary to employ operation recognition techniques (MATHIS et al., 2007) to identify when a hole was being drilled in fact.

Table 4.5 – Real-time drilling data from Norway published by Donne (2017). The abbreviation meaning used in these dataset can be found in Annex B

Group	Drilling Variables		Well			Hole depth (m)		Points	Selected?
	#	description	#	Formation	Encountered Lithology	Initial	Final	#	
1st	8	Time, DBTM, DMEA, ROP, WOB, RPM, BPOS	1	Viking, Brent	marlstones	2 553	2 637	5 973	✓
			2	Brent	claystone, coal, claystones	3 597	3 770	14 992	✗
2nd	10	Time, DBTM, DMEA, ROP, WOB, HKL, BPOS, TRQ, RPMB, RPMA	3	Cromer, Viking, Brent, Dunlin	calcitucite, claystones	3 100	3 420	10 933	✓
			4	Viking, Brent	coal, limestone, silstone, sandstones, claystones	4 300	4 800	13 659	✓
			5	Not Available	claystones, marlstones	2 500	3 000	20 210	✓
			6	Viking, Brent, Dunlin, Statfjord	claystones, sandstones	2 700	3 100	14 737	✓
			7	Viking, Brent, Dunlin, Statfjord	claystones, silstones, limestones, sandstones	3 002	3 640	32 837	✓
3rd	27	2nd Group + SPP, CEPP, KLP, CHP, CFI, CTVL, TVA, TPVT, MFO, MFO, MFI, MDO, MDI, MTO, MTI, ECDB, ECDM, GAS	8	Brent, Dunlin	sandstones, shale, silstones	2 600	2 800	17 107	✓
			9	Brent	coal, claystones, calcitucite, silstones	4 338	4 700	5 013	✓
			10	Viking, Brent, Dunlin	coal, silstones, claystones	2 600	2 900	14 417	✓

Source: dataset available in the work of Donne (2017)

Figure 4.2 – A sample of real-time drilling data considering the first approx. 8 hours of available recording — Well 4 from Norway



Source: dataset available in the work of Donne (2017)

4.1.3 Software and Hardware

The current work employed MATLAB® software, version R2017b, in order build the predictive models and to use them to optimize the drilling variables. The main packages and toolboxes employed were: Global Optimization Toolbox™, Statistics and Machine Learning Toolbox™, Parallel Computing Toolbox™, and Signal Processing Toolbox™. The simulations were performed in a simple notebook with Intel® Core™ i5-4210U CPU 1.70 GHz, memory-RAM of 4 GB, and operating at 64 bits.

4.2 PREDICTION DRILLING PERFORMANCE-INDICATORS

Here, we present the adopted training method to obtain predictive models of the drilling performance indicators. This workflow can be employed to predict the rate of penetration or specific energy. With some adjustments, this method can also be applied to predict any other drilling parameters.

4.2.1 Data Pre-Treatment

Measurement errors can occur in drilling data provided from down-hole and even surface equipment. These errors can lead e.g. to observations with noisy data and missing values. As not many works treated formally these problems, as seen in the Chapter 3, a systematic approach for data pre-treatment is presented in this section (see Figure 4.3). The proposed approach is able to treat both problem in drilling data, outliers and missing values, regardless of domain used to record the drilling data. It can be applied to both common drilling data indexes: depth basis domain (pre-salt recordings) and time basis domain (Norwegian recordings).

A simple assumption of independent data channels was adopted. This enables to apply some common approaches to treat outliers and missing values in studies on energy forecasting (GONZÁLEZ-ORDIANO et al., 2017; GUAN et al., 2013).

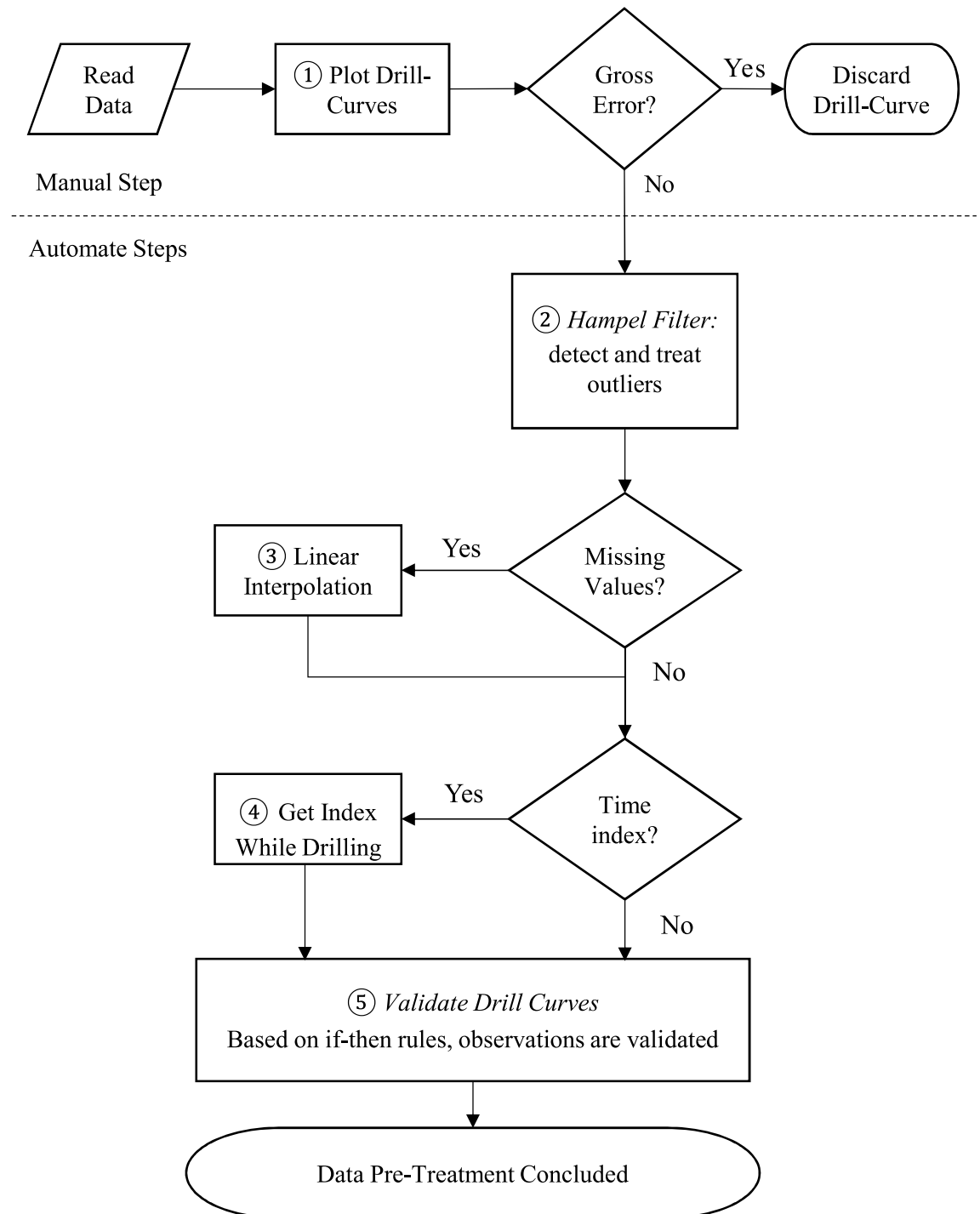
① *Plot Drill-Curves*

The data pre-treatment process starts with a manual step, where the drill-curves are plotted. This enables to identify if there are gross errors in the drilling data. If there are such errors, any attempt to perform data analysis will be prejudiced. For this reason, the drilling data from the Well 2 was not considered in this study (see Section 4.1.2).

In additional, cross-analysis with drill reports can be carried out. This enables to extract important information about: the events occurred during drilling, drill bits employed, bottom-hole-assembly, drilling dysfunctions. Gandelman (2012) carefully proceeded a pre-treatment analysis of data. In his work, a special attention was given to exclude from drilling data those observations if the drill bit had signs of great wear.

In the current work, such cross-analysis could be partially performed. For drilling data from pre-salt, information about drill bits and some important drill events were extracted from

Figure 4.3 – Data Pre-Treatment Method



Source: own authorship.

Nascimento (2016). For drilling data from Norway, information about drill bits are not available in the original work from Donne (2017). However, the author provided details about the geological formation.

② *Hampel Filter*

The first step to treating outliers is having a robust method to find them. Leys et al. (2013) showed that the use of mean plus/minus a coefficient (normally three) standard deviation is very sensitive to outliers, while the use of Median Absolute Deviation (MAD) is a more robust measure of dispersion. Therefore, we chose the MAD in order to identify outliers. This method was popularized by Hampel (1974).

Let X_j be a vector, $\{x_1, x_2, \dots, x_j\}$, of j elements. The $M_j(X_j)$ represents the median of X_j . When j is odd, then $M_j(X_j)$ is the simply middle order statistic of X_j . When j is even, it is possible to use the average of the order statistics with ranks $(j/2)$ and $(j/2) + 1$ (ROUSSEEUW; CROUX, 1993).

The scaled Median Absolute Deviation (MAD) is defined as follows (ROUSSEEUW; CROUX, 1993; LEYS et al., 2013):

$$\begin{aligned} MAD &= b * M_i(|x_1 - M_j(X_j)|, |x_2 - M_j(X_j)|, \dots, |x_j - M_j(X_j)|) \\ MAD &= b * M_i(|x_i - M_j(X_j)|) \end{aligned} \quad (4.1)$$

where M_i is the median of the absolute difference between each element of x_i and $M_j(X_j)$ for $i = 1, 2, \dots, j$. When assuming the condition of normality, the MAD is scaled by a factor $b = 1.4826$ (LEYS et al., 2013).

The rejection criterion is median plus/minus a coefficient multiplied by the MAD, which can be expressed by (LEYS et al., 2013):

$$\begin{aligned} M - TH * MAD &< x_i < M + TH * MAD \\ \frac{(x_i - M)}{MAD} &> |\pm TH| \end{aligned} \quad (4.2)$$

where M is the median $M = M_j(X_j)$, MAD is the scaled absolute deviation around the median and TH is the coefficient to determine the lower and upper limit for the outlier detection. The value of TH must be defined, and this “remains the unavoidable subjective aspect of the decision” (LEYS et al., 2013, p. 765). It is recommended to use plus/minus $2.5 * MAD$ (moderately conservative), but other values can be applied such as $TH = 3$ (very conservative) or $TH = 2$ (poorly conservative) (LEYS et al., 2013).

A moving median absolute deviation can be calculated over a sliding window across the neighbours elements of x_i . For that, it is only necessary to suppose that the vector X_j is obtained by this sliding window within a vector X_n with n elements, where $j < n$. In Matlab, this method is implemented by the function `hampel` (MATHWORKS, 2018a). For each element x_i , the window is centered about the element in the current position. As shown in the Code 4.1, we can

set the amount of neighbors, $numNeighbors$, on each either side of the center element x_i , when using the function `hampel`. This implies that the window size is $j = 2k + 1$, i.e. it is always an odd number. The thresholds TH is another parameter that can be set. To determine the values of both parameters (number of neighbors and threshold), the influence of them was investigated in the Appendix C.

Code 4.1 – General use of `hampel` function, adapted from Mathworks (2018a)

```

1 % Inputs
2 %   x - a matrix with drill curves
3 %   numNeighbors - the number of neighbors on either side of each sample in ...
   the window
4 %   TH - threshold (standard deviations) for outlier detecting
5 %
6 % Outputs
7 %   y - matrix with filtered data
8 %   isOutlier - logical matrix of detected outliers
9 %   xmedian - local median for each observation
10 %   xsigma - MAD value for each observation
11 [y,isOutlier,xmedian,xsigma] = hampel(x, numNeighbors, TH)

```

When an outlier is detected, it is possible to replace the detected outlier by a suitable value. Here, the `hampel` function replaces the identified outlier by the center value given by the moving median (MATHWORKS, 2018a).

③ *Treating Missing Values*

In this step, the linear interpolation is employed to impute the missing values. In this case, it is adopted the assumption of data channels independence. In Matlab, the function `fillmissing` enables to implement several methods to impute missing values, including the linear interpolation (MATHWORKS, 2018b).

④ *Index While Drilling*

Examples of drilling activities are (THONHAUSER, 2004): bottom hole assembly (BHA) runs, tripping, making connection, ream and wash, circulating, drilling in rotating or sliding mode. If the drilling data is recorded in time domain, it is possible to have drill curves recorded with all possible drilling activities. Once, the current study aims to model how some drilling variables affect the performance indicators (ROP and SE), it is important to split drill curves with time domain into two main states: drilling and not-drilling states. Some works presented automated operation recognition methods, which provide important information for this task (MATHIS et al., 2007; TAVARES, 2006).

According to Mathis et al. (2007), the rotary drilling occurs if three conditions are satisfied: i) bit must be on bottom, ii) circulation must be present, and iii) drillstring is rotating. If drillstring is not rotating, then sliding drilling is occurring.

In order to implement the aforementioned operation recognition, it is necessary to extract some rules from drill curves that have time as index (see Figure 4.2). The bit can be considered on

bottom if the measured bit depth (DBTM) is equal to hole depth (DMEA). The measurement of drill flow rate enables to determine if the circulation is present (second condition). The recording of drillstring rotation indicates if the drillstring is rotating or not (third condition). The necessary conditions to identify the effective drilling operations are summarized in the Table 4.6. These rules are simple, and have not the purpose to identify some drilling situations, such as ream and wash or run in hole/out of hole. To overcome noisy data, thresholds ε for each condition are employed. A visual approach was employed to determine the values of these thresholds, which is presented in the Appendix D.

Table 4.6 – Rules to determine drilling modes for real-time drilling data in time domain

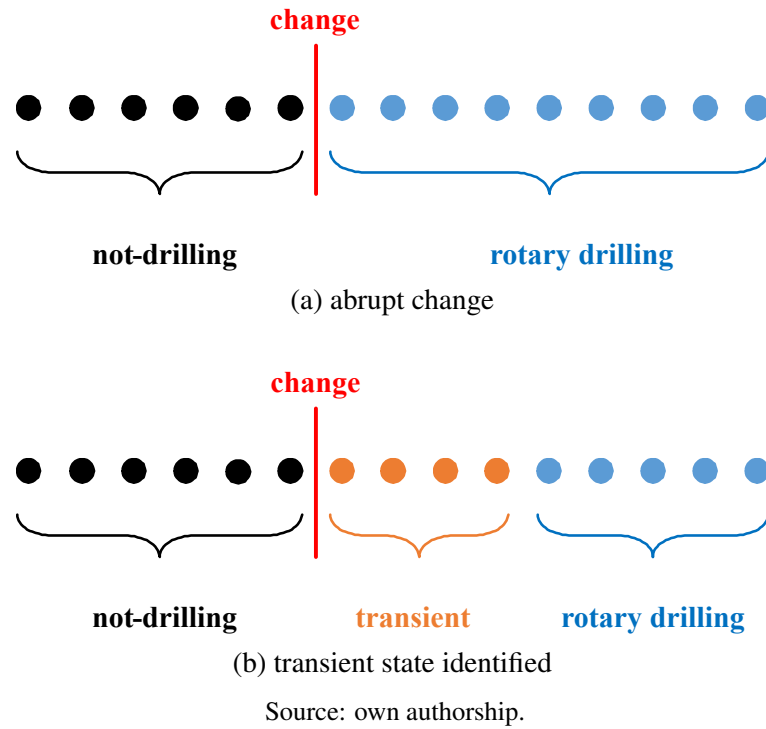
Mode	Necessary Conditions	
	According to Mathis et al. (2007)	Implementation
Rotary Drilling	i) bit on bottom	i) $DBTM - DTMEA < \varepsilon_1$
	ii) circulating present	ii) $Flow > \varepsilon_2$
	iii) drillstring rotating	iii) $RPM > \varepsilon_3$
Sliding Drilling	i) bit on bottom	i) $DBTM - DTMEA < \varepsilon_1$
	ii) circulating present	ii) $Flow > \varepsilon_2$
	iii) drillstring not rotating	iii) $RPM \leq \varepsilon_3$

Source: own authorship.

Based on the rules from the Table 4.6, the drilling data is split into three groups: rotary drilling, sliding drilling, and not-drilling. In addition to these three states, another important state is added, namely the “transient”. The transient state is a fine tuning of indexes obtained by the automated operation recognition. Two rules are employed to define a transient state. The first one considers as transient state those observations near to the operational limit for the block position. Let $BPOS$ denote the block position and ε_4 the limit for the block position. If the block position $BPOS_i$ of the i -th observation lies out of the interval $[\min(BPOS) + \varepsilon_4, \max(BOPS) - \varepsilon_4]$, then this observation is considered a “transient state”.

Another rule employed to recognize transient state is the following: if a change from not-drilling to drilling state is encountered, then a transient state is identified. A transient state from this change is only computed if the ROP value for the not-drilling state is bellow a threshold (5 m/h). When a change is confirmed, the drilling state for the next observations is considered as transient, as illustrated in Figure 4.4. The current work considers a period of 2 minutes for the transient state. The Appendix D details the transient state recognition, explaining some decision made.

Figure 4.4 – After a change from not-drilling to rotary drilling, the transient state is identified.



⑤ *Validate Drill Curves*

Even treating outliers and missing values, invalid observations may still be in recorded data. It is physically not possible to have ROP above zero when WOB is null (GANDELMAN, 2012). Therefore, invalid observations will be discarded in the final step of the data pre-treatment. Two rules are employed to check whether an observation is valid or not:

$$\begin{aligned} \text{First rule: if } ROP > 0 \text{ and } WOB \leq 0 \text{ then, the observation is invalid.} \\ \text{Second rule: if } WOB > \textit{limit} \text{ then, the observation is invalid.} \end{aligned} \quad (4.3)$$

Both rules are checked independently. The second rule avoids too high values of WOB. In the case of Norway, the WOB is considered too high if WOB is above 45 ton. For the drill curves from pre-salt region, the adopted threshold is 100 klbf; Gandelman (2012) mentioned a limit of 90 ton for the WOB. The Appendix E illustrates the impact of validation process on the resulting drill curves.

4.2.2 Training the Predictive Models

After the data pre-treatment process, we can train predictive models for the drilling performance parameters. For that, we employ a common approach of splitting at random the original dataset into two other datasets, not overlapping each other. One is called training dataset, and another called testing dataset. The predictive models for drilling performance

indicators (ROP or SE) are trained on the training dataset. Then, the generalization capability of the obtained model is assessed on the testing dataset.

We chose mainly random forests to train the predictive models, because of two reasons. The first one is their good capability to model real-world problems (BREIMAN, 2001a), as well as problems tailored to drilling engineering (HEGDE et al., 2017; HEGDE; GRAY, 2018; SOARES; GRAY, 2018). The second reason is the code simplicity of not having to normalize the drill-curves, i.e. bring all variables to the same scale. Normalization is required for other state-of-art machine learning methods, such as artificial neural networks, support vector machines. Strobl et al. (2007) reported an problem of random forests when having variables in different scales. This leads to biased feature importance measures, which, however, can be solved by employing bootstrapping samples without replacement.

There are two different functions in Matlab to train ensemble learners based on tree-models. One function is `TreeBagger`, resulting in an object with the same name. Another function is `fitrensemble`, that can generate three objects (called “Superclasses”).

The use of `TreeBagger` enables to implement either bagged trees or random forests. In addition, the use of this class enables to implement a probabilistic prediction with the function `quantilePredict` (MATHWORKS, 2018c). This function is based on Quantile Regression Forest proposed by Meinshausen (2006).

Another function from Matlab that implements ensemble learners for regression problems is the function `fitrensemble`. With that, three objects (“Superclasses”) can be obtained (MATHWORKS, 2018d):

- `RegressionBaggedEnsemble` - it implements either bagged trees or random forests, but probabilistic prediction is not possible to be carried;
- `RegressionEnsemble` - it implements gradient boosting machines with the algorithm least-squares regression (LS_Boost) from Friedman (2001);
- `RegressionPartitionedEnsemble` - it assess the generalization error capability for the previous algorithms with cross-validation process.

4.2.3 Evaluation Metrics

The current study employs the following metrics to assess the generalization capability of the learned models: R-squared, root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). There are several other metrics to evaluate regression machine learning models; the work from Botchkarev (2018) serves the basis for this section.

The Pearson's correlation coefficient can be estimated by (GUYON; ELISSEEFF, 2003):

$$R = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\left[\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \right]^{1/2}} \quad (4.4)$$

where $\mathbf{y}$ is a vector n elements, representing the actual value (e.g. ROP measured values), y_i is the actual value of the element i from the vector $\mathbf{y}$. The mean of $\mathbf{y}$ is given by $\bar{y} = \sum_{i=1}^n y_i / n$. The predicted value for i -th observation is $\hat{y}_i$, and $\bar{\hat{y}} = \sum_{i=1}^n \hat{y}_i / n$ represents the mean of the predicted vector $\hat{\mathbf{y}}$.

The squared of correlation coefficient R (equation 4.4) can be also employed as evaluation metric. It is also common to compute the R-squared based on a different formula, as done for example by Nascimento et al. (2015b). Botchkarev (2018) called this other form to compute the R-squared as “coefficient of determination”.

The mean squared error (MSE) is calculated as the mean of squared sum of residuals:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (4.5)$$

The root mean squared error (RMSE) can be calculated as the square root of MSE:

$$RMSE = \sqrt{MSE} \quad (4.6)$$

The mean absolute error (MAE) expresses the mean of absolute residuals between the predict $\hat{y}_i$ and the actual y_i :

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (4.7)$$

Hegde et al. (2017) evaluated the predictive ROP models with another metric, called normalized error rate. Botchkarev (2018) presented several metrics that compute relative error. One of them was the mean absolute percentage error (MAPE), defined as:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{|\hat{y}_i - y_i|}{|y_i|} \quad (4.8)$$

One problem may occur when using relative metrics, such as the MAPE. Let us suppose an variable within a common range for the ROP values, i.e. from 0 to 50m/h. It is possible to exist an valid observation whose actual value is near to zero, e.g. 0.01m/h. If the predictive model estimates an value of 0.10m/h for this observations, the absolute error is 0.09m/h. This can be considered either an good prediction or very bad estimation, depending on the metric employed. In terms of absolute difference or means squared error, this difference is small, so that

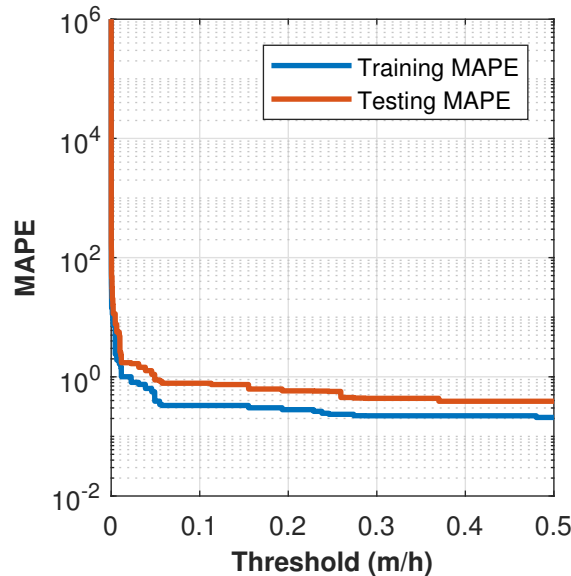
the predicted output can be considered as an good estimation, since the model could estimate an output also near to zero. However, in relative term, the predicted value is ten times higher then the actual value, and the relative absolute difference is $MAPE = (|0.10 - 0.01|/0.01) * 100 = 900\%$.

This problem happened, especially, for the drill-curves from Norway, since the data pre-treatment process is not failure free. For example, the drilling data for Well 5 had some observations considered as valid and belonging to rotary drilling state whose ROP values were around $10^{-9}m/h$. For such values, it is very likely to obatin very high relative errors. Since the MAPE is an mean metric, even an small quantity of outliers can corrupt this metric; the problem of mean statistics is better discussed in the work from Leys et al. (2013). As an alternative, a simple way to overcome this problem is to compute the MAPE if, only if, the actual value is above an threshold. In the Figure 4.5, we illustrate the influence of the threshold on the MAPE. We observed that not considering actual values below to $0.1m/h$, the adjusted MAPE was around 10^6 smaller than the original MAPE, resulting in an useful metric.

To generate the Figure 4.5, we first employed the data pre-treatment process, using the same setting of thresholds employed in the development of the results (see Chapter 5). Then, we trained a random forest with 100 trees to predict the ROP. The learning process occurred on half of observations from the data set drew at random, leaving out other 50% of observations to assess the generalization capability. Four inputs were employed in obtaining this predictive model, namely the depth, drillstring rotary speed, weight on bit and torque. The MAPE was calculated on the training dataset and the testing dataset.

In the current work, when we use the adapted MAPE, a subscript, a , is added to the acronym, resulting: $MAPE_a$.

Figure 4.5 – MAPE against the cut-off threshold, which enables to avoid the problem of small values in computing this relative metric.



Source: own authorship.

4.2.4 Hyperparameters Optimization

As discussed in the Section 2.2.7, ensemble models have several internal parameters (called hyperparameters) that influence on the prediction performance. In order to tune the hyperparameters, some possible approaches are: grid search, random search or Bayesian optimization. The function `fitensemble` has already implemented these optimization methods.

The `TreeBagger` function has not an optimizer implemented. In this case, an algorithm needs to be developed in order to tune the hyperparameters. The Code 4.2 is an example of optimization of two hyperparameter of a random forest: ‘numPTS’ (the amount of variables to be randomly split) and the depth of the trees in the forest, given by the fields ‘minLS’ (minimum leaf size). For that, two objects from `optimizableVariable` class are created, and gathered in the single variable ‘hyperRF’. Then, the Bayesian optimization function is called. The `bayesopt` function attempts to minimize the function handle `oobErrorRF`, that returns the out-of-bag mean-squared-error. In this case, the acquisition function of Bayesian optimizer is expected-improvement-plus. The readers may refer to works from Snoek, Larochelle and Adams (2012), Gelbart, Snoek and Adams (2014) for more details about the acquisition functions.

Code 4.2 – Example of hyperparameters tuning with Bayesian optimization for the TreeBagger object. Adapted from Mathworks (2018e).

```

1 % complexity (depth) of the trees in the forest
2 minLS = optimizableVariable('minLS', [1,30], 'Type', 'integer');
3
4 % the number of predictors to sample at each node
5 % where 'numInputs' is the amount of inputs selected
6 numPTS = optimizableVariable('numPTS', [1,numInputs-1], 'Type', 'integer');
7
8 %% 2x1 array OptimizableVariable objects
9 hyperRF = [minLS; numPTS];
10
11 %% Calling bayesopt to optimize the function handle '@oobErrorRF'
12 BayOpt = bayesopt(@(params) oobErrorRF(params,...
13     numTrees,...
14     dataTrain,...
15     y),...
16     hyperRF,...
17     'AcquisitionFunctionName', 'expected-improvement-plus',...
18     'Verbose',0);
19
20 %% Defining the Objective function
21 function oobEr = oobErrorRF(params, numTrees, dataTrain, y)
22 %   INPUTS:
23 %       - params: class of optimizableVariable
24 %       - numTrees: number of trees of the random forest
25 %       - dataTrain: a table having X and Y variables
26 %       - y: a string with the y-variable name
27 %   OUTPUT:
28 %       - oobEr: mean squared error for out-of-bag observations in the ...
29 %         training data
30 randomForest = TreeBagger(numTrees, dataTrain,y,...
31     'Method','regression',...
32     'OOBPrediction','on',...
33     'MinLeafSize',params.minLS,...
34     'NumPredictorstoSample',params.numPTS);
35 oobEr = oobError(randomForest, 'Mode','ensemble');
36 end

```

4.2.5 Feature Selection

Two approaches are employed in the current work to select the most important drilling variables to predict the performance indicators (ROP and SE). One method is so-called “driller-approach”, which tries to mimic how an expert driller would train predictive models for drilling parameters. In this case, the driller selects as inputs those drilling variables that he believes to be the most important to model (e.g. the ROP). Another approach employs evolutionary optimization algorithm to determine the best combination of drilling variables.

To develop the current work, there are some essential variables, because, without them it, is not possible to train predictive models for ROP nor SE for drilling optimization tasks (see Table 4.7). The bit hydraulic plays an important role in drilling process, but it is possible to carry

out the current study without them. Other information, such as vibration or rock strength, can provide important insights, such the estimation of rock strength. However, we believe that such recordings are not as essential as the surface parameters, since we aim to develop methods to optimize the WOB and rotating speed.

Table 4.7 – Drilling parameters required for the development of the current work.

Drilling Parameter	Importance
ROP and mechanical parameters (WOB, RPM, Torque)	● ● ● essential
hydraulic parameters (mud weight, mud flow, annular pressure)	● ● ○ important
vibration (e.g. stick-slip) and geological formation (e.g. p-wave velocity, which enables to compute the rock strength properties - UCS)	● ○ ○ complement

Source: own authorship.

4.3 DRILLING PARAMETERS OPTIMIZATION

The goal of the current work is to find the optimum magnitude of controllable drilling variables, especially the bit weight and rotating speed. For that, trained models of so-called performance indicators (e.g. rate of penetration or specific energy) are employed as objective function(s), either in single-objective optimization framework or in multi-objective optimization. First we present the single-objective optimization problem adopted in the current work, followed by the multi-objective optimization. A total of four optimization problems are studied, two in a single-objective optimization framework and other two in a multi-objective optimization framework

4.3.1 Single-Objective Optimization

We saw in the review Chapter 3 that most of the cited works formulated single-objective optimization problems for drilling optimization studies, using the maximization of rate of penetration as the objective function. In the current work, we investigated also the minimization of specific energy, formulating the following problems:

- the first optimization problem consists of maximizing the rate of penetration;
- the second optimization problem consists of minimizing the specific energy.

For both cases, we assessed the influence of optimizing the drilling variables not only on the objective function itself, but also on the another function. For example, if the problem is set as maximization of rate of penetration (ROP), we assessed the improvement of ROP and also the influence of maximizing ROP on specific energy.

4.3.2 Multi-Objective Optimization

Since optimizing solely ROP can decrease the drilling efficiency, one reasonable way to optimize the drilling variables is to consider also the concept of specify energy, as done by Gandelman (2012), Hegde and Gray (2018). The current work analyzed two multi-objective optimization problems based on the ε -constraint technique, as explained in next sections.

4.3.2.1 Third optimization problem

The third optimization problem considers also both functions, ROP and SE. The multi-objective optimization problem was formulated based on the ε -constraint technique, as follows:

$$\begin{aligned} \min \quad & SE(\mathbf{x}) \\ \text{s.t.} \quad & ROP(\mathbf{x}) \geq ROP_{actual} * \varepsilon \end{aligned} \tag{4.9}$$

where $\mathbf{x}$ is a vector from the searching space (i.e., WOB and RPM) subject to $x_i^{min} \leq x_i \leq x_i^{max}$, $i = \{1, 2\}$, $ROP(\mathbf{x})$ is the estimated ROP-value for any $\mathbf{x}$, ROP_{actual} is the actual ROP-value, and ε determines the magnitude of acceptable predicted-ROP. If $\varepsilon > 1$, we seek to decrease SE and, at the same time, improve the ROP. If $\varepsilon = 1$, the optimum pair of WOB and RPM is the one that can reduce the SE-value, maintaining the drill rate. If $\varepsilon < 1$, the minimization of SE accepts an decrease in ROP.

It is not expected, for some observations, that both objectives can be achieved (improvements in ROP and drilling efficiency) at the same time. If we employ only $\varepsilon > 1$, not every observation will have an optimum solution. Therefore, the following stepwise algorithm was proposed:

- a) first, an optimum solution is determined as the one that can improve the drilling efficiency and ROP at the same time. In this case, we have $\varepsilon_a > 1$;
- b) if there is no optimum solution that satisfies the previous condition, then it is accepted that minimization of SE resulted to, at least, the same ROP-value. In this case, we have $\varepsilon_b = 1$;
- c) if there is still no optimum solution, then an decrease in ROP is accepted, i.e., $\varepsilon_c < 1$.

This algorithm requires suitable values for the ε_a and ε_c . The sensitive analysis, carried out in the next chapter 5, helped us in determining these values.

4.3.2.2 Fourth optimization problem

The fourth optimization problem considers also both functions, ROP and SE. The multi-objective optimization problem was formulated based on the ε -constraint technique, as follows:

$$\begin{aligned}
& \max \quad \text{ROP}(\mathbf{x}) \\
& \text{s.t.} \quad \text{SE}(\mathbf{x}) \leq \text{SE}_{\text{actual}} * \varepsilon
\end{aligned} \tag{4.10}$$

The main goal of this problem is to optimize the ROP. If $\varepsilon < 1$, it is desired to improve the ROP and, at the same time, improve the drilling efficiency, expressed by decrease in SE-value. If $\varepsilon = 1$, the optimum pair of WOB and RPM is the one that can increase the ROP, maintaining, at least, the SE-value. If $\varepsilon > 1$, maximizing ROP accepts an decrease in drilling efficiency, expressed by higher SE-values.

If we employ only $\varepsilon < 1$, not every observation will not have an optimum solution. Therefore, the following decision workflow was proposed:

- a) first, an optimum solution is determined as the one that can improve the ROP and SE at the same time. In this case, we have $\varepsilon_a < 1$;
- b) if there is no optimum solution that satisfy the previous condition, then it is accepted an improvement in ROP, maintaining, at least, the SE-value. In this case, we have $\varepsilon_b = 1$;
- c) if there is still no optimum possible solution, then an decrease in drilling efficiency is accepted, i.e., $\varepsilon_c > 1$.

4.3.3 Grid-Search Strategy

Several techniques can be employed to search the optimum solution in the decision space. Gandelman (2012) investigated two approaches: one was a grid search in combination with *if-then* rules, and another the particle swarm optimization (PSO). Hegde, Daigle and Gray (2018) have recently investigated several optimization techniques for the problem of maximization of ROP, including: eyeball method, random search, simplex (amoeba) method, differential evolution, and PSO.

The current work employed the approach of grid search, as explained in the following paragraphs. The implementation of this approach is simple, but it can be very expensive in terms of computational costs, depending on how fine is the adopted grid.

Let us consider a optimization problem with only one variable to be optimized, denoted by x . Let us suppose the variable x is continuous, and this variable must lay within the interval $[x_{\min}, x_{\max}]$. If we split this interval with k -grid points, the amount of possible values of x are $k + 2$. In this case, the searching space $\mathcal{X}$ is defined as

$$\mathcal{X} \in \{x_1, x_2, \dots, x_n\} \tag{4.11}$$

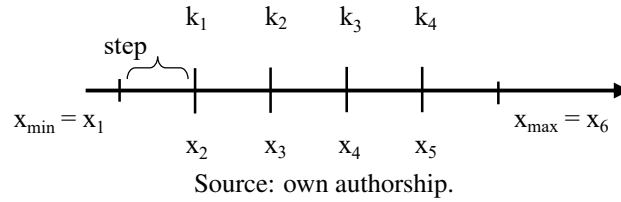
where $n = k + 2$, $x_1 = x_{\min}$. If the grid is uniform, i.e., equidistant grid points, any element x_i

from $\mathcal{X}$ can be obtained as following:

$$x_i = x_{min} + (i - 1) * s \quad (4.12)$$

where $i = \{1, 2, \dots, n\}$ and s denotes “step”, which can be calculated $s = (x_{max} - x_{min}) / (k + 1)$. The Figure 4.6 illustrates an generic variable x with $k = 4$ equidistant grid points.

Figure 4.6 – Equidistant grid points between lower and upper limits of an variable x .



If more variables need to be optimized, the searching space of each individual variables can be split in any k grid points. Then, it is possible to create all combinations of all variables to be optimized. If j represents the amount of variables to be optimized and all variables have the same value of k grid points, the amount of points in the search space becomes $(k + 2)^j$. In the current work, we employed $k = 98$, resulting in 10^4 possible combinations of WOB and RPM to be assessed, for each observation.

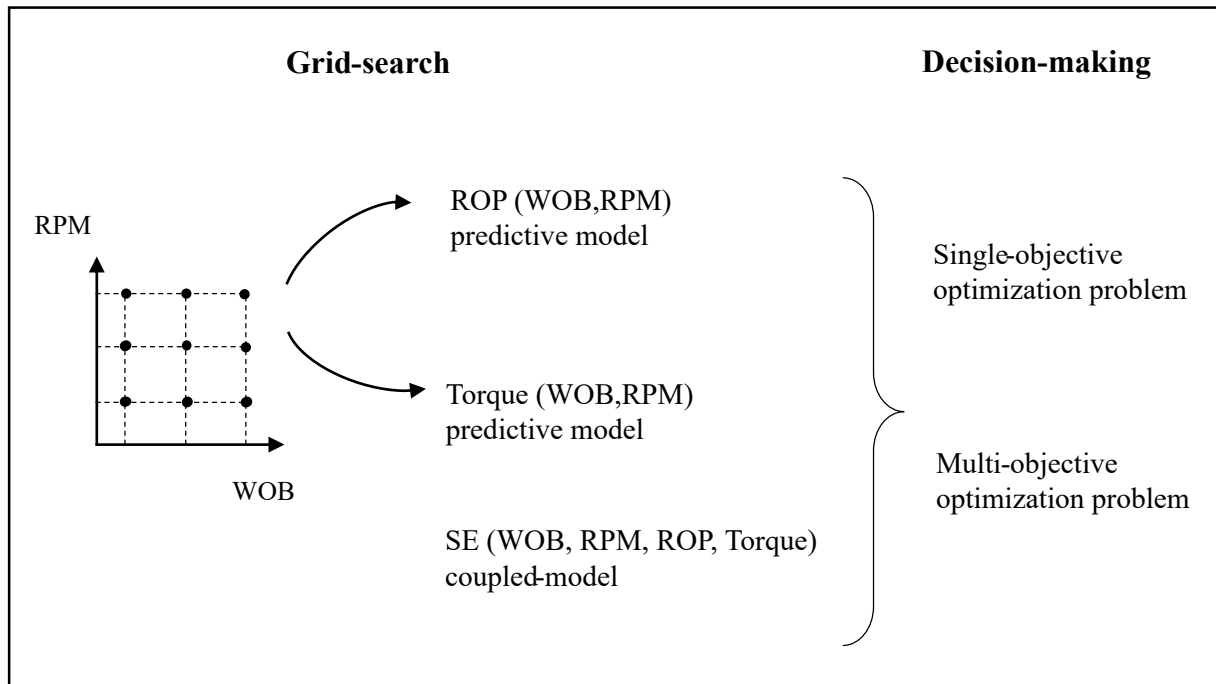
4.3.4 Optimization Workflow

The current work developed the optimization problems, aiming to investigate different approaches for decision-making. To speed up the simulations, the investigation of several combinations of WOB and RPM on objectives functions, e.g. ROP and SE, was carried out apart from the decision making. The Figure 4.7 illustrates the employed workflow employed in the optimization studies. First, the influence of several pairs of WOB and RPM on the ROP and Torque (if available) based on predictive models were computed. Then, the predictive values of ROP and torque in combination with pairs of WOB and RPM were used as inputs in SE-formulation. This process was done for each observation, resulting in matrixes $(k + 2)^j \times observations$. For some cases, the obtained matrixes had the magnitude of $10^4 \times 10^4$. After obtaining such matrixes, they could be used to assess different optimization problems, speeding up this step, since it was not required anymore to map again how decision space influenced on objective space (ROP and SE).

The downside of this approach is the difficulty in assessing the computational effort. It could be possible to check the elapsed time required to generate the matrixes of ROP, Torque and SE. However, a more realistic measure of simulation time can be achieved by checking elapsed time during the whole optimization process: from training the predictive models to the decision-making algorithm, passing through the search step and evaluation on the respective objective functions.

However, assessing the computational effort was not the main goal of the current work. Instead, the main goal was to assess different ways to formulate drilling optimization problems. Therefore, the described workflow was adopted.

Figure 4.7 – Optimization workflow adopted in the current work.



Source: own authorship.

5 RESULTS AND DISCUSSION

This chapter presents the results, starting with the analysis for predictive models of the so-called *drilling performance indicators*, namely the rate of penetration and the specific energy. Then, this chapter covers the use of these predictive models in a real-time optimization of controllable drilling variables.

5.1 RATE OF PENETRATION PREDICTION

5.1.1 Data Pre-Treatment Setting

The data pre-treatment method, presented in Section 4.2.1, was the first step prior to building the predictive models. The parameters for the data pre-treatment used in the further analysis are listed in the Table 5.1. The hampel filter parameters were determined based on the influence of them on the amount of the detected outliers, and its influence on the missing values (see Appendix C). The Appendix D illustrates the analysis performed to determine the threshold for the automated operation recognition. The Appendix E illustrates the importance of validation of observations (the last step of the data pre-treatment).

Table 5.1 – Data pre-treatment setting

Parameter	Value	Unit
<i>Hampel</i>		
Number of Neighbors	5	-
Threshold	3	-
<i>Operation Recognition</i>		
ε_1 - bit on bottom	0.15	<i>m</i>
ε_2 - circulation	100	<i>lpm</i>
ε_3 - drillstring rotating	10	<i>rpm</i>
ε_4 - transient (lower and upper limit for the block position)	0.5	<i>m</i>
transient period - after a change from not-drilling to rotary drilling	2	<i>min</i>
<i>Validation</i>		
2 nd rule for Norway	45	<i>ton</i>
2 nd rule for Pre-Salt	100	<i>klbf</i>

Source: own authorship.

The Table 5.2 gives an overview of the amount of observations resulted from the data pre-treatment process, highlighting the automated operation recognition step and the validation step. The data pre-treatment for the Well 1 from Norway resulted in a small dataset. For this

reason, this dataset will not be considered any more for the following analysis in the current work.

Some drill-curves from Pre-Salt, namely the Wells A and D, were employed partially. For the Well A, we will use the drill-curve from the section from 4 127 m to 4 185 m in the following analysis, discarding the drill-curves from the previous run. For the Well D, a data transmission loss occurred at the final of dataset. As consequence, we will not consider the final section of this dataset from here on. These observations are better explained in the Appendix B.

Table 5.2 – Amount of observations for each well from the original dataset to the final dataset after the data pre-treatment.

Well	Original	Operation Recognition				Not Valid		Final	Selected?
		Rotary	Sliding	Transient	Not-Drilling	1 st rule	2 nd rule		
Pre-Salt									
A ^(a)	1 326	-	-	-	-	860	0	466	✓
B	509	-	-	-	-	0	0	509	✓
D ^(b)	927	-	-	-	-	0	0	927	✓
H	5 184	-	-	-	-	140	0	5 044	✓
Norway									
1	5 973	204	2 483	313	2 973	160	0	44	✗
3	10 933	5 561	911	727	3 734	0	23	5 538	✓
4	13 659	7 499	140	1 167	4 853	26	21	7 452	✓
5	20 210	3 623	7 279	826	8 482	126	20	3 477	✓
6	14 737	6 278	2 337	1 266	4 856	78	104	6 096	✓
7	32 837	10 940	766	2 479	18 652	646	193	10 101	✓
8	17 107	4 724	6 245	700	5 438	24	0	4 700	✓
9	5 013	3 057	0	507	1 449	9	0	3 048	✓
10	14 417	8 253	0	854	5 310	27	0	8 226	✓

^(a) this dataset will be partially considered, i.e., only the section from 4 127 m (depth) to 4 185 m will be used. ^(b) Drill-curves for this dataset will not be considered after the depth of 4 537 m. See Appendix B for more details. For both cases, the amount of observations at the final stage from data pre-treatment is smaller than appeared in this table.

Source: own authorship.

5.1.2 Simple Comparison between ROP Models

As seen in the Chapter 3, the use of machine learning results in more accurate ROP models than those ROP models obtained by the traditional models (analytic equations). In this section, we compared a well-known ROP model proposed by Bourgoyne and Young (1974) with a data-driven model, learned from historical drilling data.

Nascimento (2016) predicted the ROP for the first run of Well B, based on the following models: Cunningham, Maurer and Bourgoyne & Young. The last model could better fit the drill-

Table 5.3 – Simple comparison between the Bourgoyne and Young Model and a random forest regression for ROP prediction, using the drilling data of the Well B. The standard deviation is given in parentheses.

Work	Train					Test			
	MAPE (%)	MAE	MSE	OOB	R	MAPE (%)	MAE	MSE	R
<i>Nascimento (2016)</i>									
BYM	36.5	-	-	-	-	-	-	-	-
BYM Optimized	23.1	-	-	-	-	-	-	-	-
<i>Current Work</i>									
Random Forest	10.5 (0.4)	0.272 (0.013)	0.158 (0.013)	0.335 (0.028)	0.911 (0.008)	15.5 (0.8)	0.404 (0.023)	0.339 (0.055)	0.776 (0.029)

Source: own authorship.

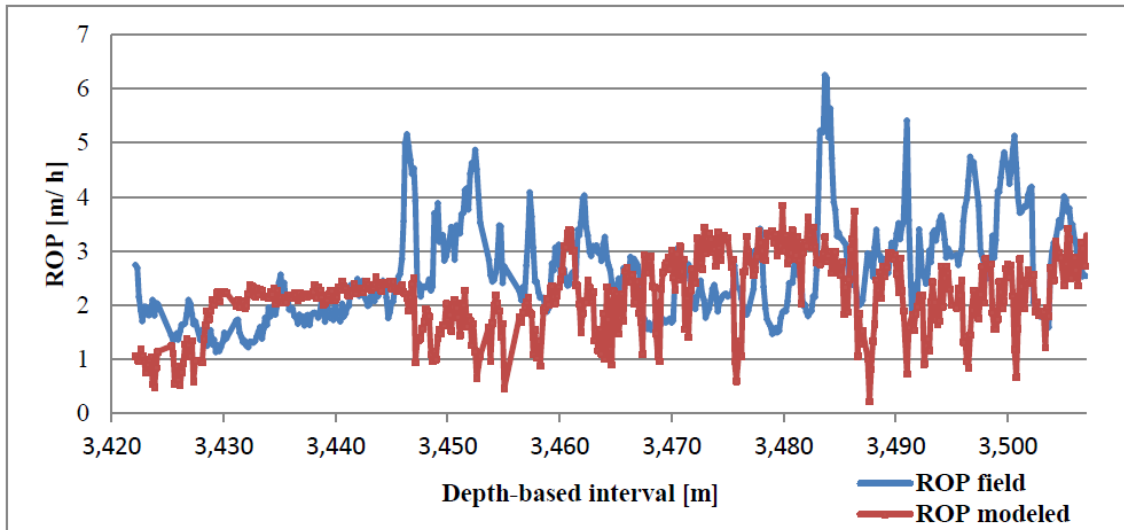
curves. In addition, the author proposed some adjustments for the Bourgoyne & Young model, adapting this model for pre-salt region. Among other adjustments, the author employed a wider coefficients ranges. As consequence, the relative error decreased from 36.52% to 23.1%. There is no mention about leaving out any part of this dataset to assess the generalization capability. Therefore, we classified the reported relative errors as *training errors*, since the relative errors were calculated on observations employed to build the predictive models.

The ROP model based on Bourgoyne & Young without adjustments from (NASCIMENTO, 2016) enables to perform the following comparison, as seen in the Figure 5.1. Here, we compared the Bourgoyne & Young model obtained in that work with a random forest regression, using only four variables as inputs: depth, surface weight on bit, drillstring rotary speed and total flow rate of all active pumps. The number of variables selected at random for each decision split was changed from the default condition (one third of number of inputs, i.e. one variable) to two. After leaving out a random 50% of observations, we trained random forests with 100 trees on the remaining 50%, and used the left-out 50% as a test set. We repeated this procedure 100 times, and at each time with a different data partition, but maintaining the 50-50 data partition. We computed the average of training and testing errors for each well, as well as the standard deviation (values shown in parentheses). The results are presented in the Table 5.3.

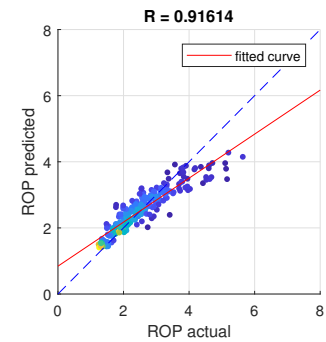
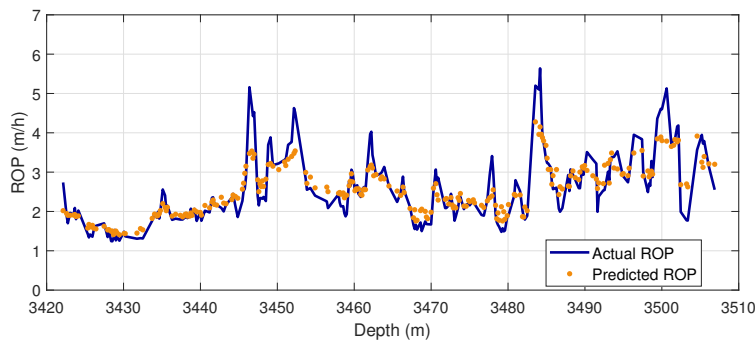
The current work employs a systematic data pre-treatment before analyzing the drill-curves. In the work from Nascimento (2016), the data cleanness process was not detailed. Therefore, we called this comparison as a *simple* one, since this difference in data pre-treatment may be a source of bias. However, in this simple comparison, we observed a better estimation capability of the machine learning algorithm over the Bourgoyne & Young model to predict the ROP, as expected according to the literature review (see Chapter 3). Another advantage was the better accuracy obtained with simpler ROP models in terms of amount inputs, since we

employed only four inputs to train random forests, which are easily accessible in drill rigs.

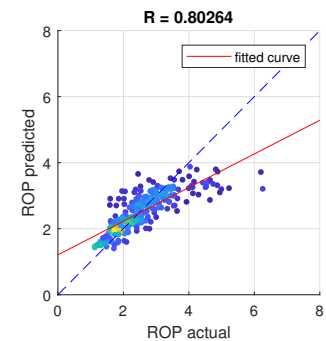
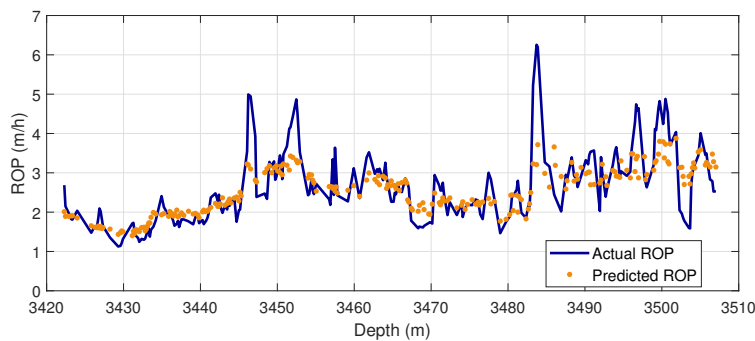
Figure 5.1 – Comparison between ROP Models



(a) Bourgoyne and Young model not optimized



(b) Random Forest - Training Set



(c) Random Forest - Testing Set

Source: a) Nascimento (2016); b-c) own authorship. To plot the regression between the actual ROP and the predicted ROP, we have employed the function `dscatter` developed by Eilers and Goeman (2004).

5.1.3 Selection of Most Important Inputs in the Prediction Task

Some works detailed the process of selecting inputs in the task of ROP prediction (ANEMANGELY et al., 2018; ESKANDARIAN; BAHRAMI; KAZEMI, 2017; HEGDE et al., 2017).

In the current work, three experiments based on different strategies were carried out. The goal was to investigate the selection of the most important drilling parameters, as following:

- a) Experiment 01 - the driller's approach with only four inputs which are easily accessible in drill rigs (e.g. depth, weight on bit, rotary speed, total flow rate);
- b) Experiment 02 - the previous surface variables, plus two more variables (e.g. pump pressure);
- c) Experiment 03 - binary genetic algorithm, where all drilling variables are possible candidates to model the ROP.

It is important to note that the rock strength was not considered in ROP modeling, as Hegde et al. (2017) did by estimating UCS with log-curves. Estimating this parameter was not possible due to lack of sonic-logs in the drill-curves employed in the current work. However, Bilgesu et al. (1997) showed the possibility of prediction the ROP without any parameter related to the drilled formation (e.g. formation abrasiveness and drillability) or to bit status (e.g. bit tooth, bearing wear). Therefore, omitting the information of bit status and formation does not invalidate the current work.

5.1.3.1 Driller's Approach

The first two approaches are called here as *driller's approach*, because it tries to mimic an expert driller dealing with the task of ROP modeling. In this case, the driller, based on his knowledge, selects the inputs which he believes to be the most important drilling parameters for ROP modeling. Here, it is important to note that such expert could select as many as possible variables to model the ROP due to the complex nature of drilling. However, in our experiments, the driller would be encouraged to do the opposite, i.e., to add as few as possible variables, being able to exclude even the very important parameters, such as the drillbit type or rock strength. To convince him, the work from Bilgesu et al. (1997) would be of a great value.

Two experiments were carried out based on this approach. The Experiment 01 one employed only four inputs, which are easily accessible in drill rigs. In the Experiment 02, we employed six inputs, by adding two more variables to the selected inputs from the Experiment 01. The variables selected in these experiments are listed in Table 5.4.

For both experiments, the configuration of random forests was the same. The number of variables to be selected at random for each decision split was set to two. We trained random forests with 100 trees on 80% of observations selected at random. The left-out 20% was used as a test set. We repeated this procedure 100 times for each well, splitting randomly the drilling data into training and testing dataset at each time, but maintaining the same 80-20 proportion. Then, we averaged the training and testing errors, computing also the standard deviation. For clarity, only the average values of metrics on the testing dataset are shown in the Table 5.5. We performed a paired-sample T-Test for each metric in order to determine whether adding

two extra-inputs improved the generalization capability or not. A significance level of 1% was considered.

Table 5.4 – Inputs selected for each experiment. The inputs of the Experiment 02 consist of adding the listed variables to the group of the Experiment 01.

Well	Experiment 01 4 inputs	Experiment 02 Adding two more inputs
<i>Pre-Salt</i>		
A	Depth, SWOB, RPM, TFLO	ECD, SPPA
B	Depth, SWOB, RPM, TFLO	ECD, SPPA
D ^(a)	Depth, SWOB, CRPM	ECD, SPPA
H	Depth, SWOB, RPM, TFLO	ECD, SPPA
<i>Norway</i>		
3-10	DMEA, WOB, RPMA, TRQ	HKL, BPOS

^(a) the measurement of TFLO, total flow rate, is not available. Therefore, it was left out.
Source: own authorship.

Table 5.5 – Average of evaluation metrics for ROP prediction on testing dataset.

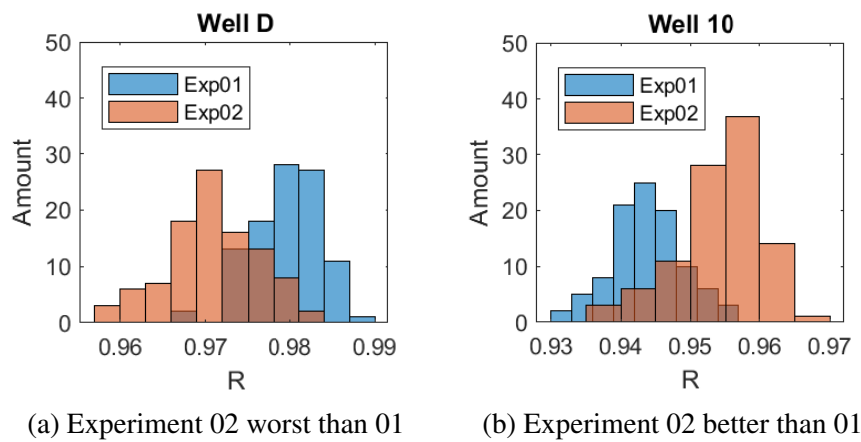
Well	Experiment 01				Experiment 02			
	MAPE _a	MAE	MSE	R	MAPE _a	MAE	MSE	R
<i>Pre-Salt</i>								
A	0.111	0.150	0.042	0.938	0.115 ≈	0.152 ≈	0.040 –	0.942 ≈
B	0.136	0.353	0.254	0.839	0.147 ✗	0.385 ✗	0.297 ✗	0.808 ✗
D	0.154	0.564	0.677	0.979	0.178 ✗	0.701 ✗	1.014 ✗	0.971 ✗
H	0.210	0.628	1.147	0.980	0.218 ✗	0.663 ✗	1.230 ✗	0.978 ≈
<i>Norway</i>								
3	0.405	2.995	27.383	0.934	0.336 ✓	2.819 ✓	24.455 ✓	0.942 ✓
4	0.082	1.890	11.634	0.948	0.089 ≈	1.791 ✓	10.689 ✓	0.953 ✓
5	0.612	4.596	53.436	0.896	0.411 ✓	4.003 ✓	41.904 ✓	0.919 ✓
6	0.057	1.065	6.053	0.935	0.055 ≈	1.001 ✓	5.480 ✓	0.942 ✓
7	0.121	1.111	7.188	0.978	0.133 ≈	1.061 ✓	6.272 ✓	0.981 ✓
8	0.086	0.759	3.095	0.971	0.093 ✗	0.731 ✓	3.112 ≈	0.971 ≈
9	0.539	8.443	149.324	0.895	0.465 ≈	8.119 ✓	137.677 ✓	0.905 ✓
10	0.235	1.604	8.516	0.944	0.207 ✓	1.411 ✓	7.063 ✓	0.954 ✓

The symbols ✓, ✗, ≈ mean, respectively, that the Experiment 02 resulted in a better prediction of ROP than the Experiment 01, worst, or not statistically different. Tests performed at the 1% significance level.
Source: own authorship.

In order to understand the paired-sample T-Test, we plotted histograms of R-correlation on testing dataset, comparing both experiments against each other, as seen in the Figure 5.2. The

paired-sample T-Test enables to statistically determine whether adding two extra-inputs improved the generalization capability or not. For the case of Well D, adding two more inputs resulted in worst generalization capability according to the selected metric. In this case, the p-value was $3.4169\text{e-}24$, so that we could reject the null-hypothesis (the sample's mean of Experiment 02 was not less than the sample's mean of Experiment 01). However, the opposite occurred for the Well 10, since adding two more variables improved the model accuracy. With a p-value of $9.4259\text{e-}22$, we could reject the null-hypothesis (the sample's mean of the Experiment 02 was not greater than the sample's mean of the Experiment 01).

Figure 5.2 – Histograms of R-correlation on testing dataset.



Source: own authorship.

This difference illustrates the difficulty in selecting manually the inputs. While adding more variables to model the ROP for the drill-curves from Pre-Salt decreased the prediction accuracy, extra inputs improved, in most cases, the prediction accuracy for Norwegian drilling data. Therefore, it is worth investigating sophisticated techniques to determine the best combination of inputs.

5.1.3.2 Evolutionary Feature Selection

In the Chapter 2, a brief overview of some possible approaches to select the best (or a good one) subset of features was given. Here, the Binary Genetic Algorithm is employed to determine the selection of the best inputs. This method can be implemented with a Matlab function called, `ga`, by setting the population type to 'bitstring'. The employed code resembles the one developed by Oluleye et al. (2014a), Oluleye et al. (2014b). In the current work, the fitness score was the out-of-bag estimation of MSE. Generally, the GA resulted in small subsets with one to three variables. Seldom, the mutation process led to an empty chromosome, i.e. without any inputs. When this case occurred, the fitness function returned an out-of-bag MSE value of 100 (much higher than the average for all wells, except for the Wells 5 and 9, which both require an even higher penalization).

To run this simulation, random forests were employed with the same configuration from the Experiments 01 and 02, expect for the amount of weak learners, which was reduced from 100 to 30 in order to speed up the simulations. The configuration of Binary Genetic Algorithm is listed in the Table 5.6.

Table 5.6 – Options for Feature Selection based on Binary Genetic Algorithm

Variables	Setting
<i>Binary Genetic Algorithm</i>	
population type	'bitstring'
population size	50
generations	50
parents selection	tournament selection with size 2
amount of elite children	2
crossover function	scattered
crossover fraction	0.80
mutation	uniforme distribution
mutation fraction	0.15
<i>Searching Space</i>	
possible features	all inputs

Source: own authorship.

The initial population was created based on an function proposed by (OLULEYE et al., 2014a), as seen in the Code 5.1. The initial step has an great impact in this evolutionary algorithm. This could be could adapted by setting a fix cut-off value instead of a random value RD, avoiding the problem of starting a population with only few inputs or too many inputs (see Figure 5.3). An alternative for RD is to set a constant values, e.g. 0.5. For this case, the expected mean for the amount of inputs selected in the initial population is the half of genome's cardinality, since the function `rand` generates uniformly distributed random numbers between 0 to 1. However, the current Experiment 03 employed the proposed function from Oluleye et al. (2014a) to initialize the population.

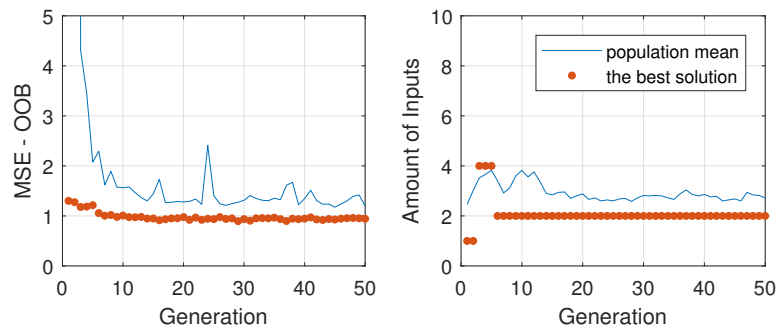
Code 5.1 – Function to initialize the population. Source: Oluleye et al. (2014a)

```

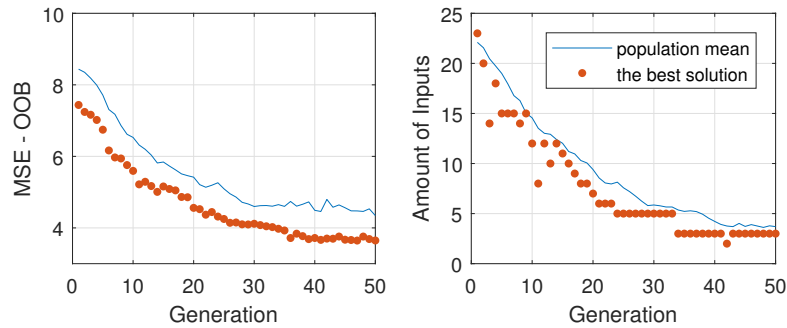
1 function [pop] = PopFunction(GenomeLength,~,options)
2     RD = rand;
3     pop = (rand(options.PopulationSize, GenomeLength)> RD); % Initial Population
4 end

```

Figure 5.3 – The initial population influenced by the random starting for the Well 8 and 10. For both cases, the cardinality of possible inputs was 25.



(a) Well 8 - Initial population smaller than the half of genome's cardinality



(b) Well 10 - Initial population higher than the half of genome's cardinality

Source: own authorship.

The best subsets obtained for each well is listed in the Table 5.7. The out-of-bag errors of the optimum solution indicate a better prediction capability than those models from the Experiments 01 and 02. However, the obtained subsets are not useful, as explained in the following paragraph.

As can be seen in the Table 5.7, all cases have an common input related to the depth. Sometimes, similar variables related to the depth was selected in the optimum subset (the Well D is an example). Other variables which, in principle, are not directly related to the rate of penetration were selected in the optimum subset of features, e.g. the block position. In additional, it is also important to note the fact of not having anymore those variables, which are believed to be important for the ROP modeling, such as the weight on bit and rotary speed for many wells. Since a ROP model without these controllable variables does not enable to optimize them (the main objective of the current work), such subsets will not be considered in the following analysis.

Therefore, the obtained subsets are not useful for the current work purpose. Perhaps, a better pre-selection of the possible candidates to used be as inputs in the prediction of ROP could improve the subset selection.

Table 5.7 – Best subsets obtained by binary genetic algorithm.

Well	Optimum Subset	Amount of Inputs	OOB - MSE	Simulation Time
<i>Pre-Salt</i>				
A	Depth, BLKP	2	0.017	258 s
B	Depth, BLKP	2	0.096	236 s
D	Depth, CDEPTH, STUCK	3	0.255	289 s
H	Depth	1	0.522	983 s
<i>Norway</i>				
3	DBTM, DMEA, BPOS, RPMB	4	18.10	1 493 s
4	DBTM, DMEA, BPOS, HKL	4	8.65	1 791 s
5	DBTM, DMEA, RPM	3	35.32	910 s
6	DBTM	1	2.90	1 139 s
7	DBTM, DMEA, BPOS, HKL	4	4.73	2 618 s
8	DMEA, CTVL	2	0.94	954 s
9	DBTM, DMEA, BPOS, HKL, CTVL	5	98.02	789 s
10	DMEA, BPOS, CTVL	3	3.65	2 032 s

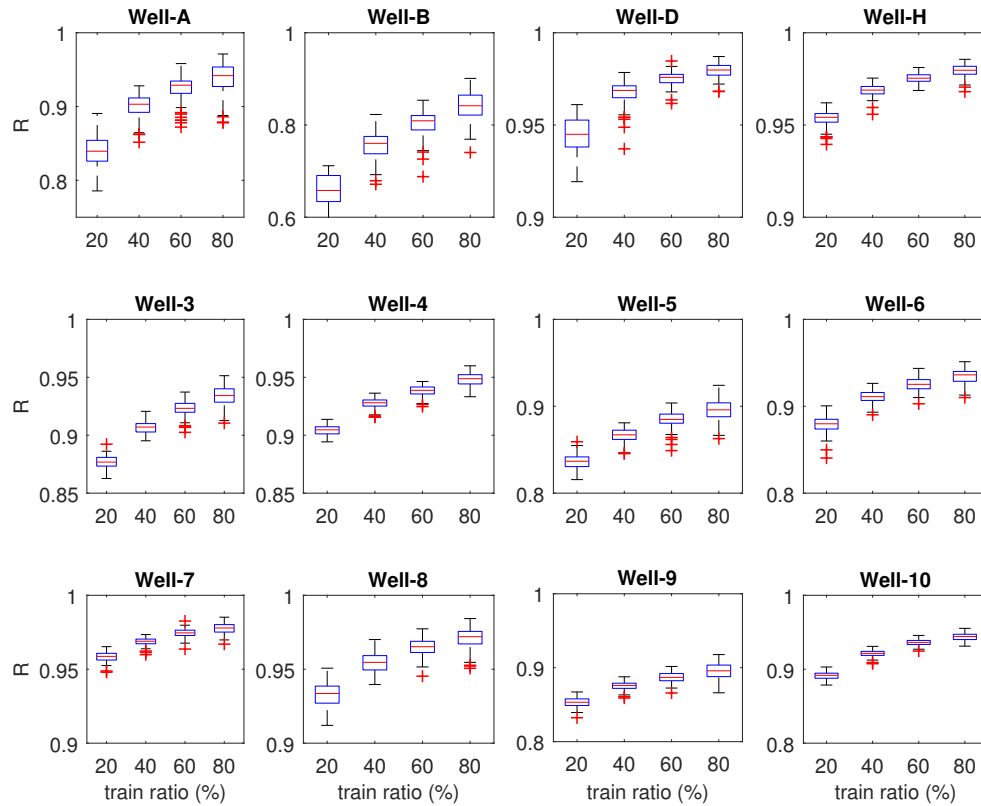
Source: own authorship.

5.1.4 Ratio Influence of Training Dataset to Overall Dataset

Here is investigated the influence of amount of training dataset on the performance of the ROP predictive model for each well. For that, we employed the same procedure from the Experiment 01, varying only the amount of observations in the learning phase. The term *train ratio* denotes the relative amount of observation employed in the training phase to the overall amount of observations in the dataset. For clarity, boxplots only for the R-correlation are shown in the Figure 5.4.

As expected, an improvement in the ROP models accuracy was observed as more observations were available in the training phase. While some predictive models achieved good generalization capability with small values of train ratio (e.g. Well H, 7 and 8), other models required a higher train ratio (e.g. Well A and B). In general, a train ratio of 60 % or even 40 % provided an acceptable prediction accuracy.

Figure 5.4 – Testing R-correlation - sensitive analysis of the train ratio.

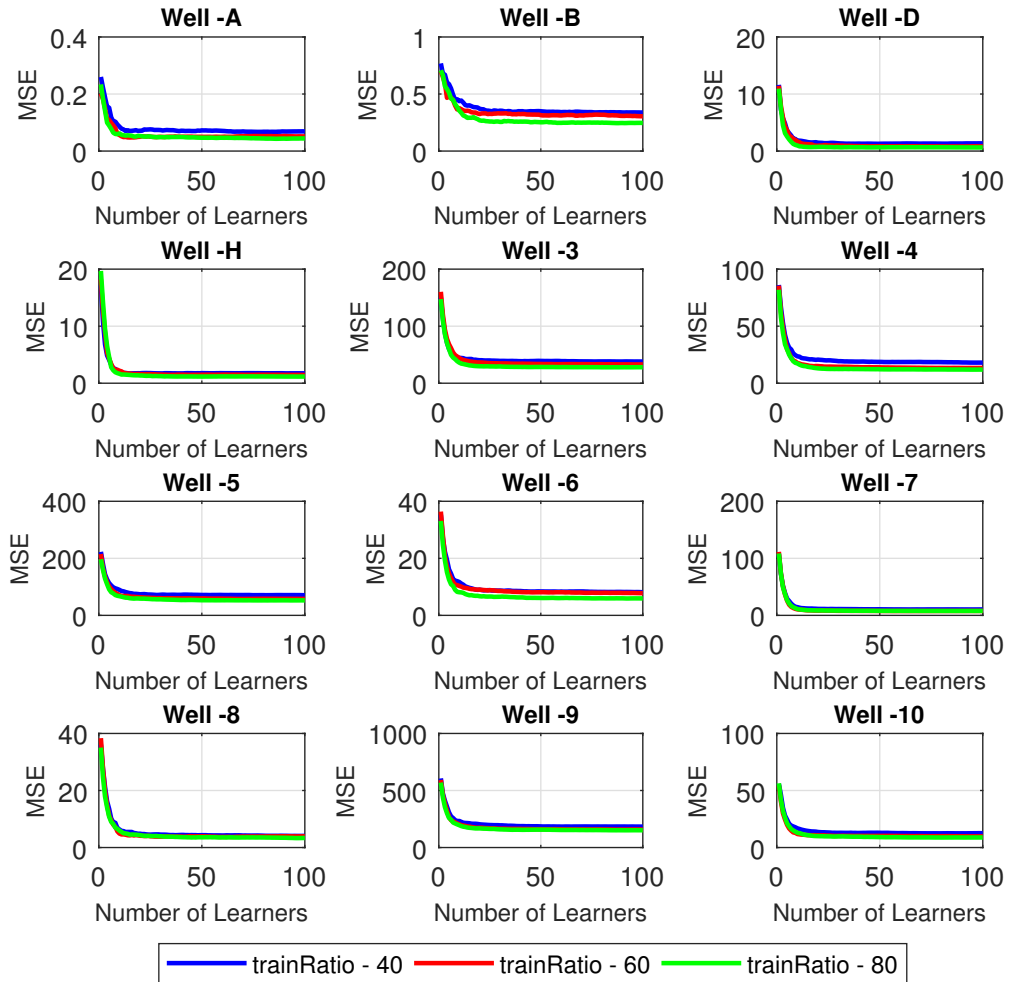


Source: own authorship.

5.1.5 Cumulative MSE of Random Forests

The cumulative mean-squared error from out-of-bag observations were investigated, and shown in the graphs forms (see Figure 5.5). Such analysis enables us to quantify how the prediction performance of random forests behaves as the amount of trees increases. We observed a error-convergence at 20 trees. After that, adding more trees to random forests did not improve the accuracy of ensemble models. Therefore, it is possible to reduce the amount of trees as done when selecting the optimum subset based on binary genetic algorithm (Section 5.1.3.2) in order to speed up the simulations.

Figure 5.5 – Cumulative out-of-bag MSE for ROP-prediction: random forests trained using the same procedure from the Experiment 01.



Source:own authorship.

5.2 SPECIFIC ENERGY PREDICTION

As seen in the drill-curves from pre-salts (e.g. Appendix B), the calculated specific energy was much higher than rock strength. Silva (2016) tested carbonates cores from pre-salt Santos Basin; the average rock strength from pre-salt carbonates was UCS 43 MPa. However, this difference is not necessary a problem, and the specific energy can be employed to assess the drilling efficiency. Hegde and Gray (2018, p. 399) stated, “for purposes of cost function, the value of MSE^1 is not as important as the relative change in MSE in this analysis”. Therefore, we can employ the specific energy as a cost function in drilling optimization, i.e., the less energy is spent, the more efficiency the formation is being drilled.

¹ In the current work, we employ the acronym of SE for the concept of (mechanical) specific energy, avoiding any misunderstanding

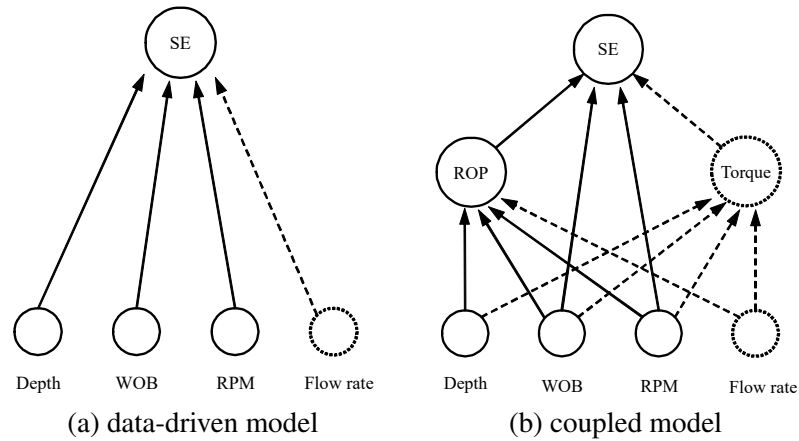
Gandelman (2012) used the estimated ROP to calculate the specific energy in his study on real-time drilling optimization. Hegde and Gray (2018) employed a similar approach, but instead of coupling the SE model with only a predictive ROP model, the authors calculated the SE based on estimated ROP and estimated Torque. The main idea of the mentioned models were to compute how changes in the controllable drilling variables (e.g. WOB and RPM) would affect the ROP. Then, the SE was calculated based on estimated values of ROP with the respective new values for the inputs, by using the Teale's formulation. Hegde and Gray (2018) called the resulting SE models as “coupled-model”.

Since the specific energy is calculated on several drilling variables, it is not common to predict the SE based on exclusively data-driven models, as done for the ROP. In literature, it was not found any work trying to estimate the SE based on exclusively data-driven models. Therefore, we investigated both approaches to model the SE, namely:

- exclusively data-driven SE models;
- coupled models, such as done by Hegde and Gray (2018), Gandelman (2012).

The goal of the comparing both approaches, illustrated in Figure 5.6, is to determine which one is more suitable to be used in the drilling optimization problem. Before presenting the results of the predictive and coupled models, we discuss some aspects regarding specific-energy calculation in the following section.

Figure 5.6 – Data-driven models against coupled models for SE prediction



Source: own authorship.

5.2.1 Calculating SE

We observed that some calculated values of the specific energy, especially, for the Norwegian wells, were too high, i.e., above 10^{30} MPa. This happened, because the drill-curves, especially from Norway, had some observations considered as rotary drilling with ROP values

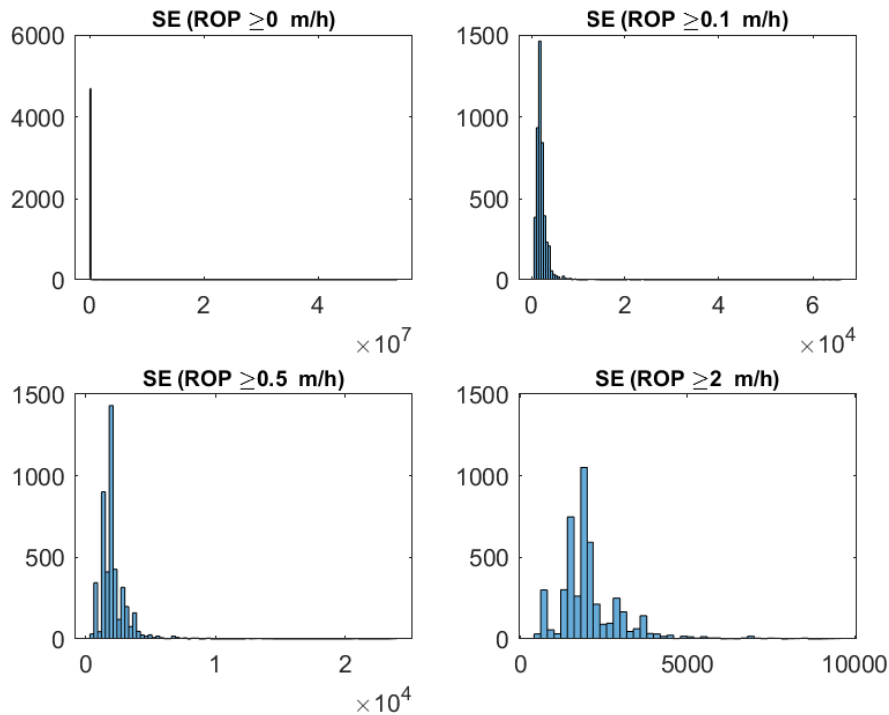
near to zero. According to Teale's formulation, a ROP-value near to zero leads to

$$\lim_{ROP \rightarrow +0} \frac{120 * \pi * RPM * Torque}{A_b * ROP} = \infty \quad (5.1)$$

As consequence, we can say $\lim_{ROP \rightarrow +0} SE = \infty$, since the term $\frac{WOB}{A_b}$ is not affected directly by the ROP value.

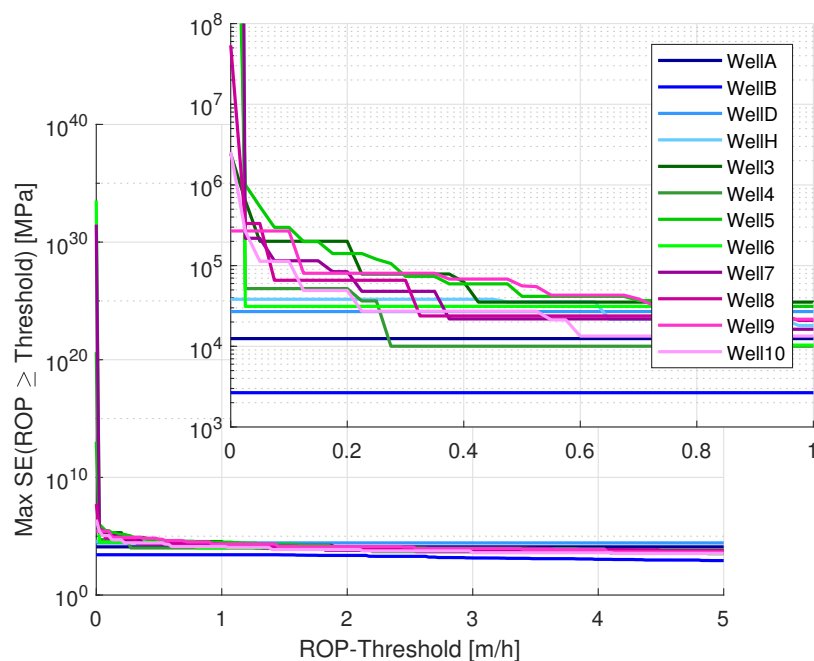
One way to avoid too high SE values is to adopted a ROP-threshold. In this case, the SE is calculated only if the ROP of the respective observations is greater than or equal to a threshold. This threshold received the name of *ROP-threshold*. In Figure 5.7, we plotted histograms of the calculated SE, considering four levels of ROP-threshold: without threshold (0 m/h), 0.1 m/h, 0.5 m/h and 2 m/h. We observed an decrease in the maximum value of SE as the ROP-thresholds increased. Even small values of ROP-thresholds decreased substantially the maximum values of SE. This happened not only for the Well 8 illustrated in Figure 5.7, but also for other Norwegian wells. This becomes clear in Figure 5.8, where we plotted the maximum calculated SE value for each well against the respective ROP-threshold.

Figure 5.7 – Histograms of SE calculated based on Teale's formulation, considering four levels of ROP-threshold - well 8



Source: own authorship.

Figure 5.8 – Maximum value of SE for each well against the ROP-threshold



Source: own authorship.

To calculate the specific energy, we selected a suitable SE model for each well, respecting the availability of drilling variables present in data-sets. For example, if the torque is not available, the Teale's equation cannot be directly applied. In this case, Rabia's model or Pessier-Fear's equation are possible alternatives to be considered (see the Section 2.1.3.2 for more details). The employed SE models and drilling variables used to calculate this parameter are listed in the Table 5.8.

Table 5.8 – Drilling variables used to calculate the SE

Well	SE Formulation	Variables used				ROP-Threshold
		WOB	Torque	RPM	ROP	[m/h]
<i>Pre-Salt</i>						
A	Rabia	SWOB	-	RPM	ROP5	-
B	Teale	DWOB	DTOR	RPM	ROP5	-
D	Rabia	SWOB	-	CRPM	ROP5	-
H ^(a)	Teale	SWOB	STOR	RPM	ROP5	-
<i>Norway</i>						
3-10	Teale	WOB	TRQ	RPM	ROP	0.5

^(a) for the Section 5.2.2, the Teale's formulation was employed for the Well H. However, for drilling optimization problems from the Section 5.3, the specific energy was estimated based on Rabia's formulation, because of the low accuracy in torque prediction for the Well H, resulting in a poor estimation for SE values.

Source: own authorship.

5.2.2 Data-driven models for the Specific Energy

The rate of penetration and torque are integral parts of the specific energy formulation. The magnitude of these variables depends also on the controllable drilling variable(s) to be optimized (in the current work, they are weight-on-bit and rotating speed). However, a SE model without the core variables as inputs (ROP and torque) simplifies, obviously, SE modeling and the implementation of drilling optimization as well. Modeling the SE becomes simpler because we will not need to couple the SE model with other predictive models (ROP and, optionally, torque), as done previously by Gandelman (2012), Hegde and Gray (2018).

Therefore, we tried to predict the SE based exclusively on data-driven models. We followed the same approach of ROP modeling. That is, we selected only some easily-accessible variables in drill rigs as inputs, mainly: depth, RPM, WOB, total-flow (if available). The selected inputs for each well are listed in the Table 5.9.

We trained random forests with 100 trees on 80% of observations drew at random, leaving out 20% to assess the generalization capability of the SE models. The number of variables to be selected at random for each decision split was set again to two. We repeated this procedure 100 times, splitting, at each time, the dataset into training and testing datasets at random, but maintaining the 80-20 partition. The average metrics of the obtained SE data-driven models were presented in the Table 5.10.

Table 5.9 – Inputs selected for the not-coupled and coupled models to predict the SE.

Well	SE data-driven models			SE Model
	4 inputs	5 inputs	6 inputs	
<i>Pre-Salt</i>				
A	Depth, SWOB, RPM, TFLO	ROP5	-	Rabia
B	Depth, SWOB, RPM, TFLO	ROP5	ROP5 and Torque	Teale
D ^(a)	Depth, SWOB, CRPM	ROP5	-	Rabia
H	Depth, SWOB, RPM, TFLO	ROP5	ROP5 and Torque	Teale
<i>Norway</i>				
3-7 ^(a)	DMEA, WOB, RPMB	ROP	ROP and Torque	Teale
8-10	DMEA, WOB, RPMB, MFI	ROP	ROP and Torque	Teale

^(a) the of total flow rate is not available. Therefore, it was left out.

Source: own authorship.

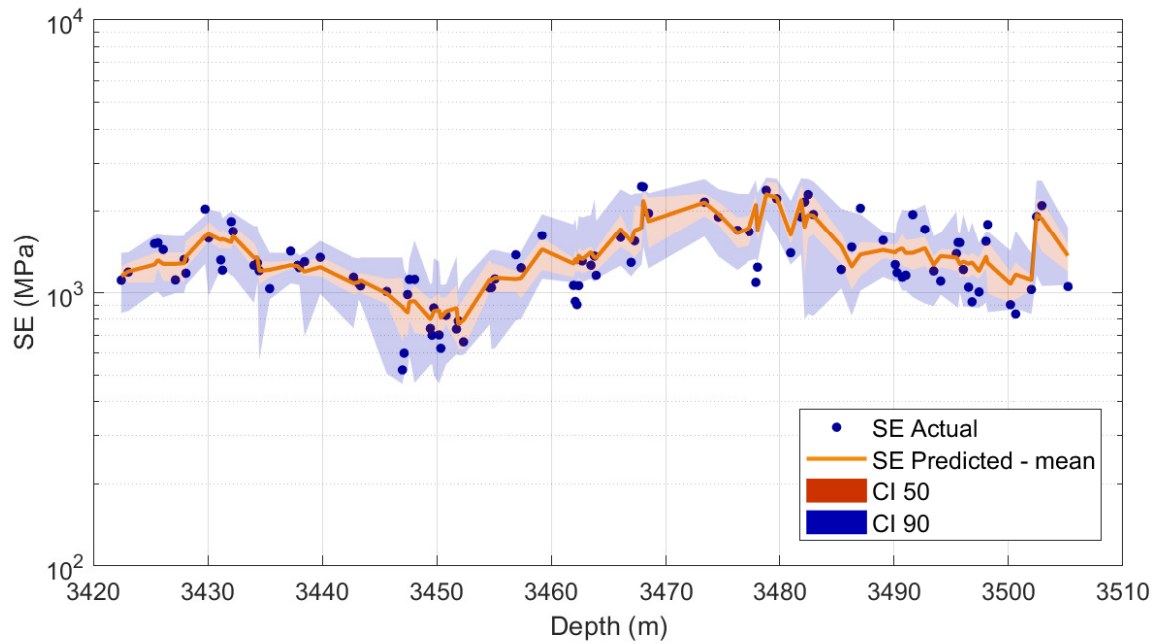
Table 5.10 – SE prediction based on data-driven models without the core variables ROP and torque as inputs: average of evaluation metrics on train and test dataset.

Well	Train dataset				Test dataset			
	MAPE _a	MAE	MSE	R	MAPE _a	MAE	MSE	R
<i>Pre-Salt</i>								
A	0.109	323.286	2.845E+05	0.977	0.175	497.989	6.364E+05	0.951
B	0.097	123.268	2.794E+04	0.932	0.148	186.958	6.305E+04	0.823
D	0.154	623.541	1.566E+06	0.961	0.225	981.082	3.679E+06	0.902
H	0.766	585.716	1.355E+06	0.948	1.179	855.727	2.769E+06	0.882
<i>Norway</i>								
3	0.126	219.523	1.224E+06	0.844	0.183	305.845	2.131E+06	0.650
4	0.038	30.891	3.214E+04	0.880	0.057	44.620	5.518E+04	0.772
5	0.257	423.138	2.966E+06	0.865	0.369	596.903	5.548E+06	0.661
6	0.040	47.247	2.362E+05	0.818	0.055	63.023	3.498E+05	0.670
7	0.050	46.803	1.727E+05	0.866	0.068	62.466	2.827E+05	0.746
8	0.050	138.573	2.993E+05	0.926	0.076	204.556	6.165E+05	0.839
9	0.209	245.887	2.532E+06	0.920	0.291	333.278	4.634E+06	0.810
10	0.094	103.920	3.742E+05	0.870	0.134	147.917	6.812E+05	0.709

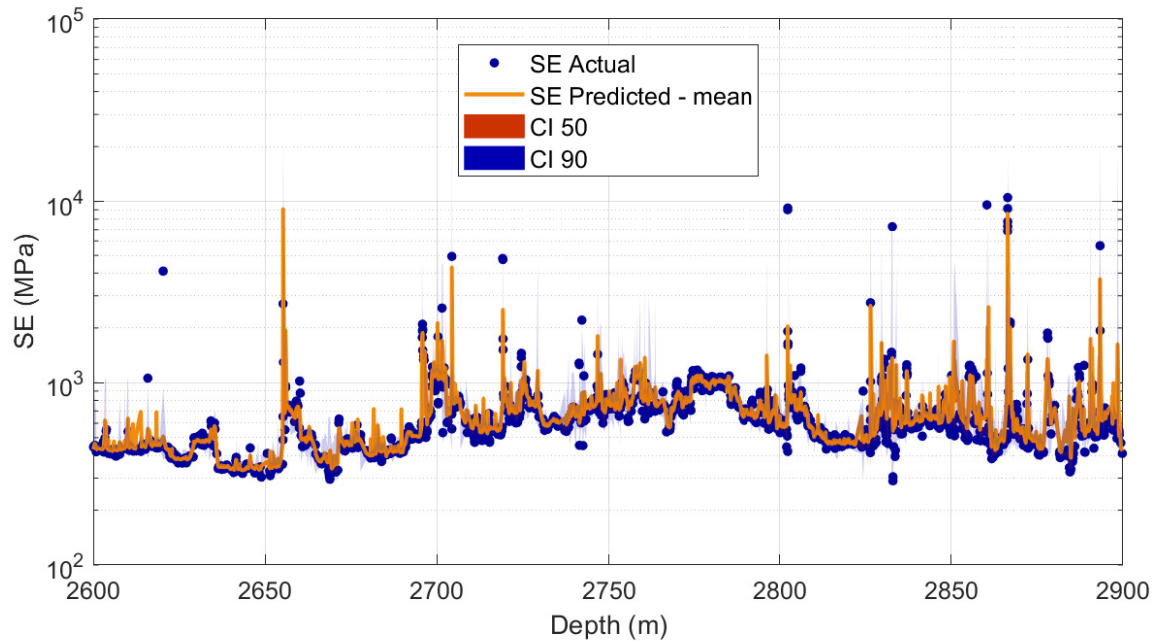
Source: own authorship.

Since random forest were trained with the `TreeBagger` function, it was possible to estimate the mean and the uncertainty around the estimated mean. This function implements the quantile random forests regression, proposed by Meinshausen (2006). In Figure 5.9, we plotted the actual SE against the predicted mean, as well as the prediction intervals of 50% in red and 90% in blue. We used the left-out observations during the learning phase. For clarity, we selected only two wells, one from each region.

Figure 5.9 – SE data-driven models: actual SE against prediction intervals on test dataset.



(a) Well B



(b) Well 10

Source: own authorship.

We observed a high variance of the SE values. This can be a possible reason for not obtaining SE models as accurate as ROP models. By comparing the R-coefficient between datasets from Norway and pre-salt, we observed a better accuracy of SE models for pre-salt region. Perhaps, this difference is due to the data spikes of SE values from Norwegian drill-curves, originated from the automated operation recognition, increasing the calculated errors.

Before testing the coupled SE models, we trained SE-predictive models with more inputs. The extra inputs were the core variables from SE formulation, i.e., ROP and torque. For the case of 5 inputs with only ROP as an extra input, we observed an substantial improvement in the prediction accuracy, as expected, since the ROP is an integral component of SE formulation. As consequence, having ROP as one of inputs improved the prediction of SE. We trained also SE-predictive models based on random forests by adding ROP and torque, the case of 6 inputs.

For both cases (5 and 6 inputs), we followed the same training and testing procedure as we did to obtain the SE models based on exclusively data-driven models with 4 inputs. For clarity, only the evaluation metrics on test datasets are shown in the Table 5.11. We analyzed through T-Test whether adding torque as input increased the model accuracy or not. While adding the torque as input improved the prediction accuracy for some cases (e.g. Well H), for other cases, it did not improve (e.g. Well 5). We observed also some indifferent results (e.g. Well 10). Therefore, a general pattern was not recognized.

Table 5.11 – SE prediction based on data-driven models with the core variables ROP and torque as inputs: average of evaluation metrics on test dataset.

Well	5 inputs (with ROP)				6 inputs (with ROP and Torque)			
	MAPE _a	MAE	MSE	R	MAPE _a	MAE	MSE	R
<i>Pre-Salt</i>								
A	0.149	307.494	2.640E+05	0.980	-	-	-	-
B	0.063	80.751	1.248E+04	0.971	0.065 ≈	81.615 ≈	1.325E+04 ≈	0.971 ≈
D	0.162	442.116	8.338E+05	0.981	-	-	-	-
H	1.060	648.908	1.938E+06	0.919	0.319 ✓	278.769 ✓	5.422E+05 ✓	0.982 ✓
<i>Norway</i>								
3	0.029	54.892	1.548E+05	0.984	0.032 ✗	64.759 ✗	2.260E+05 ✗	0.979 ✗
4	0.014	11.825	1.511E+04	0.959	0.013 ≈	11.930 ≈	1.595E+04 ≈	0.951 ≈
5	0.050	91.643	4.484E+05	0.983	0.065 ✗	113.997 ✗	5.211E+05 ≈	0.980 ≈
6	0.016	24.771	2.049E+05	0.942	0.014 ✓	21.567 ≈	1.926E+05 ≈	0.948 ≈
7	0.022	15.699	3.151E+04	0.983	0.017 ✓	16.500 ≈	4.664E+04 ≈	0.977 ≈
8	0.019	53.269	9.179E+04	0.983	0.020 ✗	57.075 ≈	1.169E+05 ≈	0.978 ✗
9	0.089	127.231	1.033E+06	0.979	0.103 ✗	138.639 ≈	1.072E+06 ≈	0.973 ≈
10	0.041	40.626	5.092E+04	0.987	0.035 ✓	38.305 ≈	7.972E+04 ✗	0.980 ✗

The symbols ✓, ✗, ≈ mean, respectively, that the experiment with 6 inputs resulted in a better prediction of SE than the experiment with 5 inputs, worst, or not statistically different. Tests performed at the 1% significance level. Source: own authorship.

5.2.3 Coupled Models for the Specific Energy

We investigated the approach of coupling the SE formulation with predictive models of the core variables (ROP and torque), as done by Gandelman (2012), Hegde and Gray (2018). To be

more specific, we investigated the calculation accuracy of SE values based on predictive models. For that, two coupled models were analyzed:

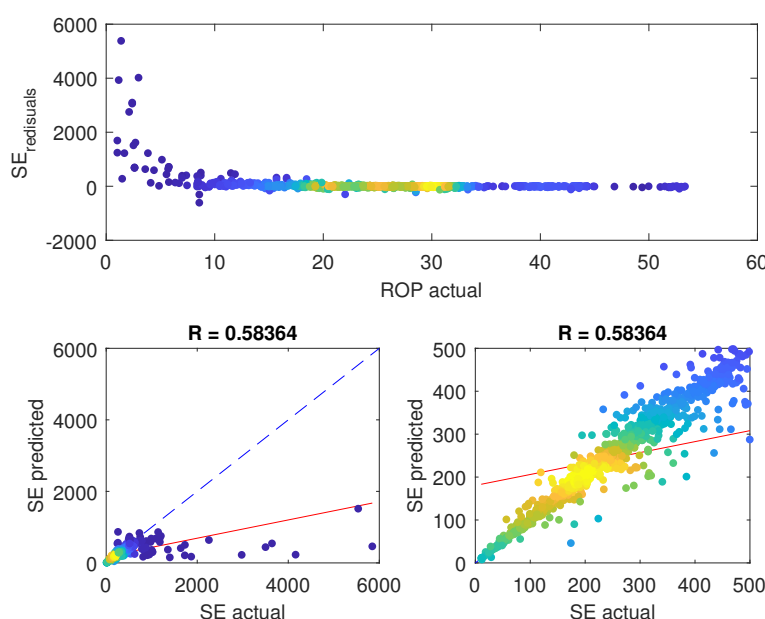
- SE-coupled models based on Rabia's formulation (ROP as coupling model);
- SE-coupled models based on Teale's formulation (ROP and torque as coupling models)

5.2.3.1 Rabia's Model

Two experiments were conducted to assess the SE coupled-models based on Rabia's formulation. In the first one, the SE calculation is coupled with ROP model and, in another, coupled with inverse-ROP model (i.e. with a predictive model of ROP^{-1}).

As already mentioned, the drill-curves from Norway were recorded in time domain. For this reason, a simple method to automatically identify the operations was employed in the current work. This led to some observations considered as rotary drilling operation, but with ROP values near to zero. As consequence, ROP values near to zero yielded too high SE values. For these observations, a small difference between the predicted ROP and the actual ROP can yield high difference between the actual SE (calculated on field data) and predicted SE (calculated on predicted ROP values). The Figure 5.10 illustrates that low values of ROP yielded too high residuals between the actual SE (calculated on field data) and coupled SE (calculated on predicted ROP). In the left bottom plot, we observed that a small quantity of observations with too high residuals corrupted the calculation of R-coefficient. In right bottom plot, we observed that the majority of observations had a small difference between the SE-predicted and SE-actual values.

Figure 5.10 – Density scatter assessing the SE coupled-model based on Rabia's formulation for the Well 10 - coupling with ROP predictive-model.



Source: own authorship. The function `dscatter` developed by Eilers and Goeman (2004) was used.

For this reason, we investigated also an alternative for SE-coupled models. The alternative consists of obtaining predictive models for the inverse-ROP, that is, ROP^{-1} . In this case, the Rabia's formulation for the specific energy becomes:

$$\widehat{SE}_c = 2.35 \left(\frac{WOB.N}{d_b} \right) * \widehat{ROP}^{-1} \quad (5.2)$$

where $\widehat{SE}_c$ denotes the SE-coupled model and $\widehat{ROP}^{-1}$ denotes the predictive model for ROP^{-1} .

For both experiments, we followed the same training and testing procedure of the previous predictive models. We trained random forests with 100 trees, with two variables to be selected at random for each node split. The predictive models were trained to estimate: (i) in the first-case the ROP, (ii) in the second case, the inverse-ROP. The selected inputs for ROP models were the same from Experiment 01 (see Table 5.4), except for Norwegian wells; we left out the torque. Then, ROP-values estimated on predictive models were used to compute the “predicted SE”. We compared the “predicted SE” with “actual SE”, calculated on the field data. Random forests were trained on 80% of observations, and the left-out 20% assessed the generalization capability. We repeated the training and testing procedure 100 times. The comparison between both approaches is presented in the Table 5.12.

Table 5.12 – Evaluation metrics on test datasets for SE coupled-models based on Rabia formulation: ROP predictive model against inverse-ROP predictive model

Well	Coupled with ROP model				Coupled with ROP ⁻¹ model			
	MAPE _a	MAE	MSE	R	MAPE ^(a)	MAE	MSE	R
<i>Pre-Salt</i>								
A	0.105	465.017	3.882E+07	0.952	0.113 ✗	454.044 ≈	3.268E+07 ✓	0.956 ≈
B	0.128	272.384	1.541E+07	0.927	0.134 ✗	265.393 ≈	1.432E+07 ✓	0.929 ≈
D	0.129	844.732	6.484E+08	0.919	0.149 ✗	817.526 ≈	4.794E+08 ✓	0.927 ✓
H	0.173	1331.401	9.487E+09	0.931	0.201 ✗	1137.798 ✓	5.581E+09 ✓	0.952 ✓
<i>Norway</i>								
3	0.092	77.571	2.896E+08	0.483	0.186 ✗	91.761 ✗	2.181E+08 ✓	0.665 ✓
4	0.045	16.321	8.763E+06	0.603	0.059 ✗	17.664 ✗	6.589E+06 ✓	0.710 ✓
5	0.152	120.127	3.752E+08	0.451	0.339 ✗	137.996 ✗	2.977E+08 ≈	0.606 ✓
6	0.031	11.671	3.234E+07	0.692	0.053 ✗	15.376 ✗	3.525E+07 ≈	0.702 ≈
7	0.032	12.174	4.158E+07	0.463	0.064 ✗	14.978 ✗	3.338E+07 ≈	0.579 ✓
8	0.055	82.992	2.119E+08	0.803	0.078 ✗	85.523 ≈	1.501E+08 ✓	0.856 ✓
9	0.119	111.667	3.589E+08	0.383	0.309 ✗	122.419 ✗	2.196E+08 ✓	0.700 ✓
10	0.071	57.883	5.156E+08	0.432	0.138 ✗	66.634 ✗	3.529E+08 ✓	0.684 ✓

^(a) we calculated the normal MAPE instead the adapted MAPE. In this case, the comparison between two approaches based on this parameter is not reliable. We added it only for illustration purpose.

The symbols ✓, ✗, ≈ mean, respectively, that the coupling the SE models with inverse-ROP resulted in a better prediction accuracy of SE than coupling with ROP, worst, or not statistically different. Tests performed at the 1% significance level.

Source: own authorship.

As seen in the Table 5.12, coupling the SE calculation with the inverse-ROP improved the R-coefficient and MSE in the majority of cases. However, the opposite happened for the MAE, i.e., the calculation of SE based on the inverse-ROP decreased the prediction accuracy. We assessed the MAPE differently for both cases: for one case we employed the adjusted MAPE and for another the not-adjusted MAPE. As consequence, a comparison between them is not reliable, and this metric was listed in the Table 5.12 only for illustration purpose.

5.2.3.2 Teale's Model

SE-coupled models based on Teale's formulation requires two coupling-predictive models: ROP and torque. This approach of coupling the SE calculation with two predictive models was adopted by Hegde and Gray (2018). By doing so, the authors considered that changes in bit weight and rotating speed would influence not only the ROP, but also the torque. As consequence, varying the controllable variables (WOB, RPM and mud flow in their case) would influence on ROP, torque and SE as well.

We compared first the actual values of SE calculated on field data with partial-predicted values. The partial-predicted values for SE were calculated on combination of field data and

predicted values only for the parameters ROP and torque². The same train and testing procedure from the previous sections was adopted. The SE-predictive accuracy between both formulations, i.e. Rabia and Teale, were compared. The results are presented in the Table 5.13. In the following section, a discussion is carried out, explaining which of SE-modeling strategies was adopted for the optimization problems.

Table 5.13 – Evaluation metrics on test dataset for SE coupled-models: a comparison between Teale’s and Rabia’s formulation.

Well	Rabia’s formulation coupled with ROP				Teale’s formulation			
	MAPE _a	MAE	MSE	R	MAPE _a	MAE	MSE	R
<i>Pre-Salt</i>								
A	0.105	465.017	3.882E+07	0.952	-	-	-	-
B	0.128	272.384	1.541E+07	0.927	0.154 ✗	211.615 ✓	8.217E+06 ✓	0.765 ✗
D	0.129	844.732	6.484E+08	0.919	-	-	-	-
H	0.173	1331.401	9.487E+09	0.931	1.007 ✗	1046.770 ✓	4.641E+09 ✓	0.812 ✗
<i>Norway</i>								
3	0.092	77.571	2.896E+08	0.483	0.099 ✗	281.781 ✗	3.322E+09 ✗	0.564 ✓
4	0.045	16.321	8.763E+06	0.603	0.048 ✗	42.856 ✗	9.410E+07 ✗	0.745 ✓
5	0.152	120.127	3.752E+08	0.451	0.155 ≈	552.038 ✗	5.645E+09 ✗	0.566 ✓
6	0.031	11.671	3.234E+07	0.692	0.036 ✗	56.448 ✗	6.690E+08 ✗	0.444 ✗
7	0.032	12.174	4.158E+07	0.463	0.037 ✗	54.613 ✗	8.077E+08 ✗	0.644 ✓
8	0.055	82.992	2.119E+08	0.803	0.062 ✗	223.048 ✗	8.699E+08 ✗	0.767 ✗
9	0.119	111.667	3.589E+08	0.383	0.129 ✗	409.596 ✗	7.448E+09 ✗	0.636 ✓
10	0.071	57.883	5.156E+08	0.432	0.094 ✗	145.968 ✗	1.673E+09 ✗	0.491 ✓

The symbols ✓, ✗, ≈ mean, respectively, that the coupled models based on Teale’s formulation resulted in a better prediction accuracy than the coupled models based on Rabia’s formulation, worst, or not statistically different. Tests performed at the 1% significance level.

Source: own authorship.

5.2.4 Final Remark Regarding SE modeling

Two decisions needed to be taken prior running the optimization routines. One was to decide how the SE would be estimated. That is, if we would follow the strategy of data-driven models or coupled models. We chose to estimate the SE based on coupled models, following the previous works from Gandelman (2012), Hegde and Gray (2018). For future works, it is interesting to compare both strategies in the optimization problems.

The second decision was related to the selection of specific energy model. Unlike the Rabia’s formulation, the Teale’s equation considers the torque on SE calculation. Therefore, the current work preferred the Teale’s formulation over Rabia’s formulation for the SE calculation.

² We evaluated the prediction accuracy of torque models, and presented the results in the Appendix F

For those wells with torque-recording available, we employed the Teale's equation, except for the Well H. The reason for it was the low accuracy in torque prediction for the Well H: the average of $MAPE_a$ was around 44% on test dataset (see Appendix F). We believed the reason for this low accuracy was resulted from too high variance of torque recording for this well (see the drill-curves in Figure 4.1). As consequence, high residuals between actual-torque values and predicted-torque values lead to high $MAPE_a$ for the SE-coupled model based on Teale's formulation (see Table 5.13). Therefore, the specific energy for the well H was calculated based on Rabia's formulation in the optimization section.

5.3 DRILLING PARAMETERS OPTIMIZATION

5.3.1 Single-Objective Optimization

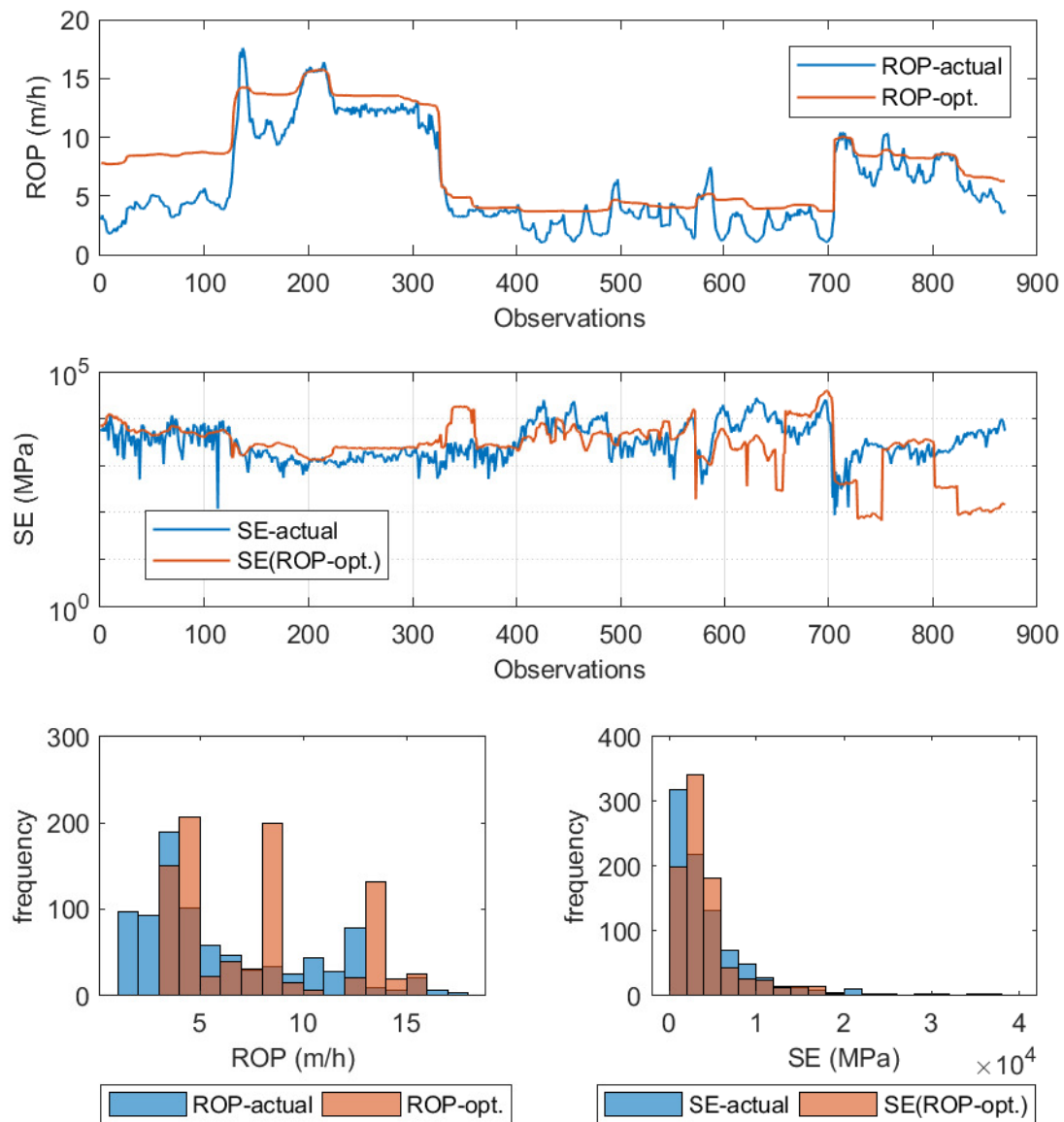
Here, the drilling optimization of a single-objective at time was investigated, comparing maximization of ROP and minimization of SE. The optimization technique employed was the grid-search.

5.3.1.1 First Optimization Problem - maximization of ROP

As seen in the review Chapter 3, a common of approach of real-time drilling optimization was to find an optimum combination of controllable inputs that could maximize the ROP. Therefore, the maximization of ROP is first analyzed. The controllable variables to be optimized were the rotating speed and weight-on-bit. In this case, it was important to set suitable lower and upper bounds on the controllable variables. The searching space is constrained by (i) equipment limits of drill-rigs (GANDELMAN, 2012; HEGDE et al., 2017), (ii) data limits (HEGDE et al., 2017). The minimum and maximum values of both variables were set as lower and upper bounds on the controllable variables. For each variable to be optimized, the searching space was split in equidistant grid points. We employed $k = 98$ grid points for each variable, resulting a total of $(98 + 2)^2 = 10^4$ possible combinations of WOB and RPM for each observations.

It was assessed also how hypothetically optimizing the ROP would influence the SE. According to specific energy formulation, it is expected an decrease of SE values as ROP increases. However, drilling dysfunction(s) may occur with ROP improvement, as consequence, more energy can be required to drill the formation, as explained by Dupriest and Koederitz (2005). This was observed for some drill-curves from pre-salt, as seen in the plots for the Well D (see the Figure 5.11). For some sections, an improvement of ROP would be only possible with an increase in the energy spent, which may not be desired. In the left bottom plot, histograms of ROP-actual values (in blue) are shown, as well as the ROP-optimized values (in red). In the right bottom plot, histograms of SE-actual values are shown together with SE with optimized variables (WOB, RPM and ROP).

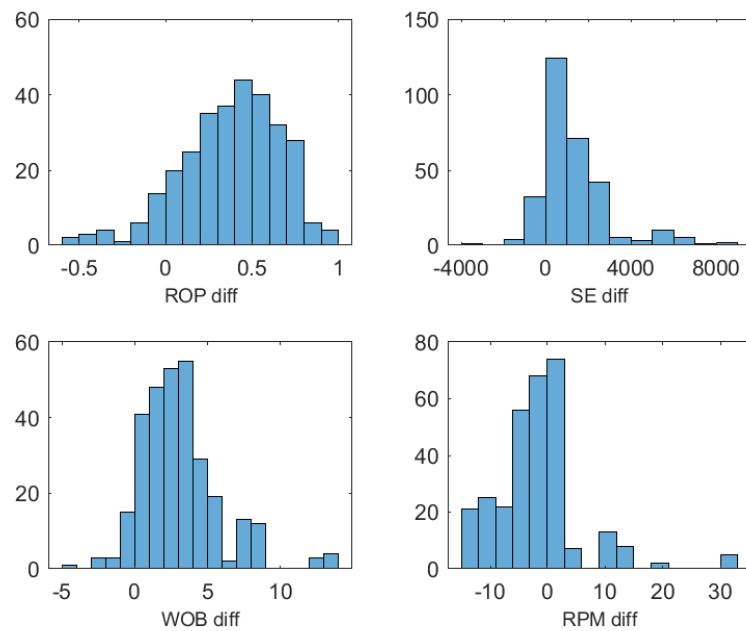
Figure 5.11 – Maximization of ROP and its influence on specific energy - Well D



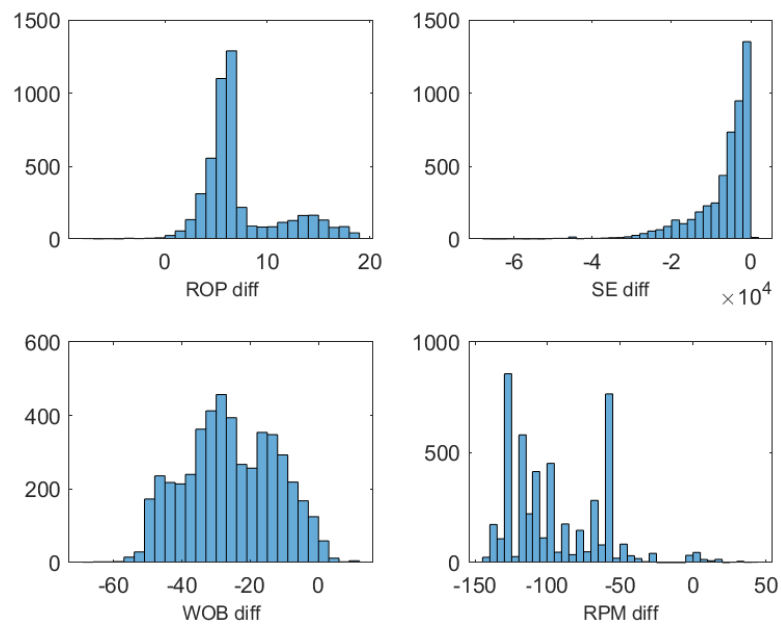
Source: own authorship. The specific energy was calculated based on Rabias' formulation.

In the previous Figure 5.11, the histograms could not provide valuable information, e.g. about how much the ROP has increased. Therefore, we compared, for each instance, the difference between the ROP-optimized values (provided from optimization routine) and ROP-actual values (from drill-curves). In a case of improvement in ROP, this difference is positive. In a case of decrease in ROP, this difference is negative. The same comparison was done for SE, and the variables to be optimized (WOB and RPM). For illustration purpose, two examples are shown in Figure 5.12. One example was for the Well A, where maximizing ROP decreased generally the drilling efficiency. Another example illustrated the opposite, i.e., the improvement of ROP resulted also in a improvement of drilling efficiency. Some statistics related to these differences are given in the Table 5.14 for all wells.

Figure 5.12 – Maximization of ROP: histograms of difference between the optimized parameters and actual values



(a) Well A



(b) Well H

Source: own authorship.

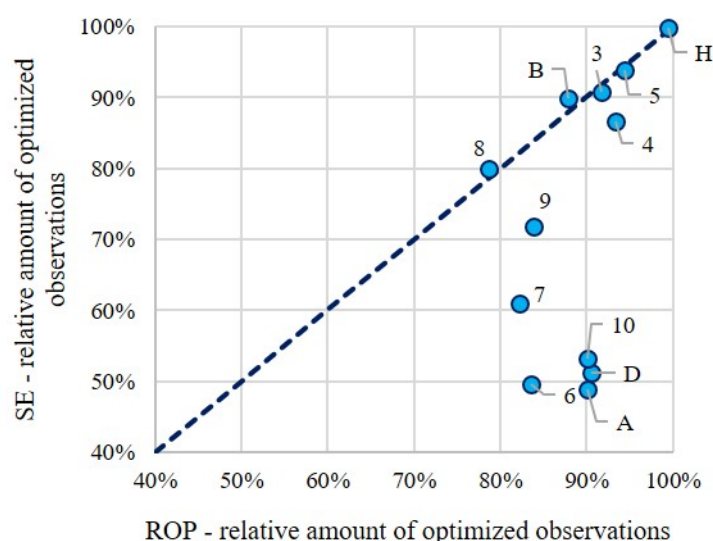
Table 5.14 – Maximization of ROP - some statistics about the difference between optimized and actual-variable values

Well	ROP_opt - ROP_actual				SE_opt - SE_actual				WOB_opt - WOB_actual				RPM_opt - RPM_actual			
	min	mean	median	max	min	mean	median	max	min	mean	median	max	min	mean	median	max
<i>Pre-Salt</i>																
A	-0.54	0.38	0.41	0.93	-6 825	-237	40	4835	-4.8	3.1	2.7	13.7	-14.1	-1.7	-0.1	30.4
B	-1.64	0.66	0.76	1.82	-1 830	-494	-386	352	-31.1	-1.8	-1.0	32.3	-71.5	-17.6	-39.9	70.4
D	-3.36	1.67	1.40	5.92	-25 723	-1615	-59	11 146	-58.3	3.6	-0.5	81.3	-84.3	2.5	0.3	119.4
H	-7.39	7.26	6.16	18.89	-66 828	-6935	-4295	100	-66.5	-25.7	-26.5	11.4	-141.4	-95.4	-103.5	44.8
<i>Norway</i>																
3	-24.71	11.83	9.25	51.84	-34 436	-499	-140	170	-7.2	3.0	2.0	13.8	-55.0	0.4	-0.5	26.6
4	-12.64	8.39	6.45	65.90	-9 544	-116	-57	164	-5.3	0.9	0.3	11.8	-12.9	-1.2	-1.2	130.4
5	-23.62	20.12	17.80	65.39	-4 1078	-813	-230	500	-11.3	2.4	2.6	8.5	-58.0	9.2	2.2	125.3
6	-13.18	1.90	0.81	39.99	-30 513	-47	0	236	-8.7	1.6	0.4	8.6	-22.6	-1.2	0.3	106.4
7	-11.19	6.36	0.99	49.96	-21 452	-165	-5	316	-5.1	1.7	0.9	8.9	-31.3	0.7	0.2	136.0
8	-7.55	1.83	0.75	27.88	-21 120	-464	-165	700	-6.2	1.0	0.7	9.0	-60.6	-7.9	-2.1	51.0
9	-24.60	13.98	8.30	96.83	-55 628	-450	-27	148	-6.9	2.0	1.1	17.1	-21.1	1.2	0.0	73.8
10	-15.55	5.03	3.94	32.59	-26 524	-151	-6	345	-8.3	4.2	4.7	10.3	-43.4	10.0	-0.2	41.7

Source:own authorship.

One statistic interesting to analyze is the relative amount of observations, which could be optimized by the routine of maximization of ROP. The same can be done to assess the improvement in drilling efficiency, expressed by decreases in SE values. In Figure 5.13, we plotted the relative amount of observations with improvement in ROP by the relative amount of observations with drilling efficiency improved as consequence of WOB and RPM optimization. Based on this metric, we observed the best result was for the Well H, because almost all observations could have its ROP and drilling efficiency improved at the same time. However, we observed a cluster with the wells A, D, 6 and 10, where maximization ROP would not necessarily mean an improvement in drilling efficiency for most instances.

Figure 5.13 – Maximization of ROP: relative amount of observations with improvement in ROP versus relative amount of observations with drilling-efficiency improved as consequence of WOB and RPM optimization



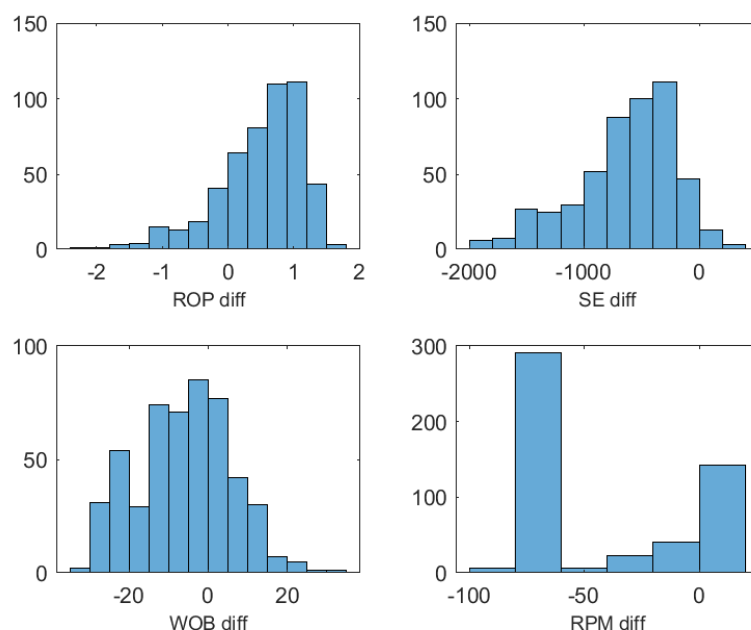
Source: own authorship.

5.3.1.2 Second Optimization Problem - minimization of SE

As done in the work from Hegde and Gray (2018), we investigated also the minimization of SE. We approached it the same way as we did in optimization the ROP. The only difference was the objective function employed. Instead of maximizing the ROP, the objective function was the minimizing of SE. That is, the optimum pair of controllable drilling parameters, WOB and RPM, was the one that resulted in the minimum value of SE for each observation.

For almost all instances, we observed that the grid search strategy could find combinations of WOB and RPM whose SE-estimated values were less than the SE-actual values. While the optimum pair of WOB and RPM resulted also in improvement of ROP for some instances, minimizing SE yielded lower penetration rates for other instances. In Figure 5.14, histograms were plotted, showing the difference between the optimized parameters and actual parameters for the well B.

Figure 5.14 – Minimization of SE: histograms of difference between the optimized parameters and actual values - Well B



Source: own authorship.

Some statistics about the difference between optimized and actual parameters are listed in the Table 5.15. It was observed an decrease in the ROP as consequence of minimizing SE for some wells (e.g. A, 4, 6 and 7). It becomes clearer in the Figure 5.15, where the relative amount of observations with SE improved was plotted against the relative amount of observations with ROP improved as consequence of minimizing SE.

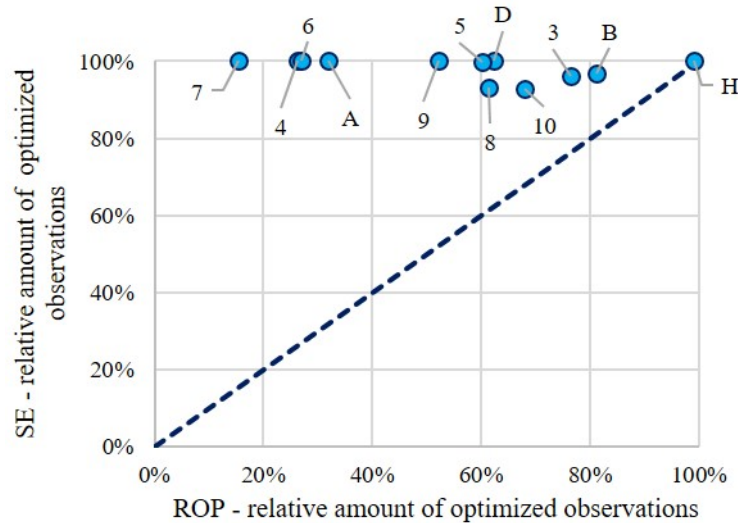
We observed, however, the minimization of SE resulted in unrealistic optimum solutions. That is, low values of WOB and RPM (near to zero) were determined as optimum solutions. In the Section 5.3.2.1, a discussion about this problem is carried out, explaining some possible solutions.

Table 5.15 – Minimization of SE - some statistics about the difference between optimized and actual variable-values

Well	SE_opt - SE_actual				ROP_opt - ROP_actual				WOB_opt - WOB_actual				RPM_opt - RPM_actual			
	min	mean	median	max	min	mean	median	max	min	mean	median	max	min	mean	median	max
<i>Pre-Salt</i>																
A	-12 276	-3 820	-2 917	-10	-1.53	-0.20	-0.15	0.41	-14.8	-8.6	-8.9	0.0	-44.9	-33.3	-30.5	0.0
B	-1 924	-642	-569	219	-2.12	0.50	0.64	1.68	-31.1	-6.3	-5.6	31.0	-80.8	-40.7	-70.3	12.7
D	-26 940	-4 448	-2 665	-86	-6.57	0.50	0.32	4.82	-91.4	-33.0	-29.4	0.0	-181.8	-109.5	-118.6	0.0
H	-66 828	-6 957	-4 313	0	-12.38	7.19	6.14	18.89	-66.5	-27.5	-27.1	0.0	-141.4	-98.0	-104.3	0.0
<i>Norway</i>																
3	-34 579	-525	-165	170	-24.71	10.10	8.07	51.79	-7.2	2.3	1.4	11.2	-55.0	-11.4	-6.8	26.6
4	-9 936	-466	-391	-175	-26.82	-2.02	-3.62	47.14	-4.9	1.1	0.1	10.8	-140.3	-128.1	-128.9	0.0
5	-41 140	-1 068	-454	196	-41.70	7.60	3.61	47.03	-11.1	2.5	3.4	7.8	-126.5	-102.2	-116.1	0.0
6	-30 747	-287	-234	-20	-18.45	-0.49	-0.88	31.37	-9.4	-0.6	-0.5	3.4	-112.1	-104.2	-104.2	0.0
7	-21 789	-455	-299	-28	-28.54	-8.97	-13.24	36.50	-4.7	0.6	0.1	7.8	-164.2	-126.2	-133.4	0.0
8	-21 477	-614	-328	234	-7.58	1.41	0.51	27.69	-10.0	1.3	0.6	9.0	-62.8	-28.7	-13.1	47.9
9	-55 814	-583	-158	-1	-41.61	5.28	0.67	87.94	-10.8	0.5	0.1	17.5	-74.5	-70.8	-72.3	0.0
10	-26 703	-276	-108	176	-15.55	2.95	2.19	30.79	-12.7	2.5	1.7	10.9	-45.4	-23.6	-39.4	1.5

Source:own authorship.

Figure 5.15 – Minimization of SE: relative amount of observations with improvement in SE versus relative amount of observations with ROP improved as consequence of WOB and RPM optimization



Source: own authorship.

5.3.2 Multi-objective optimization

The current work tried to optimize the controllable drilling variables (RPM and WOB) based on the following objective functions:

- minimization of SE;
- maximization of ROP;

In the previous sections, it was observed that the functions SE and ROP could be either conflicting to each other or not. Therefore, it is required to apply the formal formulation of multi-objective optimization problems for those instances that both objectives are not possible to be achieved at the same time. In the following sections, the results from the stepwise algorithms are presented, which were proposed to handle this multi-objective optimization problem with.

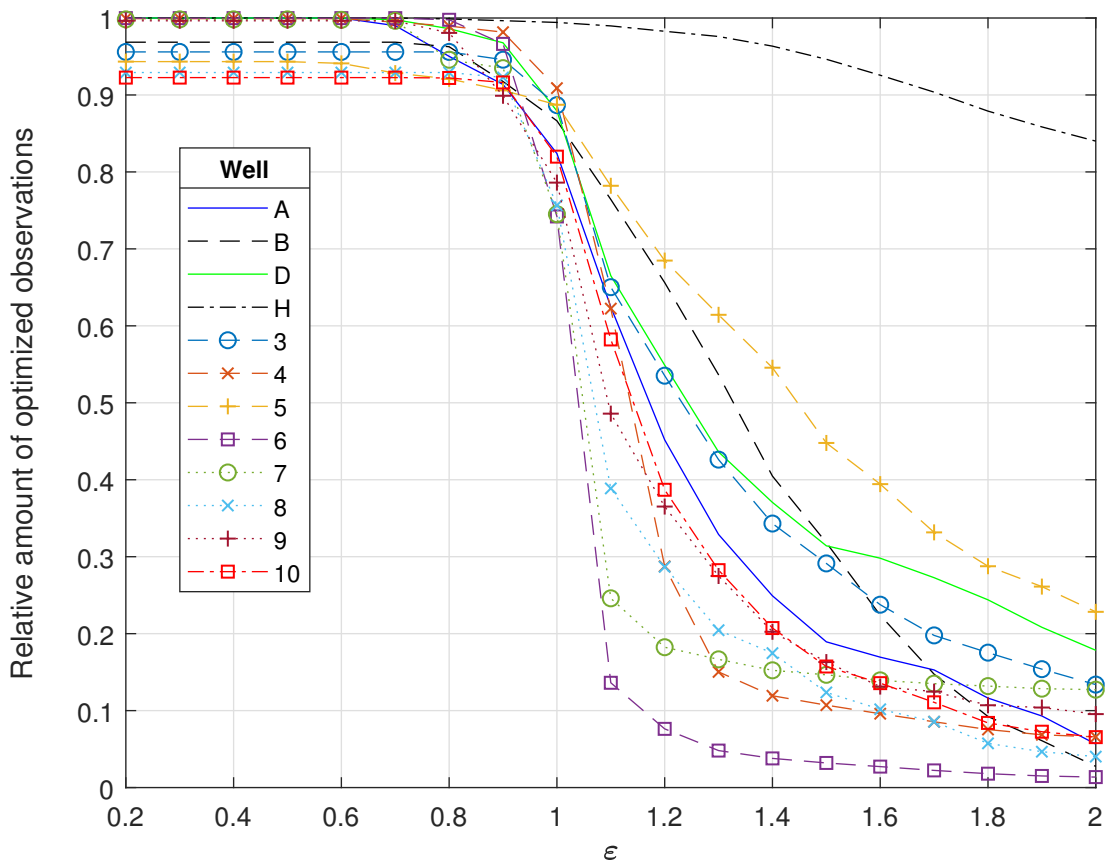
5.3.2.1 Third Optimization Problem - minimization of SE, transforming maximization of ROP into an inequality

The third optimization problem was formulated with the ε -constraint technique, expressed by the equation (4.9). The idea behind the third optimization problem is to minimize the specific energy, transforming the maximization of ROP into an inequality.

The downside of the ε -constraint technique is to determine a suitable ε -value. One way to assess the influence of ε -value on the optimization problem is to count the relative amount of observations that could be optimized, respecting the inequality and lower and upper limits of controllable drilling variables, as seen in Figure 5.16. In this plot, ε is based on the formulation

presented in the equation (4.9). For example, if $\varepsilon = 1$, then the feasible objective space is subject to $ROP(\mathbf{x}^*) \geq ROP_{actual} * \varepsilon$. In this case, an observation is considered optimized if there is at least one optimum solution $\mathbf{x}^*$ whose predicted-SE is less the actual-SE and meets the inequality condition given by $ROP(\mathbf{x}^*) \geq ROP_{actual} * \varepsilon$. The higher the ε is, the more difficult is the task of finding an optimum solution, because the feasible region becomes smaller, as observed in Figure 5.16.

Figure 5.16 – Influence of ε on the relative amount of observations that could be optimized for the problem of minimization of $SE(\mathbf{x})$ subject to $ROP(\mathbf{x}) \geq ROP(\mathbf{x}) * \varepsilon$.



Source: own authorship.

As observed in the Figure 5.16, it is not always possible to find an optimum pair of WOB and RPM that optimizes both objective functions at the same time. For this reason, we developed the stepwise algorithm explained in the Section 4.3.2.1. It starts searching for an optimum solution that can improve the drilling efficiency and drilling rate. If no solution is found, then we search for an solution that is able to minimize the SE, maintaining, at least, the ROP-value. If there is still no optimum solution, then an decrease in ROP-value is accepted. We set the values of $\varepsilon_a = 1.5$, $\varepsilon_b = 1.0$ and $\varepsilon_c = 0.75$. That is, first, we want to minimize the SE and increase the ROP by at least 50%. In the third and last case, an decrease in ROP of up to 25% is accepted.

For the second and third optimization problems (main objective was minimization of SE), we observed unrealistic-optimum solutions. To be more specific, we observed SE-optimum

values near to zero, WOB- and RPM-optimum values near to zero. Such solutions are in fact not realistic, because it is expected that low WOB and RPM results also in low-ROP values. As consequence, both objectives (min SE and max ROP) are expected to not be achieved for too low values of WOB-RPM. Therefore, we analyzed further these optimization problems in the following paragraphs.

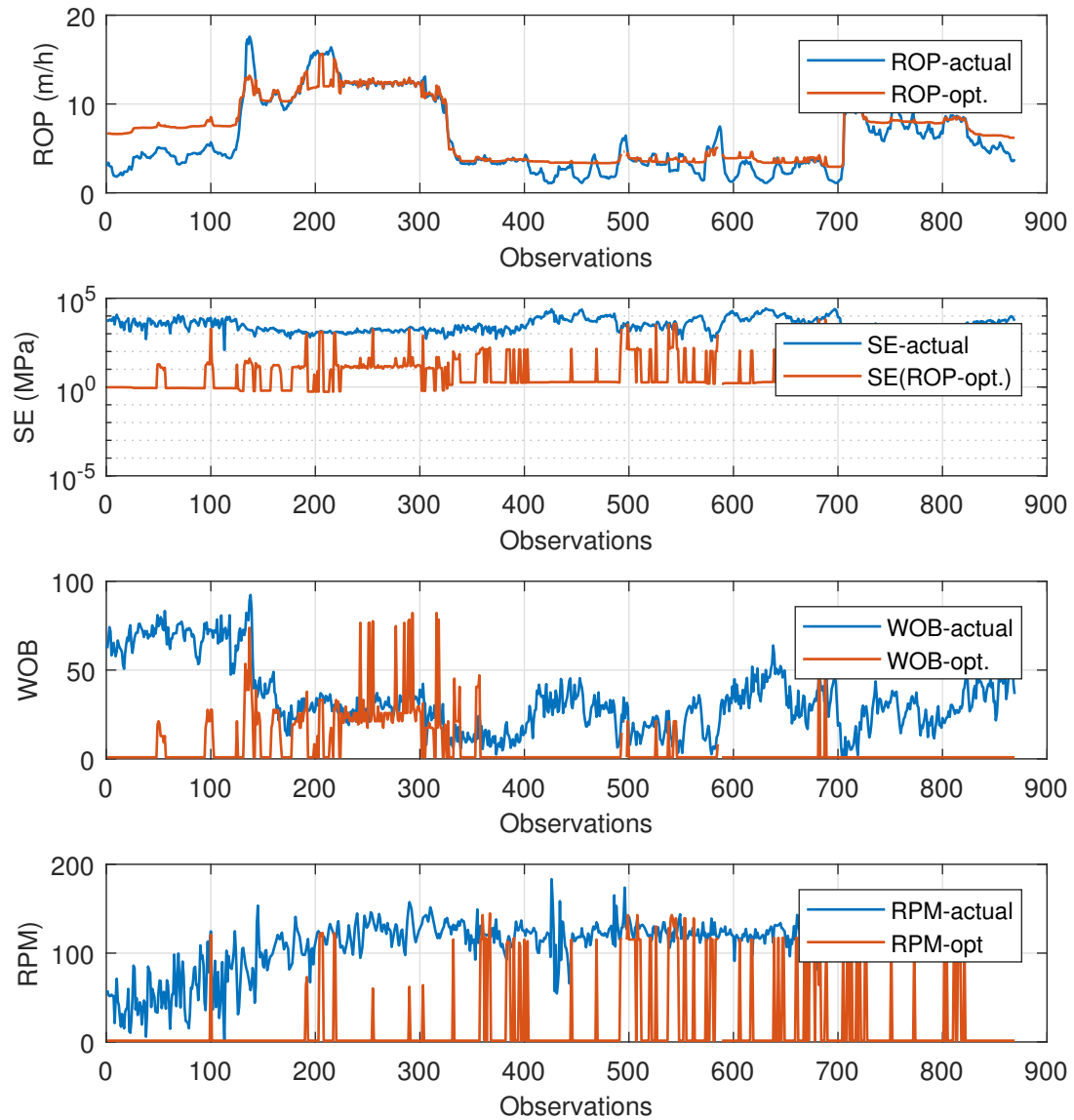
A proposal for new lower-limits in the searching space

The actual values of the drilling parameters, including ROP, SE, WOB and RPM, were analyzed and compared with the optimized values. For example, Figure 5.17 shows in blue the actual values and in red the optimized values for the Well D. We selected this well as an example to illustrate all possible problems occurred in the drilling optimization in second and third problem (main objective was minimization of SE). The problems were: SE-optimum values near to zero and WOB- and RPM-optimum values near to zero.

Before explaining why this problem occurred, it is important to remember, that the SE was estimated based on Rabia's formulation for the wells A, D, H. For some observations, solutions considered as optimum had WOB and RPM near to zero. This might happen because the SE equation from Rabia's formulation is direct proportional to WOB and RPM. If a pair of low WOB and RPM lies in a feasible region Ω , then this pair can be considered as an optimum solution, since its SE value is very low. As consequence, it reaches the goal of minimizing the SE. The Teale's equation behaves similar when given low values of WOB and RPM as inputs. This explains for the second optimization problem (minimization of SE) why low values of WOB-RPM could be considered as optimum solutions

To understand why unrealistic solutions also happened in the third optimization problem, it is required to look at the predictive models for the ROP. The predictive models did not predict low ROP values when given low WOB and RPM. Let us suppose the following example. A ROP model learned that an instance (e.g. with depth of 1000m) had an output y given an pair of WOB and RPM as input. However, this model did not actually learn how this output would be for the same depth (e.g. 1 000m) with a very different pair of WOB and RPM, even this pair occurred in a different section (e.g. 1 200m) and was used in the learning phase, being therefore considered as a feasible one. Since the depth is also used as input, the learned ROP model may predict an unrealistic ROP value for this different pair of WOB and RPM. This unrealistic value might be e.g. too far away from 0 m/h.

Figure 5.17 – Minimization of $SE(x)$ subject to $ROP(x) \geq ROP(x) * \varepsilon$ - well D with not realistic optimum solutions.



Source: own authorship.

This shows a special needs to taken when predicting ROP values with very different inputs from the training phase. Some possible solutions to avoid unrealistic optimum solutions are:

- to adopt new lower limits for WOB and RPM, avoiding too low-values of WOB and RPM;
- to adopt a moving lower and upper bounds for the WOB and RPM. This can avoid the problem of predicting values with very different inputs used in the learning phase;
- to employ data-driven models to estimate SE, instead of using SE coupled-models as done in the current work;

- to add another constraint in the optimization problem, for example, $SE(x) > UCS$. By imposing this inequality, it is not accept optimum solution with estimated SE-values (i.e., the energy to destroy the rock) less than the rock strength, which is physically not possible;
- since there is no standard in the literature regarding data pre-treatment of drill curves, it always worth of investigating other methods for data cleanness.

For simplicity, only one strategy was investigated, which consists of avoiding values of WOB and RPM near to zero. In the Table 5.16, the current lower and upper limits employed until now are listed, as well as the new proposed limits. To determine the new lower bounds, we visually analyzed the results from all wells, as we did in the Figure 5.17. A summary of the detected problems is also given in the Table 5.16, when using the current limits employed until here.

Table 5.16 – Problems identified in the third optimization problem

Well	Optimized values near to zero			Current limits - range		New limits	
	WOB	RPM	SE	RPM	WOB	RPM min	WOB min
<i>Pre-Salt</i>							
A	x			[80.1 - 125.0]	[0.4 - 15.2]	81.0	3.0
B	x			[78.2 - 159.0]	[0.1 - 34.2]	79.0	3.0
D	x	x	x	[1.5 - 183.3]	[0.9 - 92.3]	50.0	5.0
H	x	x	x	[30.6 - 171.9]	[0.2 - 66.7]	50.0	5.0
<i>Norway</i>							
3				[219.9 - 278.4]	[0.4 - 14.5]	220.0	1.5
4		x		[29.3 - 169.6]	[0.1 - 15.1]	50.0	1.5
5				[167.5 - 294.0]	[0.0 - 12.1]	168.0	1.0
6				[192.1 - 304.2]	[0.0 - 10.4]	193.0	1.0
7		x		[16.4 - 180.6]	[0.0 - 10.1]	50.0	1.0
8	x			[227.3 - 290.6]	[0.0 - 11.8]	228.0	1.0
9				[117.2 - 192.7]	[0.1 - 20.0]	118.0	2.0
10				[118.8 - 164.3]	[0.0 - 15.7]	119.0	1.5

Source: own authorship.

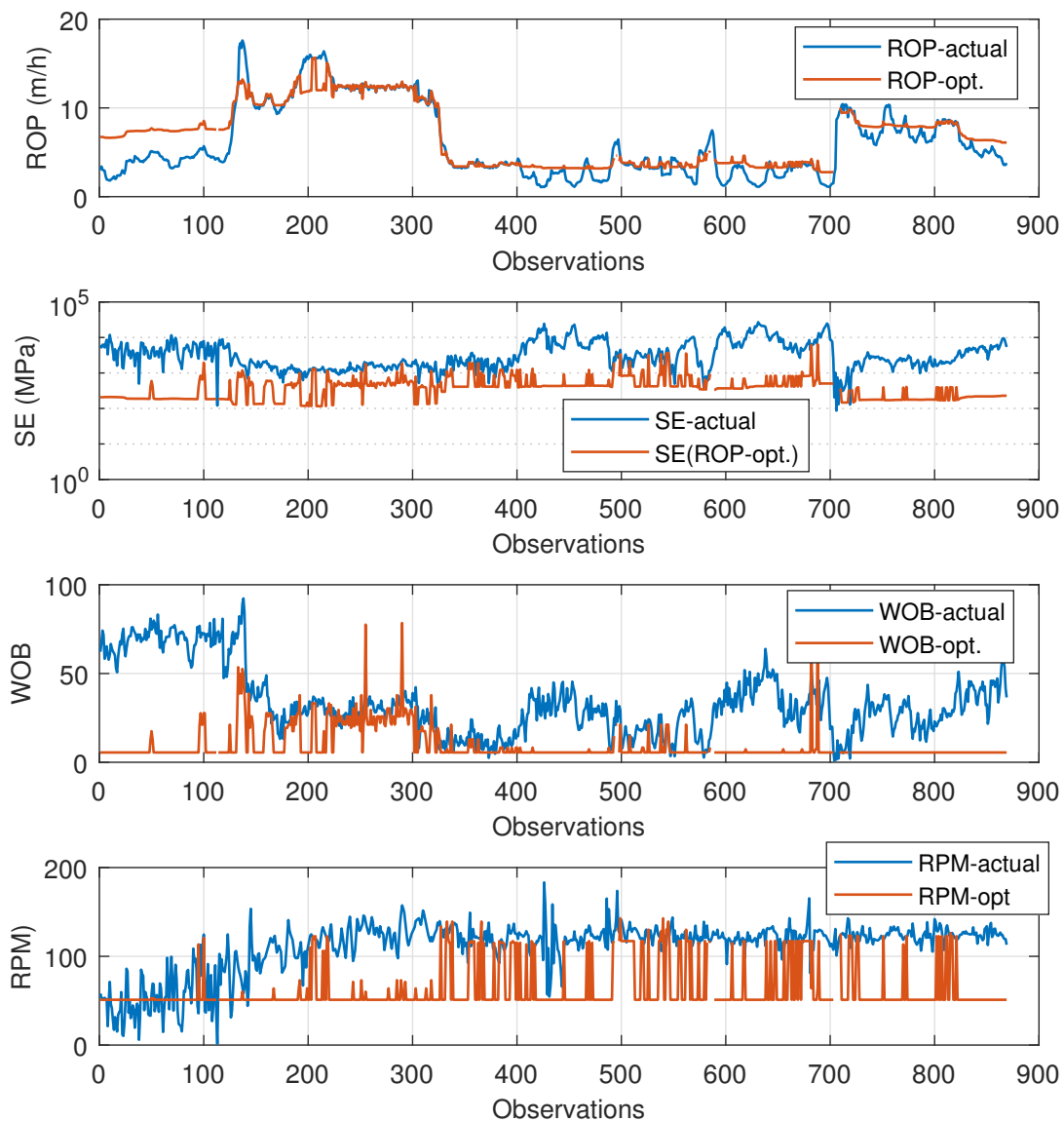
Simulations based on new lower limits in search space

To speed up the simulations based on new lower bounds, the grid-search was not performed again. Instead, we simply left out those combinations of WOB and RPM out of the new range. As consequence, the matrixes of SE and RPM were adapted to the new limits. This enabled to speed up the analysis, since it was not required to map again how several different pairs of WOB-RPM would influence on the ROP and SE.

By assessing the influence of ε on the amount of observations optimized, it was noticed a very similar behaviour from the Figure 5.16. Therefore, this plot with new lower bounds was omitted.

In the Figure 5.18, the optimum solutions was plotted against the actual-values again for the well D. Some differences can be observed between the solution with new limits (Figure 5.18) and the previous solutions (Figure 5.17). First, SE-optimized values were not near to zero anymore. This occurred because the minimum-allowed values of RPM and WOB for the well D were 50 rpm and 5 klbf, respectively. However, it is possible to note an preference on the minimum values for WOB and RPM as optimum solutions. In the Table 5.17, some statistics are listed, presenting the results for all wells.

Figure 5.18 – Minimization of $SE(x)$ subject to $ROP(x) \geq ROP(x) * \varepsilon$ - optimum solutions for the well D with new lower limits.



Source: own authorship.

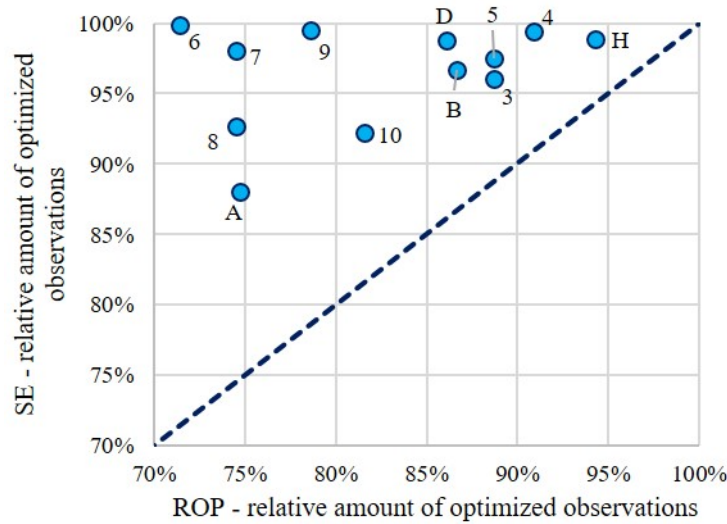
Table 5.17 – Minimization of $SE(\mathbf{x})$ subject to $ROP(\mathbf{x}) \geq ROP(\mathbf{x}) * \varepsilon$ - some statistics about the difference between optimized and actual variable-values

Well	Relative amount of optimized observations				SE_opt - SE_actual		ROP_opt - ROP_actual		WOB_opt - WOB_actual		RPM_opt - RPM_actual	
	$\varepsilon_a = 0.75$	$\varepsilon_b = 1$	$\varepsilon_c = 1.25$	not-optimized	mean	median	mean	median	mean	median	mean	median
<i>Pre-Salt</i>												
A	18.9%	55.8%	13.3%	12.0%	-2821	-1974	0.09	0.05	-5.0	-5.3	-9.8	-8.6
B	31.6%	55.0%	10.0%	3.3%	-654	-569	0.58	0.68	-5.9	-5.6	-39.4	-69.3
D	31.0%	55.1%	12.7%	1.3%	-4018	-2280	0.76	0.35	-24.5	-22.2	-48.3	-64.3
H	73.8%	20.5%	4.6%	1.2%	-6647	-3829	2.15	2.15	-18.5	-20.2	-70.0	-79.5
<i>Norway</i>												
3	29.1%	59.6%	6.9%	4.5%	-541	-175	11.22	8.46	2.5	1.6	-5.9	-0.9
4	10.7%	80.2%	8.4%	0.7%	-259	-141	4.33	2.02	0.9	-0.3	-49.2	-31.9
5	44.8%	43.9%	3.5%	7.8%	-988	-373	14.08	14.24	2.5	2.8	-48.8	-58.4
6	3.2%	68.2%	28.4%	0.1%	-188	-168	0.26	0.13	-0.2	-0.2	-57.2	-96.1
7	14.6%	59.9%	23.4%	2.1%	-232	-55	1.88	0.37	1.0	0.4	-28.7	-27.4
8	12.4%	62.1%	17.9%	7.6%	-624	-328	1.69	0.77	1.5	0.9	-24.1	-11.7
9	16.3%	62.3%	20.6%	0.8%	-555	-137	7.43	3.43	0.9	0.4	-54.3	-70.6
10	15.7%	65.8%	10.2%	8.3%	-291	-105	3.71	2.67	3.1	2.4	-21.9	-38.7

Source: own authorship.

By imposing the inequality $ROP(x) \geq ROP_{actual} * \varepsilon$ on the minimization problem of SE, it was observed an increase in the amount of optimum solutions that could also improve the ROP. In Figure 5.19, we observe that, for all wells, more than 70% of observations could have its SE and ROP improved at the same time. In the problem of solely minimizing SE, many observations would have its ROP decreased (see the previous Figure 5.15 from the second optimization problem).

Figure 5.19 – Minimization of $SE(x)$ subject to $ROP(x) \geq ROP(x) * \varepsilon$ - relative amount of optimized observations.



Source: own authorship.

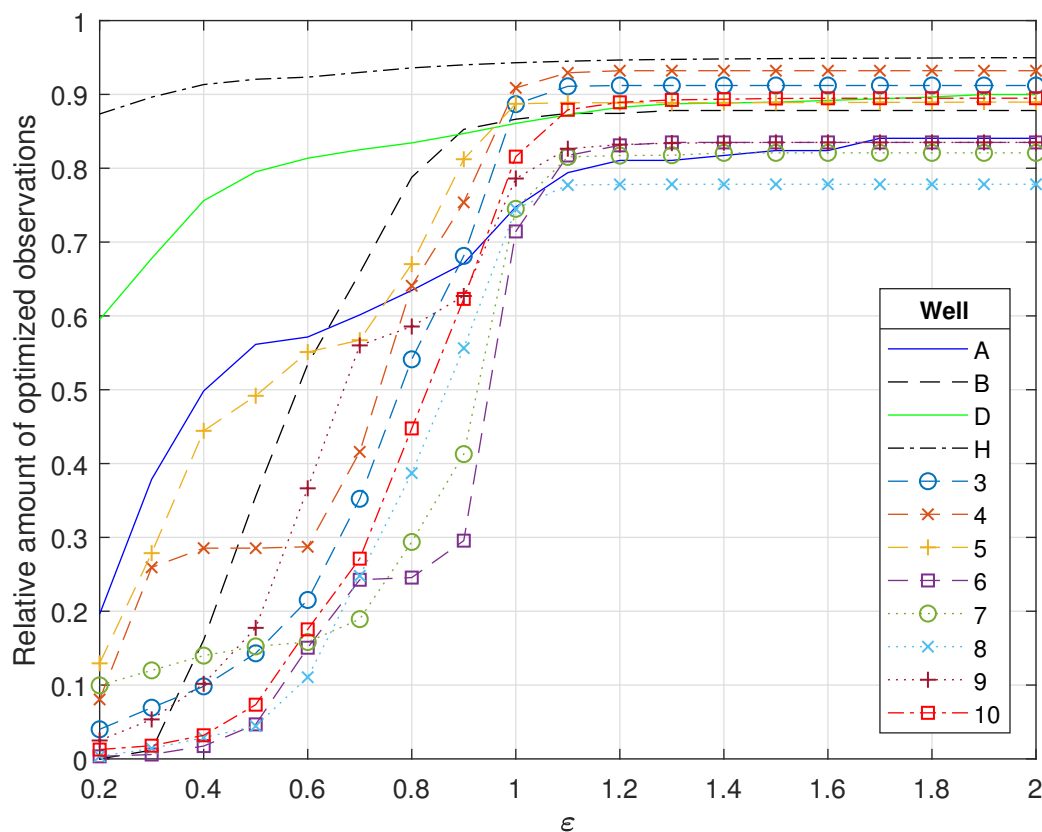
5.3.2.2 Fourth Optimization Problem - maximization of ROP, transforming minimization of SE into an inequality

Another multi-objective optimization problem was investigated. At this time, the main goal was to maximize the ROP, while the minimization of SE was transformed into an inequality. The new lower limits for WOB and RPM, given in the Table 5.16, were employed in this section. First, it was assessed the influence of ε on the amount of observations that could be optimized. As the ε decreases and becomes less than one, finding optimum solutions able to improve the ROP and drilling efficiency at the same time becomes more difficult, as seen in the Figure 5.20.

Since it is not always possible to find an optimum pair of WOB and RPM able to optimize both functions at the same time, a stepwise algorithm for this fourth optimization problem was developed (see the Section 4.3.2.2). It starts searching for an optimum solutions that can improve the ROP and drilling efficiency at the same time. If no solution is found, then we search for an solution that is able to improve the ROP, maintaining, at least, the current SE value. If there is still no optimum solution, then an decrease in drilling efficiency (expressed by an increase SE value) is accepted. In the fourth optimization problem, we set the values of $\varepsilon_a = 0.75$, $\varepsilon_b = 1.0$ and $\varepsilon_c = 1.25$. That is, first, we want to maximize the ROP and increase the drilling efficiency

by at least 25%. In the third and last case, it is accepted and decrease in drilling efficiency up to 25%.

Figure 5.20 – Influence of ε on the relative amount of observations that could be optimized for the problem of maximization of $ROP(\mathbf{x})$ subject to $SE(\mathbf{x}) \leq SE(\mathbf{x}) * \varepsilon$

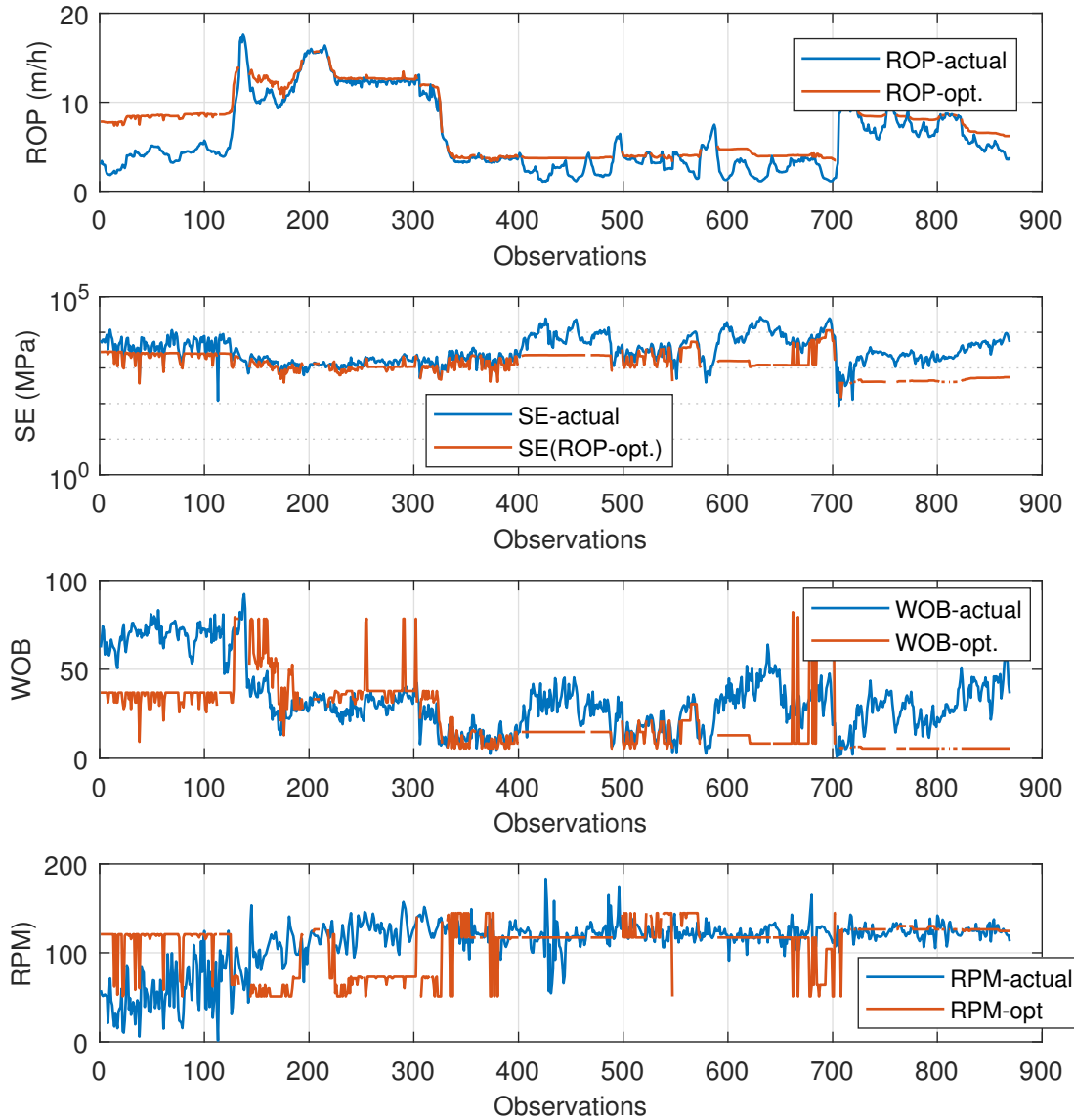


Source: own authorship.

In the Figure 5.18, the optimum solutions was plotted against the actual values again for the well D. Some differences can be observed between the first optimization problem (solely maximization of ROP - see the Figure 5.11) and this fourth optimization problem (maximization of ROP, transforming the minimization of SE into an inequality).

An interesting fact occurred for the observations from 0 to 100 and, partially, from 400-700. For such observations, substantial improvement in ROP would be possible without having to deteriorate the drilling efficiency, after implementing the stepwise algorithm in the fourth optimization problem with the constraint $SE(\mathbf{x}) \leq SE_{actual} * \varepsilon$ in the objective space. In the first optimization problem (only maximization of ROP), the optimum solutions of WOB-RPM would provide higher ROP values than the fourth problem, but the drilling efficiency would decrease.

Figure 5.21 – Maximization of $ROP(x)$ subject to $SE(x) \leq SE(x) * \varepsilon$ - optimum solutions for the well D



Source: own authorship.

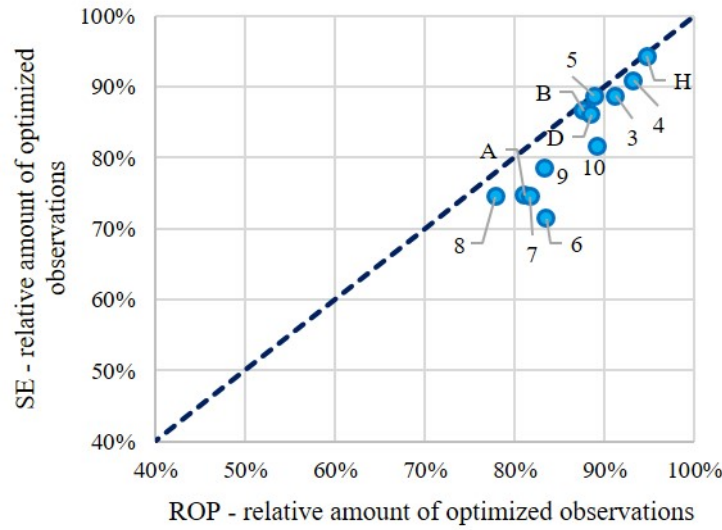
The results from the fourth optimization problem are presented in the Table 5.18. Differently from the third problem, fewer observations could not be optimized in the fourth problem. By analyzing the optimum RPM-values, we observed the third optimization problem led to greater reduction of rotating speed than the fourth optimization problem. In the Figure 5.22, we observe that imposing the inequality $SE(x) \leq SE_{actual} * \varepsilon$ increased the number observations that could have the ROP and SE improved at the same time, in comparison with the first optimization problem (only maximization of ROP - see Figure 5.13).

Table 5.18 – Maximization of $ROP(\mathbf{x})$ subject to $SE(\mathbf{x}) \leq SE(\mathbf{x}) * \varepsilon$ - some statistics about the difference between optimized and actual variable-values

Well	Relative amount of optimized observations				ROP_opt - ROP_actual		SE_opt - SE_actual		WOB_opt - WOB_actual		RPM_opt - RPM_actual	
	$\varepsilon_a = 0.75$	$\varepsilon_b = 1$	$\varepsilon_c = 1.25$	not-optimized	mean	median	mean	median	mean	median	mean	median
<i>Pre-Salt</i>												
A	61.5%	13.3%	6.3%	18.9%	0.29	0.20	-1405	-1452	-0.3	0.1	-4.5	-5.1
B	73.3%	13.4%	1.0%	12.4%	0.81	0.85	-580	-460	-0.9	0.3	-21.2	-40.2
D	83.1%	3.0%	2.4%	11.5%	1.65	1.26	-3114	-1524	-11.5	-12.3	-2.2	-0.8
H	93.2%	1.1%	0.4%	5.3%	3.24	3.29	-4848	-2336	6.8	12.6	-14.2	-9.4
<i>Norway</i>												
3	43.4%	45.3%	2.5%	8.8%	12.79	9.82	-548	-169	3.1	2.6	-0.4	-0.5
4	55.8%	35.1%	2.3%	6.8%	7.60	4.86	-197	-126	0.9	-0.3	-24.2	-4.9
5	60.6%	28.1%	0.1%	11.1%	19.00	17.81	-881	-279	2.4	2.7	2.4	0.8
6	24.3%	47.2%	12.1%	16.5%	1.81	0.58	-133	-24	0.4	0.1	-25.2	-0.6
7	21.5%	53.0%	7.2%	18.3%	7.43	1.29	-219	-15	1.8	0.9	-3.8	-1.0
8	32.4%	42.1%	3.3%	22.2%	2.36	1.28	-636	-354	1.7	1.2	-12.9	-11.4
9	57.9%	20.7%	4.8%	16.6%	15.01	6.75	-593	-133	2.2	1.5	-26.3	-1.4
10	33.9%	47.7%	7.6%	10.9%	5.24	3.80	-229	-39	4.1	4.5	-3.3	-0.6

Source: own authorship.

Figure 5.22 – Maximization of $ROP(x)$ subject to $SE(x) \leq SE(x) * \varepsilon$ - relative amount of optimized observations.



Source: own authorship.

5.3.3 On the Implementation's Feasibility of Proposed Optimization Methods

Both multi-objective optimization problems proposed here require the actual values of an instance to determine the optimum solution. In theory, this is not possible, since the optimization process should be carried out before the implementation, as consequence, before knowing the actual values. However, an alternative can be applied in practice, which consists of estimating suitable values for ROP and SE to be used in the comparison steps in the optimization routine (i.e., third and fourth optimization problems). Such suitable values can be obtained based on previous instances, e.g., through a simple moving averaging or any forecasting method. For sake of simplicity, the current work did not deal with such problem, and assumed this information is already available for the optimization algorithm.

To overcome the mentioned challenge, Hegde and Gray (2018) formulated single-objective problems, which do not require any comparison. Gandelman (2012) proposed a target value for ROP, which could be set by a driller and was used in his searching algorithm based on *if-then* rules.

The elapsed time for the optimization algorithm was accessed to estimate its feasibility. It was measured the required time to train the predictive models, as well as the required time to perform the grid-search for each observation. The results are listed in the Table 5.19. In the first two columns, the elapsed time was computed considering both data pre-treatment process and training time. For that, we ran a simulation apart from the optimization code, but with the same configuration used in that routines. For each well, we trained ten times the predictive models, considering the whole process, i.e., from the data pre-treatment to the training phase. The average of elapsed time are listed in these first two columns. Even if another data partition had been employed (e.g. the strategy proposed by Hegde et al. (2017), Soares and Gray (2018) to

obtain predictive models in real-time), the training time would not have been restrictive, because this step took less than a few seconds.

Table 5.19 – Elapsed time in seconds for training phase (including data pre-treatment) and grid-search in the optimization algorithm

Well	Average train time (seconds)		Grid-search (seconds/observations)		
	ROP	Torque	ROP	Torque	SE
<i>Pre-Salt</i>					
A	0.241	-	0.479	-	0.006
B	0.311	0.641	0.498	0.498	0.007
D	0.445	-	0.476	-	0.007
H	1.455	-	0.408	-	0.005
<i>Norway</i>					
3	1.684	1.778	0.508	0.5083	0.005
4	2.414	2.604	0.491	0.4914	0.004
5	1.122	1.249	0.473	0.4733	0.005
6	2.063	1.992	0.503	0.5031	0.004
7	3.065	3.272	0.513	0.5131	0.004
8	1.620	1.609	0.510	0.5104	0.006
9	1.028	1.081	0.512	0.5124	0.005
10	2.569	2.715	0.556	0.5557	0.005

Source: own authorship.

The most time intensive aspect of developed algorithms was the grid-search, since 10^4 possible combinations of WOB and RPM were tested for each objective function and each instance. By analyzing the elapsed time to perform the grid-search, we observed that around 0.5 second was required to predicted all possible combinations for each predictive model (ROP and/or torque). As consequence, the evaluation of a considerable amount of combinations did not take longer than 1.2 second per observation. As discussed by Hegde, Daigle and Gray (2018, p. 9), “the frequency of drilling-parameter change can be in the range of every 50 to 100 ft of drilling or every 10 minutes of drilling”. Therefore, the elapsed time to perform the grid-search would not hinder the implementation of this searching strategy, because the required time to map the decision space into the objective space would cause a delay of only a second in most cases.

6 CONCLUSIONS

The current work investigated the possibility of employing machine learning algorithms to optimize drilling operational variables, especially weight-on-bit (WOB) and rotating drillbit speed (RPM), considering maximization of rate of penetration (ROP) and/or minimization of specific energy (SE).

For that, real-time drilling data from pre-salt region and Norwegian continental shelf were employed. Random forests could learn the complex relationship among the drilling variables to estimate ROP, as expected according to the literature review carried out in the chapter 3. It was observed that using surface operational variables as inputs (depth, WOB, RPM and mud flow) resulted in predictive models with acceptable accuracy to predict ROP. Such variables are easily accessed at drill rigs. Besides that, the employment of them as inputs in the predictive models enables the optimization of them, as already in some previous works (GANDELMAN, 2012; HEGDE; GRAY, 2017; HEGDE; DAIGLE; GRAY, 2018; HEGDE; GRAY, 2018). The use of binary genetic algorithm was employed to obtain the best subset of inputs. However, the obtained optimum subsets of inputs could not be used further in the drilling optimization studies.

Extensive analysis was done regarding the prediction of specific energy (SE). To be more specific, two approaches were compared: one used exclusively data-driven models, as commonly done to estimate ROP, and another with coupled models, following the previous works from Gandelman (2012), Hegde and Gray (2018). Data-driven models could very accurately estimate the SE after adding the ROP as input, as expected according to the Teale's formulation. We assessed also the accuracy of coupling the SE calculation with predictive models for ROP and torque (if available). We observed, especially for Norwegian drill-curves, high residuals between actual values and estimated values for those observations whose ROP-values were near to zero. For the optimization problems, we followed the approach adopted by previous works (GANDELMAN, 2012; HEGDE; GRAY, 2018) to estimate the SE based on coupled models. For future works, it is interesting to investigate the employment of data-driven models to map the influence of drilling parameters on SE.

The obtained predictive models for ROP and SE were employed in optimization problems as objective functions. Four optimization problems were formulated: two in single-objective framework and other two in multi-objective framework. It was observed that maximization of ROP alone was not necessary a good optimization strategy, because the drilling efficiency could decrease (or more energy could be required). By imposing an inequality of $SE(\mathbf{x}) \leq SE_{actual} * \varepsilon$ on maximization of ROP, this problem could be avoided in many cases. The implementation of the stepwise strategy, which consisted of varying ε value, was tested with three ε -values. As extension of the proposed algorithm, it is possible to implement the same idea in an iterative loop, assessing several ε -values. By doing so, it is expected to increase the amount of observations whose optimum solutions are able to improve both objective functions. However, it is necessary

to assess the computational effort.

The current work focused on optimizing two controllable drilling variables (WOB and RPM). However, the stepwise algorithm can be employed also to optimize more drilling variables at the same time, for example by adding the mud flow to WOB-RPM, as done in some works (AWOTUNDE; MUTASIEM, 2014; HEGDE; GRAY, 2017; HEGDE; GRAY, 2018). It is possible to add other parameters to be optimizes, such as mud weight (BATAEE; IRAWAN; KAMYAB, 2014), bit hydraulic (ARABJAMALOEI; SHADIZADEH, 2011) or any other hydraulic/drilling fluid parameters.

For future works, more efforts are required to overcome the problem of having, as optimum, solutions pairs of WOB and RPM near to zero when the minimization of SE is employed as objective function. The present thesis investigated one possible solution for that, which consisted of increasing the lower bounds for WOB and RPM in order to avoid too low values. However, it is worth of studying other strategies regarding the definition of searching space, including the employment of a moving widow to determine the lower and upper bounds for WOB and RPM. A similar alternative can be adopted by adding another constraint in the objective space. This constraint can be formulated as following: the energy spent to drill a rock formation can be not less than the rock strength. Besides that, another solution can be the employment of data-driven models to map how WOB and RPM influence on objective space of SE, since random forests could accurately estimate the SE. Finally, investigating other methods for data cleanness is always worth, because there is no standard in the literature regarding data pre-treatment of drill curves.

REFERENCES

- AGGARWAL, C. C. **Outlier Analysis**. New York, NY: Springer New York, 2013.
- AGWU, O. E. et al. Artificial intelligence techniques and their applications in drilling fluid engineering: A review. **Journal of Petroleum Science and Engineering**, v. 167, p. 300–315, 2018.
- AL-SUDANI, J. A. Real-time monitoring of mechanical specific energy and bit wear using control engineering systems. **Journal of Petroleum Science and Engineering**, v. 149, p. 171–182, 2017.
- AMADI, W. K.; IYALLA, I. Application of mechanical specific energy techniques in reducing drilling cost in deepwater development. In: SPE DEEPWATER DRILLING AND COMPLETIONS CONFERENCE, 2012, Galveston, Texas. **Proceedings...** Galveston, Texas: Society of Petroleum Engineers, 2012.
- AMAR, K.; IBRAHIM, A. Rate of penetration prediction and optimization using advances in artificial neural networks, a comparative study. In: INTERNATIONAL JOINT CONFERENCE ON COMPUTATIONAL INTELLIGENCE (NCTA), 4., 2012, Barcelona. **Proceedings...** Barcelona: SciTePress, 2012. p. 647–652.
- ANEMANGELY, M. et al. Drilling rate prediction from petrophysical logs and mud logging data using an optimized multilayer perceptron neural network. **Journal of Geophysics and Engineering**, IOP Publishing, v. 15, n. 4, p. 1146–1159, 2018.
- ANSARI, H. R.; HOSSEINI, M. J. S.; AMIRPOUR, M. Drilling rate of penetration prediction through committee support vector regression based on imperialist competitive algorithm. **Carbonates and Evaporites**, v. 32, n. 2, p. 205–213, 2017.
- ANTONIO, L. M.; COELLO, C. A. C. Coevolutionary multi-objective evolutionary algorithms: A survey of the state-of-the-art. **IEEE Transactions on Evolutionary Computation**, p. 1–16, 2017. (Preprint).
- ARABJAMALOEI, R.; SHADIZADEH, S. Modeling and optimizing rate of penetration using intelligent systems in an iranian southern oil field (ahwaz oil field). **Petroleum Science and Technology**, v. 29, n. 16, p. 1637–1648, 2011.
- AREHART, R. Drill-bit diagnosis with neural networks. **SPE Computer Applications**, v. 2, n. 04, p. 24–28, 1990. SPE-19558-PA.
- ARNAOUT, A. et al. Intelligent data quality control of real-time rig data. In: SPE MIDDLE EAST INTELLIGENT ENERGY CONFERENCE AND EXHIBITION, 2013, Manama, Bahrain. **Proceedings...** Manama, Bahrain: Society of Petroleum Engineers, 2013. SPE-167437-MS.
- AWOTUNDE, A. A.; MUTASIEM, M. A. Efficient drilling time optimization with differential evolution. In: SPE NIGERIA ANNUAL INTERNATIONAL CONFERENCE AND EXHIBITION, 2014, Lagos, Nigeria. **Proceedings...** Lagos, Nigeria: Society of Petroleum Engineers, 2014. SPE-172419-MS.

BAHARI, A.; SEYED, A. B. Drilling cost optimization in a hydrocarbon field by combination of comparative and mathematical methods. **Petroleum Science**, v. 6, n. 4, p. 451–463, 2009.

BAHARI, M. et al. Determining bourgoyne and young model coefficients using genetic algorithm to predict drilling rate. **Journal of Applied Sciences**, v. 8, n. 17, p. 3050–3054, 2008.

BARBOSA, L. F. F. M. et al. Machine learning methods applied to rate of penetration prediction - a technical review. In: RIO OIL & GAS EXPO AND CONFERENCE, 2018, Rio de Janeiro, Brazil. **Proceedings...** Rio de Janeiro, Brazil: Brazilian Petroleum, Gas and Biofuels Institute (IBP), 2018. IBP1871_18.

BASARIR, H.; TUTLUOGLU, L.; KARPUZ, C. Penetration rate prediction for diamond bit drilling by adaptive neuro-fuzzy inference system and multiple regressions. **Engineering Geology**, v. 173, p. 1–9, 2014.

BATAEE, M.; IRAWAN, S.; KAMYAB, M. Artificial neural network model for prediction of drilling rate of penetration and optimization of parameters. **Journal of the Japan Petroleum Institute**, v. 57, n. 2, p. 65–70, 2014.

BEASLEY, C. et al. Brazil's presalt play. **Oilfield Review**, v. 22, n. 3, p. 28–37, 2010.

BEHERA, S.; SAHOO, S.; PATI, B. A review on optimization algorithms and application to wind energy integration to grid. **Renewable and Sustainable Energy Reviews**, v. 48, p. 214–227, 2015.

BELLO, O. et al. Application of artificial intelligence techniques in drilling system design and operations: A state of the art review and future research pathways. In: SPE NIGERIA ANNUAL INTERNATIONAL CONFERENCE AND EXHIBITION, 2016, Lagos, Nigeria. **Proceedings...** Lagos, Nigeria: Society of Petroleum Engineers, 2016. SPE-184320-MS.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **Journal of Machine Learning Research**, v. 13, p. 281–305, 2012.

BEVILACQUA, M.; CIARAPICA, F. E.; MARCHETTI, B. Acquisition, processing and evaluation of down hole data for monitoring efficiency of drilling processes. **Journal of Petroleum Science Research**, v. 2, n. 2, p. 49–56, 2013.

BEZMINABADI, S. N. et al. Effect of rock properties on rop modeling using statistical and intelligent methods: A case study of an oil well in southwest of iran. **Archives of Mining Sciences**, v. 62, n. 1, 2017.

BILGESU, H. et al. A real-time interactive drill-off test utilizing artificial intelligence algorithm for dsats drilling automation university competition. In: SPE WESTERN REGIONAL MEETING, 2017, Bakersfield, California. **Proceedings...** Bakersfield, California: Society of Petroleum Engineers, 2017. SPE-185730-MS.

BILGESU, H. I. et al. A new approach for the prediction of rate of penetration (rop) values. In: SPE EASTERN REGIONAL MEETING, 1997, Lexington, Kentucky. **Proceedings...** Lexington, Kentucky: Society of Petroleum Engineers, 1997. SPE-39231-MS.

BINGHAM, M. G. **A New Approach to Interpreting Rock Drillability**. [S.l.]: Petroleum Publishing Company, 1965. 93 p.

BISHOP, C. M. Model-based machine learning. **Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences**, The Royal Society, v. 371, n. 1984, 2013.

BOND, D. F. et al. Applying technical limit methodology for step change in understanding and performance. **SPE Drilling & Completion**, v. 13, n. 03, p. 197 – 203, 1998.

BOTCHKAREV, A. Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio. **SSRN Electronic Journal**, 2018. Available at: <<https://www.ssrn.com/abstract=3177507>>. Accessed: sep. 2018.

BOURGOYNE, A.; YOUNG, F. A multiple regression approach to optimal drilling and abnormal pressure detection. **Society of Petroleum Engineers Journal**, v. 14, n. 04, p. 371 – 384, 1974.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). **Statistical Science**, v. 16, n. 3, p. 199–231, 2001.

BYRNE, J. P.; LORUSSO, M.; XU, B. Oil prices, fundamentals and expectations. **Energy Economics**, 2018. (Article in press, DOI:10.1016/j.eneco.2018.05.011).

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16**, ACM Press, New York, New York, USA, p. 785–794, mar 2016. Available at: <<http://arxiv.org/abs/1603.02754>>. Accessed: aug. 2018.

CHEN, X. et al. Real-time optimization of drilling parameters based on mechanical specific energy for rotating drilling with positive displacement motor in the hard formation. **Journal of Natural Gas Science and Engineering**, v. 35, p. 686–694, 2016.

CHIANDUSSI, G. et al. Comparison of multi-objective optimization methodologies for engineering applications. **Computers & Mathematics with Applications**, v. 63, n. 5, p. 912–942, 2012.

CLAESEN, M. et al. Easy Hyperparameter Search Using Optunity. **ArXiv e-prints**, dez. 2014. Available at: <<http://arxiv.org/abs/1412.1114>>. Accessed: sep. 2018.

COHON, J. L.; MARKS, D. H. A review and evaluation of multiobjective programming techniques. **Water Resources Research**, v. 11, n. 2, p. 208–220, 1975.

COLEMAN, T. F.; LI, Y. An interior trust region approach for nonlinear minimization subject to bounds. **SIAM Journal on Optimization**, v. 6, n. 2, p. 418–445, 1996.

CUI, Y. et al. Review: Multi-objective optimization methods and application in energy saving. **Energy**, v. 125, p. 681–704, 2017.

DENG, Y. et al. Theoretical and experimental study on the penetration rate for roller cone bits based on the rock dynamic strength and drilling parameters. **Journal of Natural Gas Science and Engineering**, v. 36, p. 117–123, 2016.

DIAZ, M. B. et al. Drilling data from an enhanced geothermal project and its pre-processing for rop forecasting improvement. **Geothermics**, v. 72, p. 348–357, 2018.

DONG, G.; CHEN, P. A review of the evaluation, control, and application technologies for drill string vibrations and shocks in oil and gas well. **Shock and Vibration**, v. 2016, p. 1–34, 2016. Article ID 7418635.

DONNE, C. P. J. **Parameter Detection in real time Drilling Data**: Create a matlab agent to forecast changes in formation hardness. 138 p. Master's Thesis (Petroleum Engineering Master) — Faculty of Engineering Department of Geoscience and Petroleum, Norwegian University of Science and Technology, Trondheim, Norway, 2017. Available at: <<http://hdl.handle.net/11250/2450329>>. Accessed: 15 nov. 2017.

DUPRIEST, F. E.; KOEDERITZ, W. L. Maximizing drill rates with real-time surveillance of mechanical specific energy. In: SPE/IADC DRILLING CONFERENCE, 2005, Amsterdam, The Netherlands. **Proceedings...** Amsterdam, The Netherlands: Society of Petroleum Engineers, 2005.

DUTRA, B. M. **Novas metodologias e método de análise de dados de perfuração de poços de petróleo e gás natural, e de implementação de curvas de drill-rate test em tempo real visando otimização do processo**. 57 f. Trabalho de Graduação (Engenharia Mecânica) — Faculdade de Engenharia de Guaratinguetá, Universidade Estadual Paulista (UNESP), 2016. Available at: <<http://hdl.handle.net/11449/155378>> Accessed: 11 mar. 2017.

EILERS, P. H. C.; GOEMAN, J. J. Enhancing scatterplots with smoothed densities. **Bioinformatics**, v. 20, n. 5, p. 623–628, 2004.

EQUINOR. **The Statfjord area**. 2018. Available at: <<https://www.equinor.com/en/what-we-do/norwegian-continental-shelf-platforms/statfjord.html>>. Accessed: 23 jul. 2018.

EREN, T.; OZBAYOGLU, M. E. Real time optimization of drilling parameters during drilling operations. In: SPE OIL AND GAS INDIA CONFERENCE AND EXHIBITION, 2010, Mumbai, India. **Proceedings...** Mumbai, India: Society of Petroleum Engineers, 2010. SPE-129126-MS.

ESKANDARIAN, S.; BAHRAMI, P.; KAZEMI, P. A comprehensive data mining approach to estimate the rate of penetration: Application of neural network, rule based models and feature ranking. **Journal of Petroleum Science and Engineering**, v. 156, p. 605–615, 2017.

FERNÁNDEZ-DELGADO, M. et al. Do we need hundreds of classifiers to solve real world classification problems? **Journal of Machine Learning Research**, v. 15, p. 3133–3181, 2014.

FORMIGHIERI, S.; FILHO, P. J. d. F. Estimation of bourgoyne and young model coefficients using markov chain monte carlo simulation. In: 2015 WINTER SIMULATION CONFERENCE (WSC), 2015, Huntington Beach, CA, USA. **Proceedings...** Huntington Beach, CA, USA: IEEE, 2016. p. 1172–1183.

FORTIN, F.-A. et al. DEAP: Evolutionary algorithms made easy. **Journal of Machine Learning Research**, v. 13, p. 2171–2175, 2012.

FRAGA, C. T. d. C. et al. Brazilian pre-salt: An impressive journey from plans and challenges to concrete results. In: OFFSHORE TECHNOLOGY CONFERENCE, 2015, Houston, TX. **Proceedings...** Houston, TX: Offshore Technology Conference, 2015. SPE-25710-MS.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of Computer and System Sciences**, v. 55, n. 1, p. 119–139, 1997.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

FRIEDMAN, J. H. Recent advances in predictive (machine) learning. **Journal of Classification**, v. 23, n. 2, p. 175–197, 2006.

FRUHWIRTH, R. K.; THONHAUSER, G.; MATHIS, W. Hybrid simulation using neural networks to predict drilling hydraulics in real time. In: SPE ANNUAL TECHNICAL CONFERENCE AND EXHIBITION, 2006, San Antonio, Texas. **Proceedings...** San Antonio, Texas: Society of Petroleum Engineers, 2006. SPE-103217-MS.

FURLAN, F. Eles vão fundo no investimento. **Exame**, p. 44–46, jul 2018.

GANDELMAN, R. A. **Predição da ROP e Otimização em Tempo Real de Parâmetros Operacionais na Perfuração de Poços de Petróleo Offshore**. 175 f. Dissertação (Mestrado em Tecnologia de Processos Químicos e Bioquímicos) — Escola de Química, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ - Brasil, 2012.

GARCÍA, L. P. F.; CARVALHO, A. C. P. L. F. de; LORENA, A. C. Noisy data set identification. In: **Hybrid Artificial Intelligent Systems**. [S.l.]: Springer Berlin Heidelberg, 2013. p. 629–638.

GELBART, M. A.; SNOEK, J.; ADAMS, R. P. Bayesian Optimization with Unknown Constraints. **ArXiv e-prints**, mar. 2014. Available at: <<https://arxiv.org/abs/1403.5607v1>>. Accessed: 23 oct. 2018.

GHASEMLOONIA, A.; Geoff Rideout, D.; BUTT, S. D. A review of drillstring vibration modeling and suppression methods. **Journal of Petroleum Science and Engineering**, v. 131, p. 150–164, 2015.

GODHAVN, J.-M. et al. Drilling seeking automatic control solutions. **IFAC Proceedings Volumes**, IFAC, v. 44, n. 1, p. 10842–10850, jan 2011.

GONZÁLEZ-ORDIANO, J. Á. et al. Photovoltaic power forecasting using simple data-driven models without weather data. **Computer Science - Research and Development**, v. 32, n. 1-2, p. 237–246, 2017.

GRAHAM, J.; MUENCH, N. Analytical determination of optimum bit weight and rotary speed combinations. In: FALL MEETING OF THE SOCIETY OF PETROLEUM ENGINEERS OF AIME, 1959, Dallas, Texas. **Proceedings...** Dallas, Texas: Society of Petroleum Engineers, 1959. SPE-1349-G.

GUAN, C. et al. Very short-term load forecasting: Multilevel wavelet neural networks with data pre-filtering. **IEEE Transactions on Power Systems**, v. 28, n. 1, p. 30–41, 2013.

GURIA, C.; GOLI, K. K.; PATHAK, A. K. Multi-objective optimization of oil well drilling using elitist non-dominated sorting genetic algorithm. **Petroleum Science**, v. 11, n. 1, p. 97–110, 2014.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, v. 3, n. 3, p. 1157–1182, 2003.

- HAIMES, Y. Y.; LASDON, L. S.; WISMER, D. A. On a bicriterion formulation of the problems of integrated system identification and system optimization. **IEEE Transactions on Systems, Man, and Cybernetics**, SMC-1, n. 3, p. 296–297, 1971.
- HAMADA, Y. et al. Continuous depth profile of the rock strength in the nankai accretionary prism based on drilling performance parameters. **Scientific Reports**, Springer US, v. 8, n. 1, p. 2622, 2018.
- HAMPEL, F. R. The influence curve and its role in robust estimation. **Journal of the American Statistical Association**, v. 69, n. 346, p. 383–393, 1974.
- HARELAND, G.; HOBEROCK, L. Use of drilling parameters to predict in-situ stress bounds. In: SPE/IADC DRILLING CONFERENCE, 1993, Amsterdam, Netherlands. **Proceedings...** Amsterdam, Netherlands: Society of Petroleum Engineers, 1993. SPE-25727-MS.
- HARELAND, G.; RAMPERSAD, P. Drag - bit model including wear. In: SPE LATIN AMERICA/CARIBBEAN PETROLEUM ENGINEERING CONFERENCE, 1994, Buenos Aires, Argentina. **Proceedings...** Buenos Aires, Argentina: Society of Petroleum Engineers, 1994. SPE-26957-MS.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. 12. ed. New York, NY: Springer New York, 2009. (Springer Series in Statistics).
- HEGDE, C.; DAIGLE, H.; GRAY, K. E. Performance comparison of algorithms for real-time rate-of-penetration optimization in drilling using data-driven models. **SPE Journal**, jul 2018. (Preprint, DOI: 10.2118/191141-PA).
- HEGDE, C. et al. Analysis of rate of penetration (rop) prediction in drilling using physics-based and data-driven models. **Journal of Petroleum Science and Engineering**, v. 159, p. 295–306, 2017.
- HEGDE, C.; GRAY, K. Use of machine learning and data analytics to increase drilling efficiency for nearby wells. **Journal of Natural Gas Science and Engineering**, v. 40, p. 327–335, 2017.
- HEGDE, C.; GRAY, K. Evaluation of coupled machine learning models for drilling optimization. **Journal of Natural Gas Science and Engineering**, v. 56, p. 397–407, 2018.
- HWANG, C.-L.; MASUD, A. S. M. **Multiple Objective Decision Making — Methods and Applications**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1979. v. 164. 358 p. (Lecture Notes in Economics and Mathematical Systems, v. 164).
- INDEXMUNDI. **Crude Oil (petroleum); Dubai Fateh Monthly Price - US Dollars per Barrel**. 2018. Available at: <<https://www.indexmundi.com/commodities/?commodity=crude-oil-dubai&months=60>>. Accessed: 27 jul. 2018.
- INDEXMUNDI. **Crude Oil (petroleum); West Texas Intermediate Monthly Price - US Dollars per Barrel**. 2018. Available at: <<https://www.indexmundi.com/commodities/?commodity=crude-oil-west-texas-intermediate>>. Accessed: 27 jul. 2018.
- INTERNATIONAL ENERGY AGENCY. **Key World Energy Statistics 2017**. 2017. 97 p. Available at: <<https://www.iea.org/publications/freepublications/publication/KeyWorld2017.pdf>>. Accessed: jul. 2018.

INTERNATIONAL ENERGY AGENCY. **World Energy Outlook 2017**. 2017. Available at: <<https://www.iea.org/weo2017/>>. Accessed: jul. 2018.

KARAKUL, H.; ULUSAY, R. Empirical correlations for predicting strength properties of rocks from p-wave velocity under different degrees of saturation. **Rock Mechanics and Rock Engineering**, v. 46, n. 5, p. 981–999, 2013.

KHAN, M. I. Falling oil prices: Causes, consequences and policy implications. **Journal of Petroleum Science and Engineering**, v. 149, p. 409–427, 2017.

KIM, M. S. Impacts of supply and demand factors on declining oil prices. **Energy**, v. 155, p. 1059–1065, 2018.

KOTSIANTIS, S. B. Decision trees: a recent overview. **Artificial Intelligence Review**, v. 39, n. 4, p. 261–283, 2013.

KUTAS, D. T. et al. A study of the applicability of bourgoyne & young rop model and fitting reliability through regression. In: INTERNATIONAL PETROLEUM TECHNOLOGY CONFERENCE, 2015, Doha, Qatar. **Proceedings...** Doha, Qatar: International Petroleum Technology Conference, 2015. IPTC-18521-MS.

LEYS, C. et al. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. **Journal of Experimental Social Psychology**, v. 49, n. 4, p. 764–766, 2013.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical Analysis with Missing Data**. 2nd. ed. Hoboken, NJ - EUA: John Wiley and Sons, Inc., 2002.

LUMMUS, J. L. Drilling optimization. **Journal of Petroleum Technology**, v. 22, n. 11, p. 1379–1388, 1970.

LYONS, W. C.; PLISGA, G. J. **Standard Handbook of Petroleum and Natural Gas Engineering**. 2nd. ed. [S.l.]: Gulf Professional Publishing, 2004.

MA, T.; CHEN, P.; ZHAO, J. Overview on vertical and directional drilling technologies for the exploration and exploitation of deep petroleum resources. **Geomechanics and Geophysics for Geo-Energy and Geo-Resources**, Springer International Publishing, v. 2, n. 4, p. 365–395, 2016.

MARLER, R.; ARORA, J. Survey of multi-objective optimization methods for engineering. **Structural and Multidisciplinary Optimization**, v. 26, n. 6, p. 369–395, 2004.

MATHIS, W. et al. Use of real-time rig-sensor data to improve daily drilling reporting, benchmarking, and planning: A case study. **SPE Drilling & Completion**, v. 22, n. 03, p. 217–226, 2007.

MATHWORKS. **Function**: hampel. 2018. Available at: <<https://www.mathworks.com/help/signal/ref/hampel.html>>. Accessed: 19 oct. 2018.

MATHWORKS. **Function**. 2018. Available at: <<https://www.mathworks.com/help/matlab/ref/fillmissing.html>>. Accessed: 19 oct. 2018.

MATHWORKS. **Function**: quantilepredict. 2018. Available at: <<https://www.mathworks.com/help/stats/treebagger.quantilepredict.html>>. Accessed: 23 oct. 2018.

MATHWORKS. **Function:** fitrensemble. 2018. Available at: <<https://www.mathworks.com/help/stats/fitrensemble.html>>. Accessed: 23 oct. 2018.

MATHWORKS. **Tune Random Forest Using Quantile Error and Bayesian Optimization.** 2018. Available at: <<https://www.mathworks.com/help/stats/tune-random-forest-using-quantile-error-and-bayesian-optimization.html>>. Accessed: 23 oct. 2018.

MAURER, W. The “perfect - cleaning” theory of rotary drilling. **Journal of Petroleum Technology**, v. 14, n. 11, p. 1270–1274, 1962. SPE-408-PA.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 5, n. 4, p. 115–133, 1943.

MEINSHAUSEN, N. Quantile regression forests. **Journal of Machine Learning Research**, v. 7, p. 983–999, 2006.

MENDES-MOREIRA, J. et al. Ensemble approaches for regression. **ACM Computing Surveys**, v. 45, n. 1, p. 1–40, 2012.

MENG, C. U. I. et al. Maximizing drilling performance with real-time surveillance system based on parameters optimization algorithm. **Advances in Petroleum Exploration and Development**, v. 8, n. 1, p. 15–24, 2014.

MINISTÉRIO DE MINAS E ENERGIA. EMPRESA DE PESQUISA ENERGÉTICA. **Plano Decenal de Expansão de Energia 2026**. Brasília, 2017. 271 p.

MITCHELL, R. F.; MISKA, S. Z. Drilling engineering. In: COLEÇÃO, E. da (Ed.). **Petroleum Engineering Handbook**. [S.l.]: Society of Petroleum Engineers, 2007. Volume 2.

MITCHELL, R. F.; MISKA, S. Z. **Fundamentals of Drilling Engineering**. Richardson, TX - EUA: Society of Petroleum Engineers, 2011.

MOHAN, K.; ADIL, F.; SAMUEL, R. Comprehensive hydromechanical specific energy calculation for drilling efficiency. **Journal of Energy Resources Technology**, v. 137, n. 1, p. 012904, 2014.

MORAVEJI, M. K.; NADERI, M. Drilling rate of penetration prediction and optimization using response surface methodology and bat algorithm. **Journal of Natural Gas Science and Engineering**, v. 31, p. 829–841, 2016.

MOTAHHARI, H.; HARELAND, G.; JAMES, J. Improved drilling efficiency technique using integrated pdm and pdc bit parameters. **Journal of Canadian Petroleum Technology**, v. 49, n. 10, p. 45–52, 2010.

NASCIMENTO, A. **Mathematical Modeling for Drilling Optimization in Pre-salt Sections: a Focus on South Atlantic Ocean Operations**. 135 f. Tese (Doutorado em Engenharia Mecânica) — Faculdade de Engenharia de Guaratinguetá, Universidade Estadual Paulista (UNESP), 2016.

NASCIMENTO, A. et al. Reverse engineering: A new well monitoring and analysis methodology approaching playing-back drill-rate tests in real-time for drilling optimization. **Journal of Energy Resources Technology**, v. 139, n. 1, p. 12902–12902–5, 2016.

NASCIMENTO, A. et al. Dynamic drill-rate test approach applied to a pre-salt case study. In: SPE LATIN AMERICAN AND CARIBBEAN PETROLEUM ENGINEERING CONFERENCE, 2015, Quito, Ecuador. **Proceedings...** Quito, Ecuador: Society of Petroleum Engineers, 2015. SPE-177141-MS.

NASCIMENTO, A. et al. Mathematical modeling applied to drilling engineering: An application of bourgoyne and young rop model to a presalt case study. **Mathematical Problems in Engineering**, v. 2015, p. 9, 2015. Article ID 631290.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in Neurorobotics**, v. 7, p. 23, 2013.

NUNES, F.; LIMA, D. Licitações de óleo e gás atraem 32 companhias. **O Estado de São Paulo**, 2017. 24 set. 2017. Seção Economia B5.

OLULEYE, B. et al. A genetic algorithm-based feature selection. **International Journal of Electronics Communication and Computer Engineering**, v. 5, n. 4, p. 899–905, 2014.

OLULEYE, B. et al. Zernike moments and genetic algorithm : Tutorial and application. **British Journal of Mathematics & Computer Science**, v. 4, n. 15, p. 2217–2236, 2014.

ORGANIZATION OF THE PETROLEUM EXPORTING COUNTRIES. **2017 OPEC World Oil Outlook 2040**. 2017. Available at: <<http://www.opec.org>>. Accessed: jul. 2018.

OTALVORA, W. C. et al. A comprehensive approach to measure the realtime data quality using key performance indicators. In: SPE ANNUAL TECHNICAL CONFERENCE AND EXHIBITION, 2016, Dubai, UAE. **Proceedings...** Dubai, UAE: Society of Petroleum Engineers, 2016. SPE-181315-MS.

PAYETTE, G. S. et al. Real-time well-site based surveillance and optimization platform for drilling: Technology, basic workflows and field results. In: SPE/IADC DRILLING CONFERENCE AND EXHIBITION, 2017, Hague, The Netherlands. **Proceedings...** Hague, The Netherlands: Society of Petroleum Engineers, 2017. SPE-184615-MS.

PESSIER, R.; FEAR, M. Quantifying common drilling problems with mechanical specific energy and a bit-specific coefficient of sliding friction. In: SPE ANNUAL TECHNICAL CONFERENCE AND EXHIBITION, 1992, Washington, D.C. **Proceedings...** Washington, D.C: Society of Petroleum Engineers, 1992.

PONTES, F. J. et al. Artificial neural networks for machining processes surface roughness modeling. **The International Journal of Advanced Manufacturing Technology**, v. 49, n. 9-12, p. 879–902, 2010.

PREST, B. C. Explanations for the 2014 oil price decline: Supply or demand? **Energy Economics**, v. 74, p. 63–75, 2018.

QUINLAN, J. R. The effect of noise on concept learning. In: **Machine Learning: Artificial Intelligence Approach, Volume 2**. [S.l.]: Morgan Kaufmann Publishers Inc., 1986. cap. 6, p. 149–166.

RABIA, H. Specific energy as a criterion for drill performance prediction. **International Journal of Rock Mechanics and Mining Sciences and**, v. 19, n. 1, p. 39–42, 1982.

RAHMANIFARD, H.; PLAKSINA, T. Application of artificial intelligence techniques in the petroleum industry: a review. **Artificial Intelligence Review**, Springer Netherlands, p. 1–24, 2018.

REN, Y.; ZHANG, L.; SUGANTHAN, P. Ensemble classification and regression-recent developments, applications and future directions [review article]. **IEEE Computational Intelligence Magazine**, v. 11, n. 1, p. 41–53, 2016.

ROBERTS, J. J. et al. GAtoolbox: a matlab - based genetic algorithm toolbox for function optimization. In: THE 12TH LATIN-AMERICAN CONGRESS ON ELECTRICITY GENERATION AND TRANSMISSION - CLAGTEE 2017, 12., 2017, Mar del Plata. **Proceedings...** Mar del Plata, 2017. p. 1–12.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, v. 65, n. 6, p. 386–408, 1958.

ROUSSEEUW, P. J.; CROUX, C. Alternatives to the median absolute deviation. **Journal of the American Statistical Association**, v. 88, n. 424, p. 1273–1283, 1993.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. **WIREs Data Mining and Knowledge Discovery**, v. 8, n. 4, p. e1249, 2018.

SALGADO, C. M. et al. Noise versus outliers. In: _____. **Secondary Analysis of Electronic Health Records**. Cham: Springer International Publishing, 2016. p. 163–183.

SARKAR, K.; VISHAL, V.; SINGH, T. N. An empirical correlation of index geomechanical parameters with the compressional wave velocity. **Geotechnical and Geological Engineering**, v. 30, n. 2, p. 469–479, 2012.

SCHLUMBERGER. **Oilfield Glossary**: mud weight. 2018. Available at: <https://www.glossary.oilfield.slb.com/en/Terms/m/mud_weight.aspx>. Accessed: 16 aug. 2018.

SCHLUMBERGER. **Oilfield Glossary**: equivalent circulating density. 2018. Available at: <https://www.glossary.oilfield.slb.com/Terms/e/equivalent_circulating_density.aspx>. Accessed: 16 aug. 2018.

SCRUCCA, L. GA: A package for genetic algorithms in R. **Journal of Statistical Software**, v. 53, n. 4, p. 213–266, 2013.

SEBORG, D. E. et al. **Process Dynamics and Control**. [S.l.]: John Wiley & Sons Inc, 2011.

SHARMA, P. K.; SINGH, T. N. A correlation between p-wave velocity, impact strength index, slake durability index and uniaxial compressive strength. **Bulletin of Engineering Geology and the Environment**, v. 67, n. 1, p. 17–22, 2008.

SHOKOUHI, S. V.; SKALLE, P.; AAMODT, A. An overview of case-based reasoning applications in drilling engineering. **Artificial Intelligence Review**, v. 41, n. 3, p. 317–329, 2014.

SILVA, C. F. D. **Análise geomecânica dos carbonatos do pré-sal da bacia de santos**. 139 f. Dissertação (Mestrado em Engenharia Civil) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil, nov 2016. Available at: <http://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=30291@1>. Accessed: 5 aug. 2018.

SKALLE, P. **Drilling data some doubts**. 2018. [e-mail]. Message received by the author on 28 April 2018.

SKALLE, P.; AAMODT, A.; ERIKGUNDERSEN, O. Experience transfer for process improvement. **Engineering Applications of Artificial Intelligence**, v. 26, n. 9, p. 2206–2214, 2013.

SKJERPEN, T. et al. Modelling and forecasting rig rates on the norwegian continental shelf. **Resource and Energy Economics**, v. 53, p. 220–239, 2018.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. **ArXiv e-prints**, jun. 2012. Available at: <<https://arxiv.org/abs/1206.2944v2>>. Accessed: 22 sep. 2018.

SOARES, C.; DAIGLE, H.; GRAY, K. Evaluation of pdc bit rop models and the effect of rock strength on model coefficients. **Journal of Natural Gas Science and Engineering**, v. 34, p. 1225–1236, 2016.

SOARES, C.; GRAY, K. Real-time predictive capabilities of analytical and machine learning rate of penetration (rop) models. **Journal of Petroleum Science and Engineering**, n. July, p. 1–26, 2018. (Article in press, DOI: 10.1016/j.petrol.2018.08.083).

STAVELEY, C.; THOW, P. Increasing drilling efficiencies through improved collaboration and analysis of real-time and historical drilling data. In: SPE INTELLIGENT ENERGY CONFERENCE AND EXHIBITION, 2010, Utrecht, The Netherlands. **Proceedings...** Utrecht, The Netherlands: Society of Petroleum Engineers, 2010. SPE-128722-MS.

STORN, R.; PRICE, K. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. **Journal of Global Optimization**, v. 11, n. 4, p. 341–359, 1997.

STROBL, C. et al. Bias in random forest variable importance measures: Illustrations, sources and a solution. **BMC Bioinformatics**, v. 8, n. 1, p. 25, 2007.

SZLEK, J.; MENDYK, A. **CRAN-R project fscaret**: Automated feature selection from 'caret'. 2018. Available at: <<https://CRAN.R-project.org/package=fscaret>>.

TANSEV, E. A heuristic approach to drilling optimization. In: **Proceedings...** Dallas, Texas: Society of Petroleum Engineers, 1975. p. 18.

TAVARES, R. M. **Interpretação e Análise de Dados de Perfuração em Poços de Petróleo**. 145 f. Dissertação (Mestrado em Ciências e Engenharia de Petróleo) — Universidade Estadual de Campinas, Faculdade de Engenharia Mecânica e Instituto de Geociências, 2006. Available at: <http://repositorio.unicamp.br/bitstream/REPOSIP/263676/1/Tavares_RogérioMartins_M.pdf>. Accessed: sep. 2018.

TEALE, R. The concept of specific energy in rock drilling. **International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts**, v. 2, n. 1, p. 57–73, 1965.

THONHAUSER, G. Using real-time data for automated drilling performance analysis. **OIL GAS European Magazine**, v. 4, p. 170–173, 2004.

VESCONTE, M.-J. L.; TINKHOF, R.; HARDMAN, P. The majnoon field: A case study of drilling operations in a remote area of iraq. In: IADC/SPE DRILLING CONFERENCE AND EXHIBITION, 2014, Fort Worth, Texas. **Proceedings...** Fort Worth, Texas: Society of Petroleum Engineers, 2014. SPE-167949-MS.

WANG, Y.; SALEHI, S. Application of real-time field data to optimize drilling hydraulics using neural network approach. **Journal of Energy Resources Technology**, v. 137, n. 6, p. 062903–062903–9, 2015.

WARREN, T. Penetration rate performance of roller cone bits. **SPE Drilling Engineering**, v. 2, n. 01, p. 9–18, 1987. SPE-13259-PA.

WOLPERT, D. H.; MACREADY, W. G. An efficient method to estimate bagging's generalization error. **Machine Learning**, v. 35, n. 1, p. 41–55, 1999.

XUE, B. et al. A survey on evolutionary computation approaches to feature selection. **IEEE Transactions on Evolutionary Computation**, v. 20, n. 4, p. 606–626, 2016.

YOUNG, F. Computerized drilling control. **Journal of Petroleum Technology**, v. 21, n. 04, p. 483–496, 1969.

YUAN, P. et al. Application of case-based reasoning method on drilling parameter optimization. In: 2009 INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE AND SOFTWARE ENGINEERING, 2009, Wuhan, China. **Proceedings...** Wuhan, China, 2009.

ZADEH, L. Optimality and non-scalar-valued performance criteria. **IEEE Transactions on Automatic Control**, v. 8, n. 1, p. 59–60, 1963.

ZAMORA, M.; ROY, S. The top 10 reasons to rethink hydraulics and rheology. In: IADC/SPE ASIA PACIFIC DRILLING TECHNOLOGY, 2000, Kuala Lumpur, Malaysia. **Proceedings...** [S.l.]: Society of Petroleum Engineers, 2000.

ZHANG, Y.; WANG, S.; JI, G. A comprehensive survey on particle swarm optimization algorithm and its applications. **Mathematical Problems in Engineering**, v. 2015, p. 1–38, 2015. Article ID 931256.

ZHOU, A. et al. Multiobjective evolutionary algorithms: A survey of the state of the art. **Swarm and Evolutionary Computation**, v. 1, n. 1, p. 32–49, 2011.

ZHU, X.; TANG, L.; YANG, Q. A literature review of approaches for stick-slip vibration suppression in oilwell drillstring. **Advances in Mechanical Engineering**, v. 6, 2014. Article ID 967952.

APPENDIX A – NEW SUGGESTION FOR THE VARIABLE NAMES OF THE WELL 6 FROM NORWAY

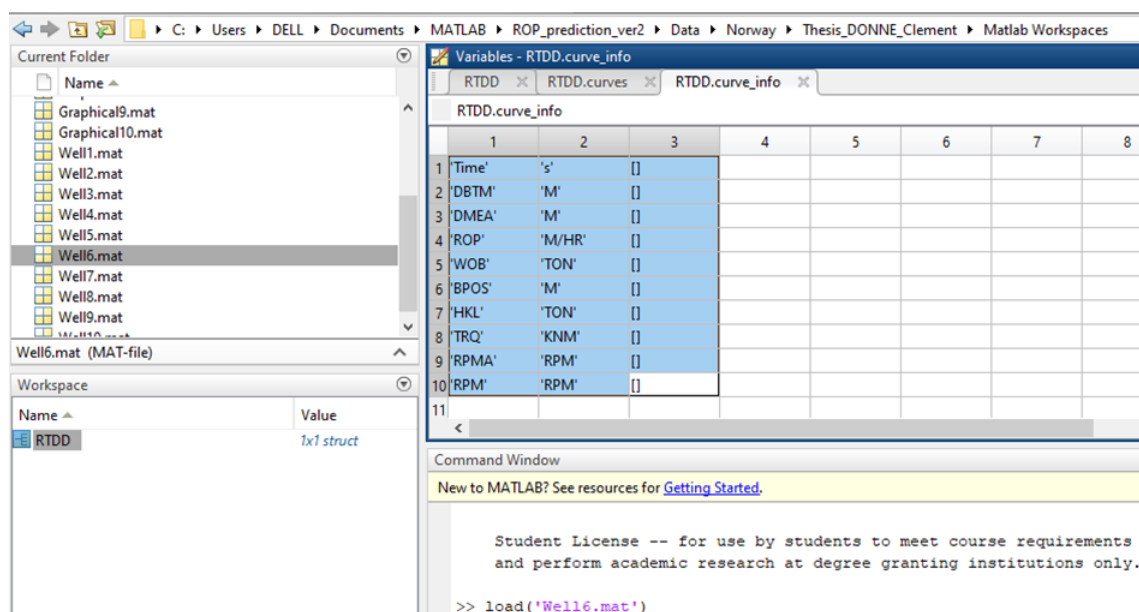
Considering the amount of variables recorded for each well drilled in the data Real-Time Drilling Data (RTDD.mat) used by (DONNE, 2017), it is possible to classify them into three different groups:

- first group – with 8 variables recorded (wells 1-2)
- second group – with 10 variables recorded (wells 3-7)
- third group – with 36 variables recorded (wells 8-10)

The sequence in which the variable names appear in the `RTDD.cuve_info` (see. Figure A.1) were listed in the Table A.1, where it is possible to observe, that the wells 4, 5 and 7 have the same sequence of the variables. The well 3 has a different sequence, but it does not lead to a problem, because the curves shape of the well 3 is similar to the curves obtained for the wells 4, 5 and 7. A problem only occurs when we plot the curves from the well 6, following the variable names given in the `RTDD.cuve_info` of the file `Well6.mat`.

It was observed that the drilling curves of all wells, except the 6, have a common pattern. We assume that this “common pattern” is the right one. This assumption is the basis for the

Figure A.1 – Getting the sequence of the variable names in `RTDD.cuve_info`; showing an example for the well 6. The first line of `RTDD.cuve_info` indicates that the variable 'Time' is recorded in the first column; the variable 'DBTM' is recorded in the second column and so on



Source: own authorship.

Table A.1 – Sequence in which the variable names appear in the drilling data with 10 variables recorded (that is, the 2nd group). We highlighted the columns with the problem

	Date	Column									
		# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10
well 3	02-Apr-2003	'Time'	'DBTM'	'DMEA'	'ROP'	'WOB'	'RPMB'	'HKL'	'BPOS'	'TRQ'	'RPMA'
well 4	16-Jan-2004	'Time'	'DBTM'	'DMEA'	'ROP'	'WOB'	'RPM'	'BPOS'	'HKL'	'TRQ'	'RPMA'
well 5	19-Jun-1993	'Time'	'DBTM'	'DMEA'	'ROP'	'WOB'	'RPM'	'BPOS'	'HKL'	'TRQ'	'RPMA'
well 6	17-Apr-2005	'Time'	'DBTM'	'DMEA'	'ROP'	'WOB'	BPOS'	HKL'	TRQ'	RPMA'	RPM'
well 7	14-Nov-2004	'Time'	'DBTM'	'DMEA'	'ROP'	'WOB'	'RPM'	'BPOS'	'HKL'	'TRQ'	'RPMA'

Source: own authorship.

following comparison. An example of this “common pattern” is given in Figure 4.2, where the curves of the well 4 are plotted.

We believe that this assumption is right, because, when the block position goes down, the measured depth of the well increases, and the ROP values indicate also that the well is being drilled. When the block position goes up, the measured depth of the well remains the same, and the ROP goes to zero, as expected. The other parameters, such as hock load, WOB, RPM and torque seem to be adequate for both states, drilling and not-drilling.

The plots of all wells have all the same pattern, except the well 6. If we now compare the common pattern (Figure 4.2) with the curves of the well 6 (Figure A.2), according to the sequence of variables presented in `RTDD.cuve_info`, we realize that the hock load is acting like the block position, and the torque is acting like rpm. Following the suggestion presented in the Table A.2 results in the same curves pattern from all other wells (Figure A.3, which was called here as the “common pattern”. The bit rotation magnitude is now around 300 rpm, the same magnitude found in the original graphical study of the real-time drilling data (see Figure A.4).

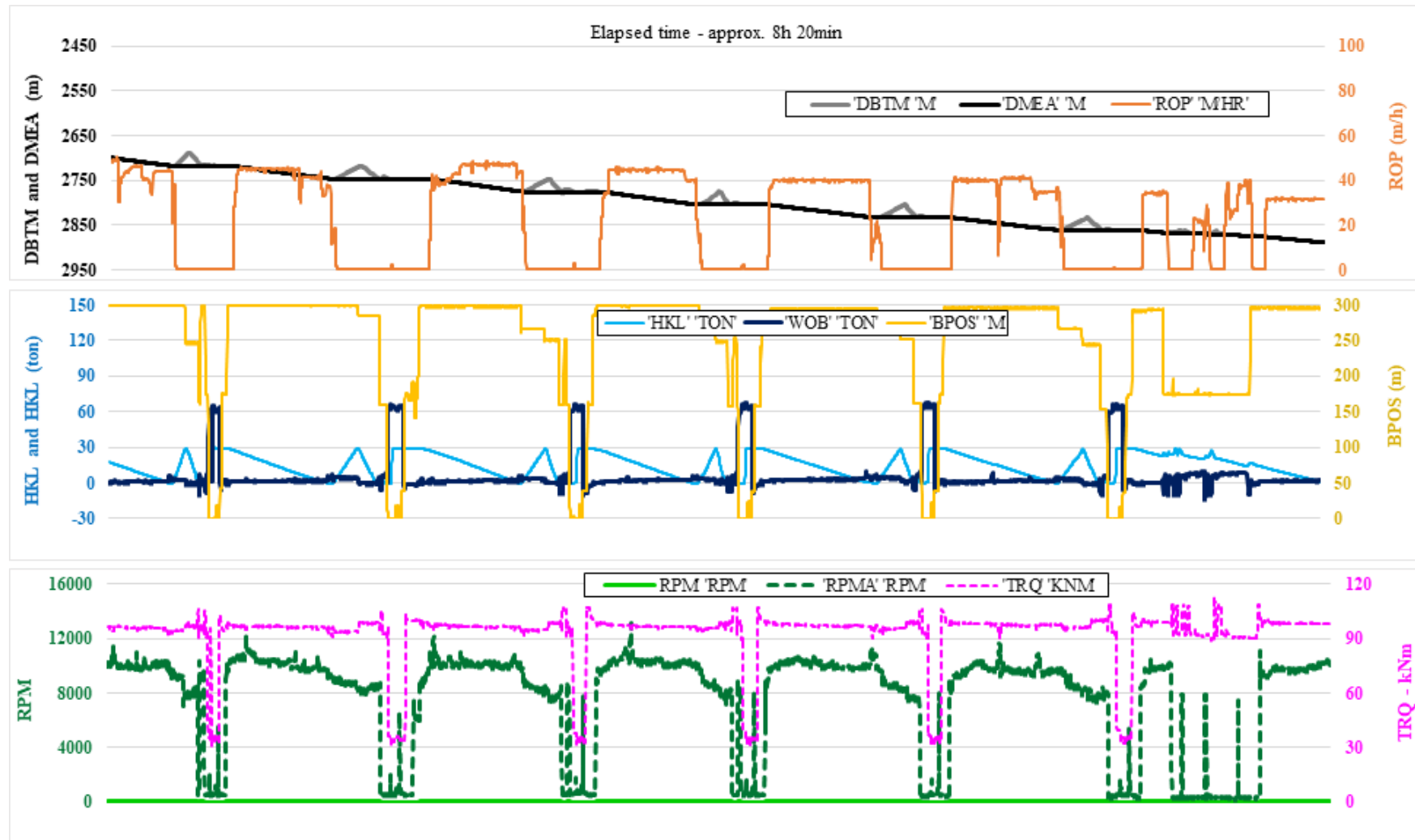
By observing the previous curves, it is possible to conclude that the curves pattern of the well 6 would be the same of other wells, if the sequence of variables were according to the suggestion presented in the Table A.2.

Table A.2 – Suggestion for the variable names of the well 6. We believe that the sequence of the variable names for the well 6 should be the same as the wells 4, 5 and 7, which can be achieved by only sliding the RPM to the column # 6, and moving the yellow block downwards

How it is			Our suggestion
Column #	Variable	Unit	Variable
1	'Time'	's'	Time
2	'DBTM'	'M'	DBTM
3	'DMEA'	'M'	DMEA
4	'ROP'	'M/HR'	ROP
5	'WOB'	'TON'	WOB
6	BPOS	'M'	RPM
7	HKL	'TON'	BPOS
8	TRQ	'KNM'	HKL
9	RPMA	'RPM'	TRQ
10	RPM	'RPM'	RPMA

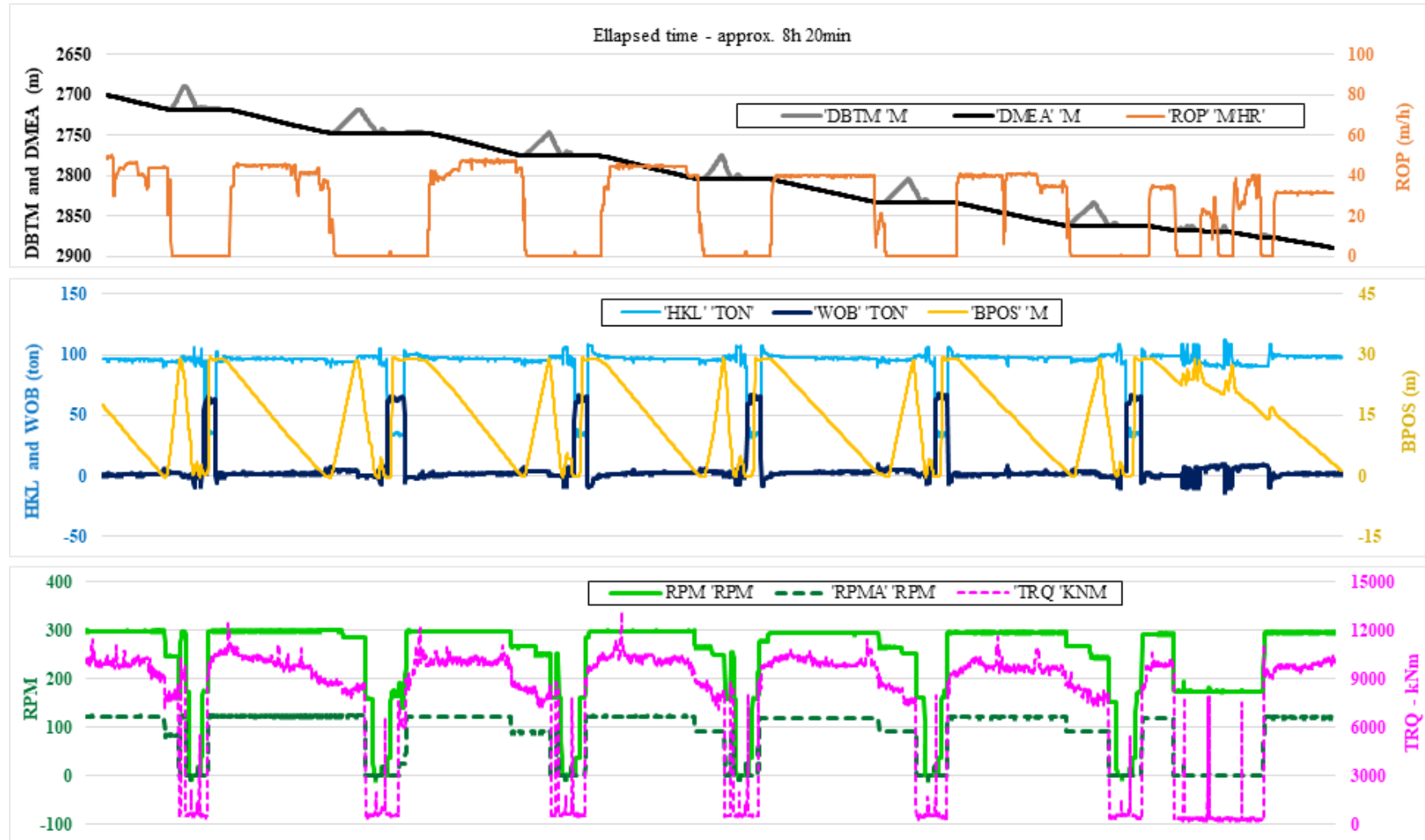
Source: own authorship.

Figure A.2 – Real-time drilling data of well 6 considering the first approx. 8 hours of available recording according to the informed sequence of variable names – state: “how it is”. It is possible to observe that the HKL is acting like the BPOS, and the TRQ like the RPMA.



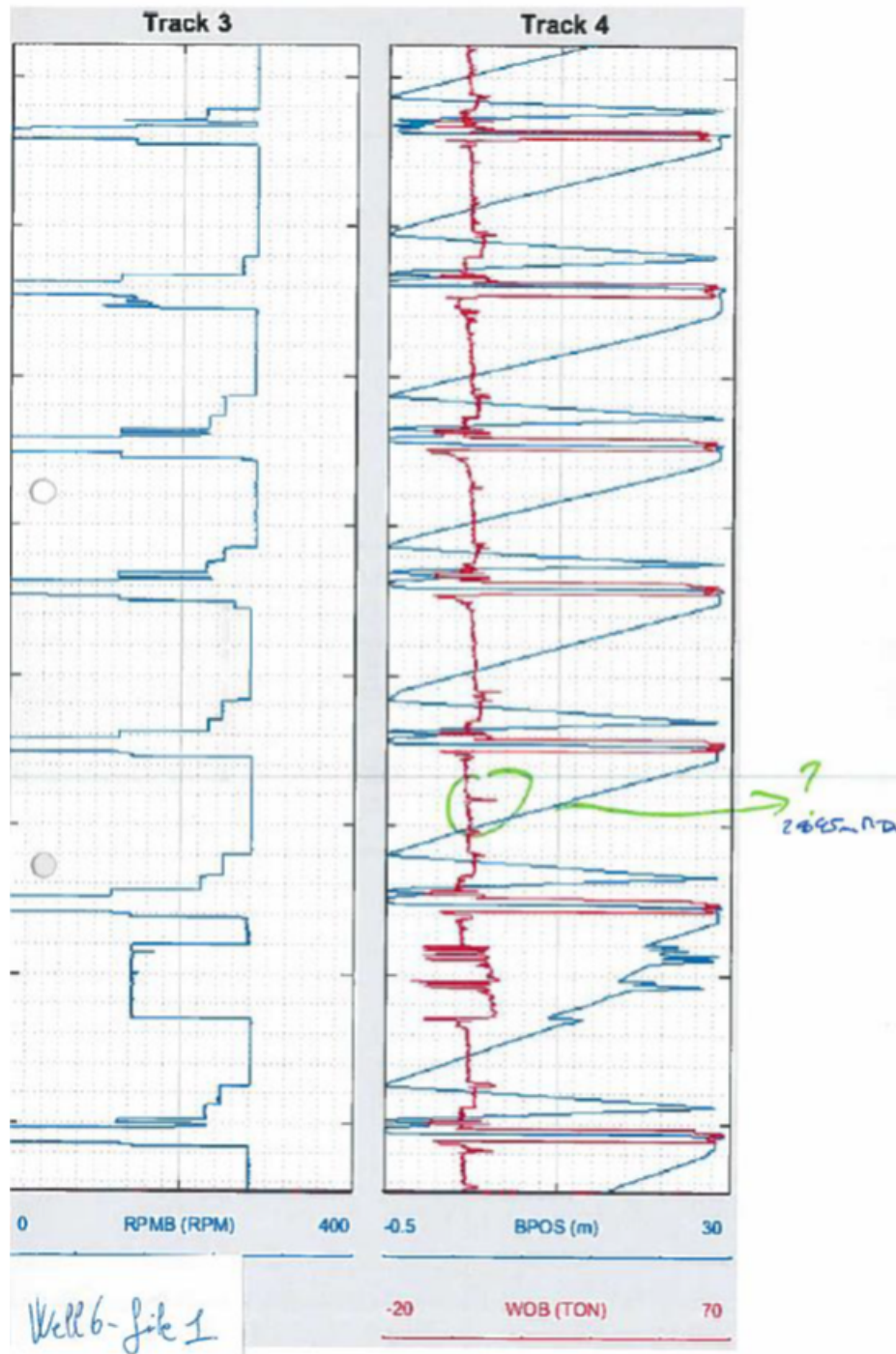
Source: dataset available in the work of Donne (2017)

Figure A.3 – Real-time drilling data of well 6 considering the first approx. 8 hours of available recording according to our suggestion for the variable names – Our suggestion for the sequence of the variable names



Source: dataset available in the work of Donne (2017)

Figure A.4 – Graphical Study of the real-time drilling data – well 6 – file 1.

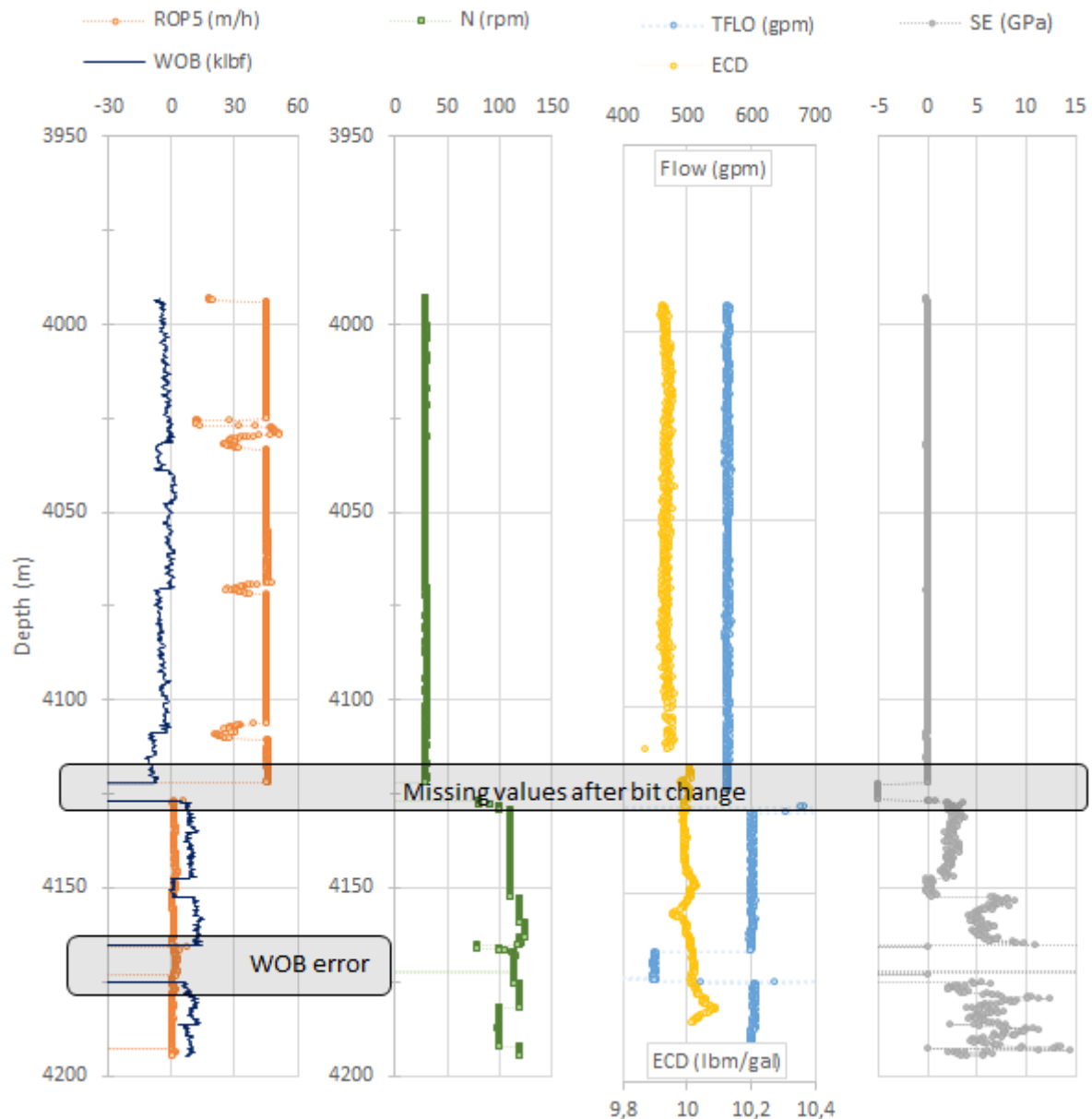


Source: Donne (2017, Appendix VII, pp. XXXII)

APPENDIX B – DRILL-CURVES FROM PRE-SALT

The remaining drill-curves from pre-salt not shown in the Chapter 4 are presented in Figure B.1, Figure B.2 and Figure B.3 for respectively Well A, B and D.

Figure B.1 – Raw recording of drill curves from Well A. The specific energy was calculated based on Rabia's formulation.

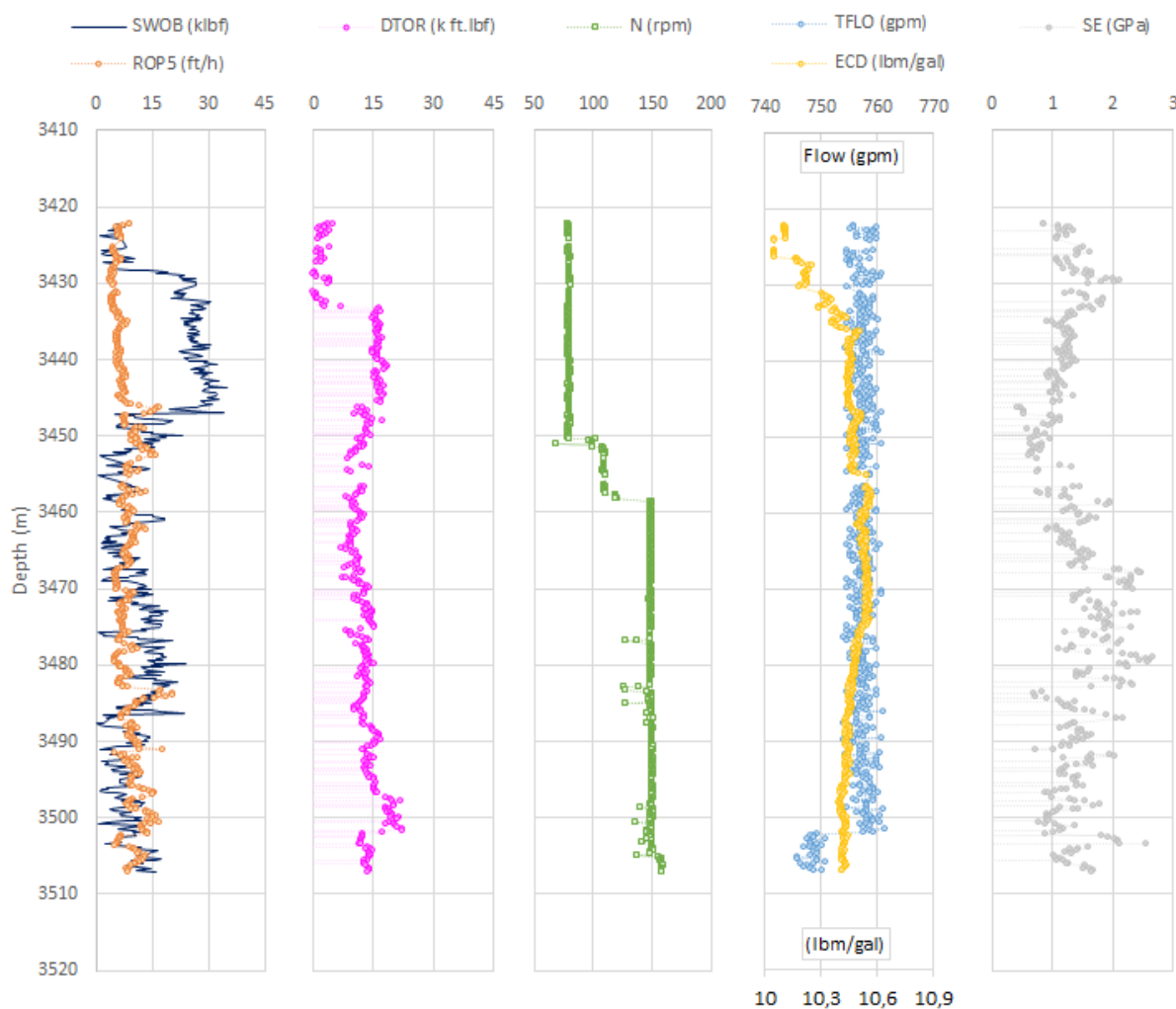


Source: own authorship.

In the Figure B.1, it is possible to observe, at the depth of 4 127 m, an abrupt change in the magnitude of drilling variables. This change coincides with a bit change (NASCIMENTO, 2016). It will be considered that after this change, the carbonate formation started to be drilled. Therefore, the depth of 4 127 m is considered as an cut-off for this dataset. That is, everything

above it is discarded. It is also possible to observe that the WOB recording went to round - 130 klbf in the section from 4165 m to 4174 m. In drill-reports, no events related to hydraulics were reports for this well (NASCIMENTO, 2016).

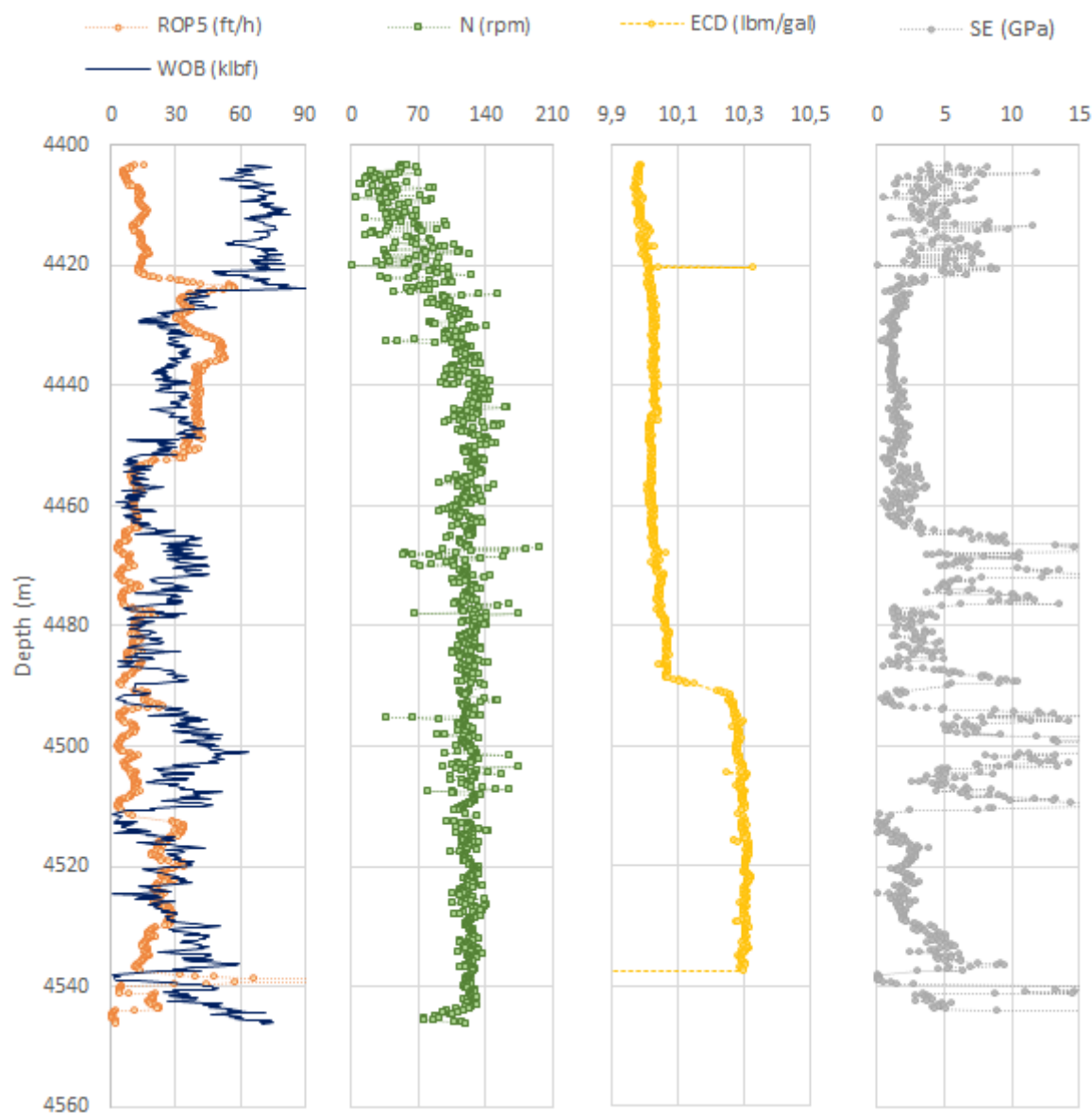
Figure B.2 – Raw recording of drill curves from Well B. The specific energy was calculated based on Taele's formulation.



Source: own authorship.

Some missing values are present in the dataset from Well B, as seen Figure B.2. However, it is believed that the proposed data pre-treatment method is able to handle this issue, since there is not any long period of missing values. In this well, an kick event was reported at the depth 3 436 - 3 447 m (NASCIMENTO, 2016).

Figure B.3 – Raw recording of drill curves from Well D. The specific energy that was calculated based on Rabia's formulation.



Source: own authorship.

In the Well D, a loss event was reported at the depth of 4 543 m 4 546 m (NASCIMENTO, 2016). Three meters above it, a data transmission loss of ECD occurred. The data transmission loss occurred also for some other variables, such as the downhole annular pressure and gamma-ray. For this reason, the drill-curves from this dataset is considered until this data transmission loss occurs (i.e. until 4 537 m).

APPENDIX C – DATA PRE-TREATMENT ANALYSIS OF HAMPEL FILTER

In Matlab, there is a function called `filloutliers` to identified and replace the outliers. For that, several techniques can be employed to identify the outliers, including the moving median absolute deviation. Other function from Matlab is `hampel` that employs the hamper filter (i.e. moving median absolute deviation).

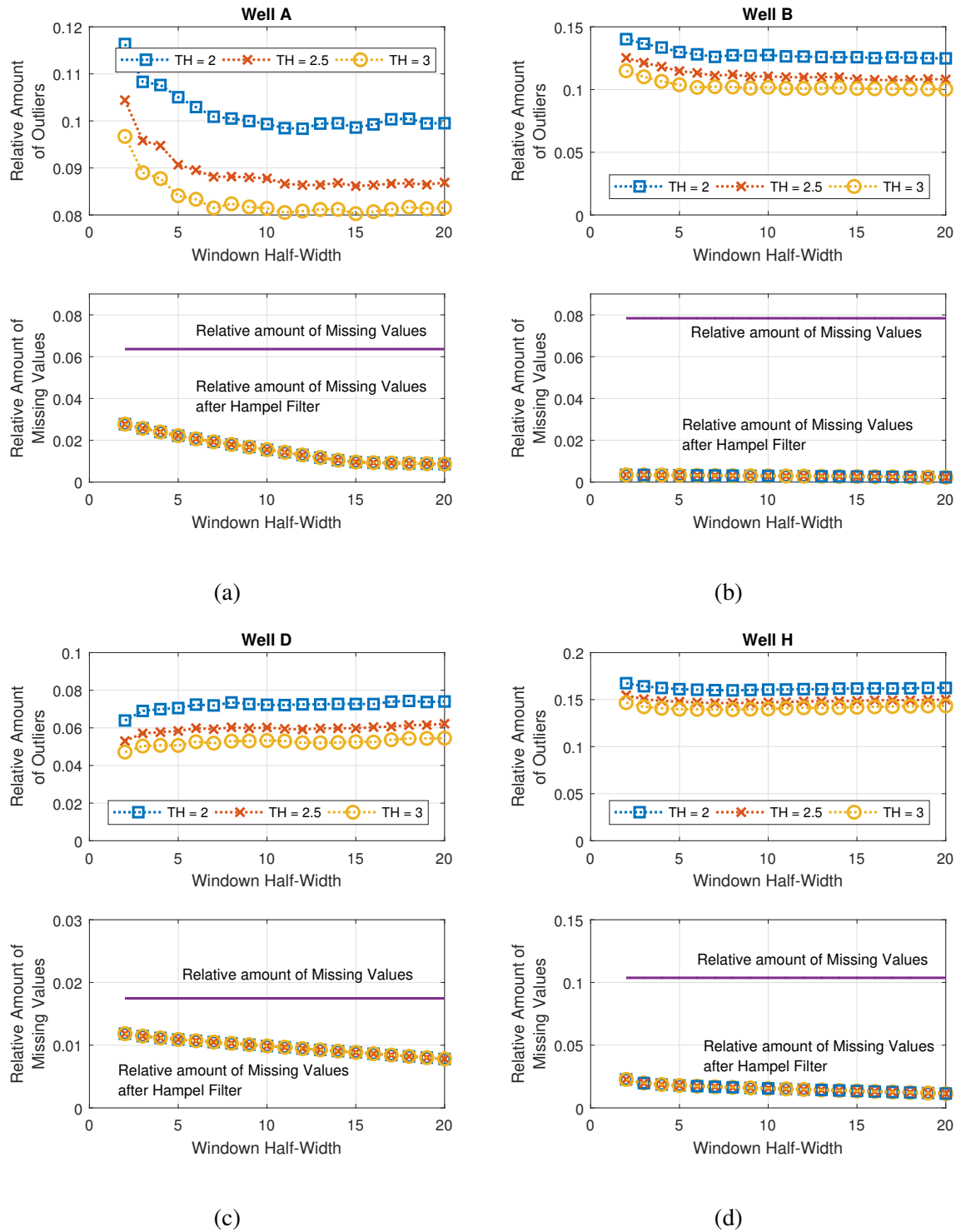
The main difference between both functions is that not-a-number (NaN) instances can be treated as outliers only in the `hampel` function. With `hampel` filter, some NaN are identified as outliers, and then replaced by the center value of the moving median. The `filloutliers` does not identified NaN as outliers. The second difference is the amount of identified outliers differs according to window width. With `hampel` function, the amount of detected outliers normally decreases with a wider window, while the amount of detected outliers increases with wider window for `filloutliers`.

In the current work, the `hampel` function is applied in the data pre-treatment process. Two parameters are required to be set: the amount of neighbors in the moving window, and the threshold for outlier detection. The Figure C.1 shows the influence of both parametrics on the amount of outliers detected and on the amount of missing values for pre-salt wells. The amount of detected outliers normally decreases with a higher number of neighbors in the moving window. On the other hand, the amount of missing values (NaN) treated increases with the wider moving window. The higher the threshold TH , the less outliers are identified. However, the threshold has no influence on the amount of NaN treated as outliers.

The relative amount of outliers for all drill-curves from pre-salt are shown in Figure C.2 with box-plots. For the drilling data from Norway, such box-plots are in the Figure C.3, where a similar behaviour to pre-salt data can be observed. Based on the analysis of these curves, the selected parameters for the current work are:

- number of neighbors in the moving window: 5;
- threshold: 3.

Figure C.1 – Parametric influence of Hampel filter on detecting outliers. Some missing values (not-a-number instances) are detected as outliers.



Source: own authorship.

Figure C.2 – Relative amount of outliers for different levels of window half-width and thresholds (TH) - all drill curves from pre-salt.

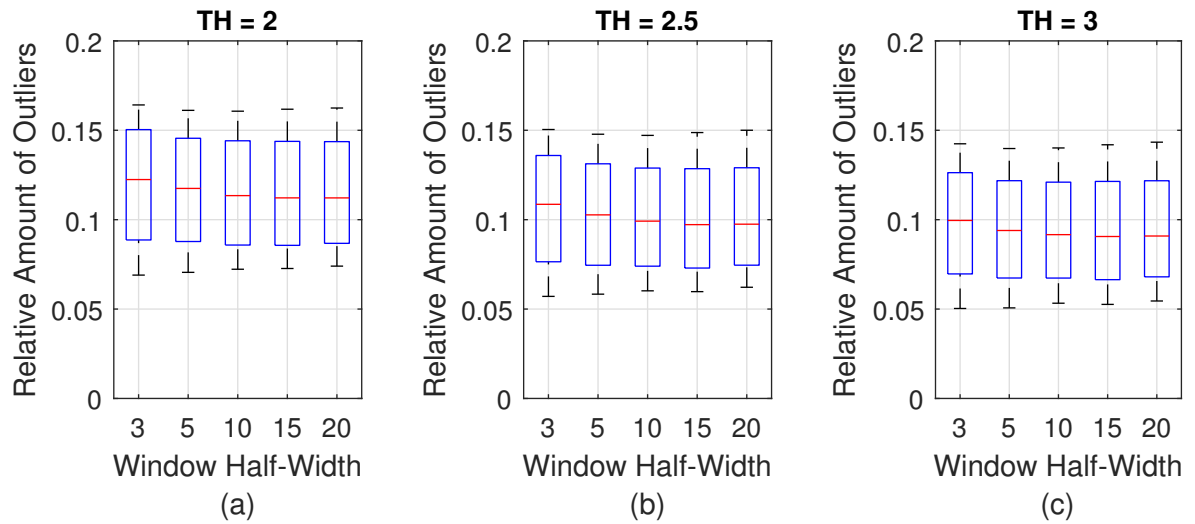
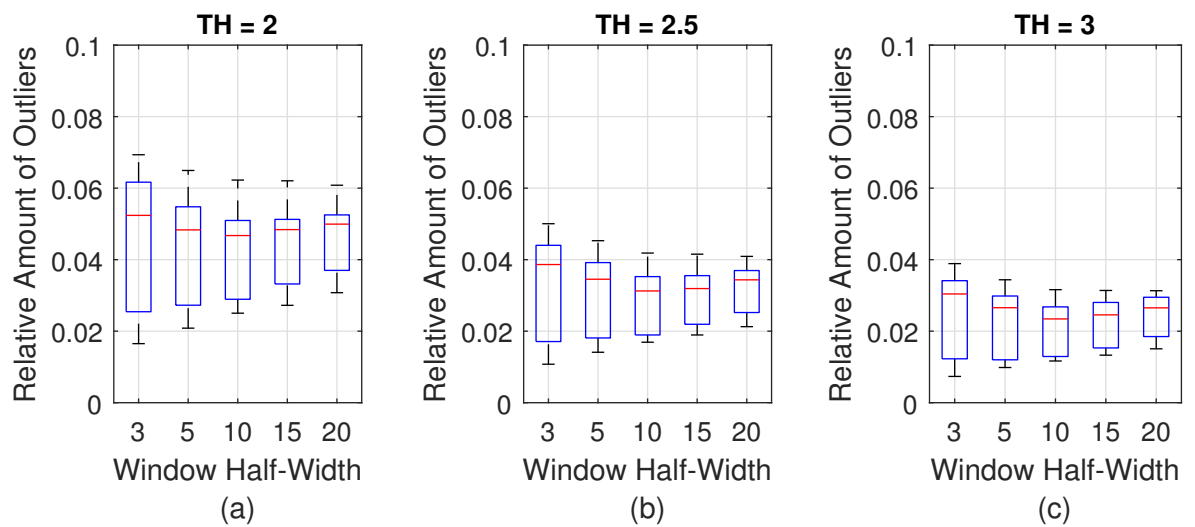


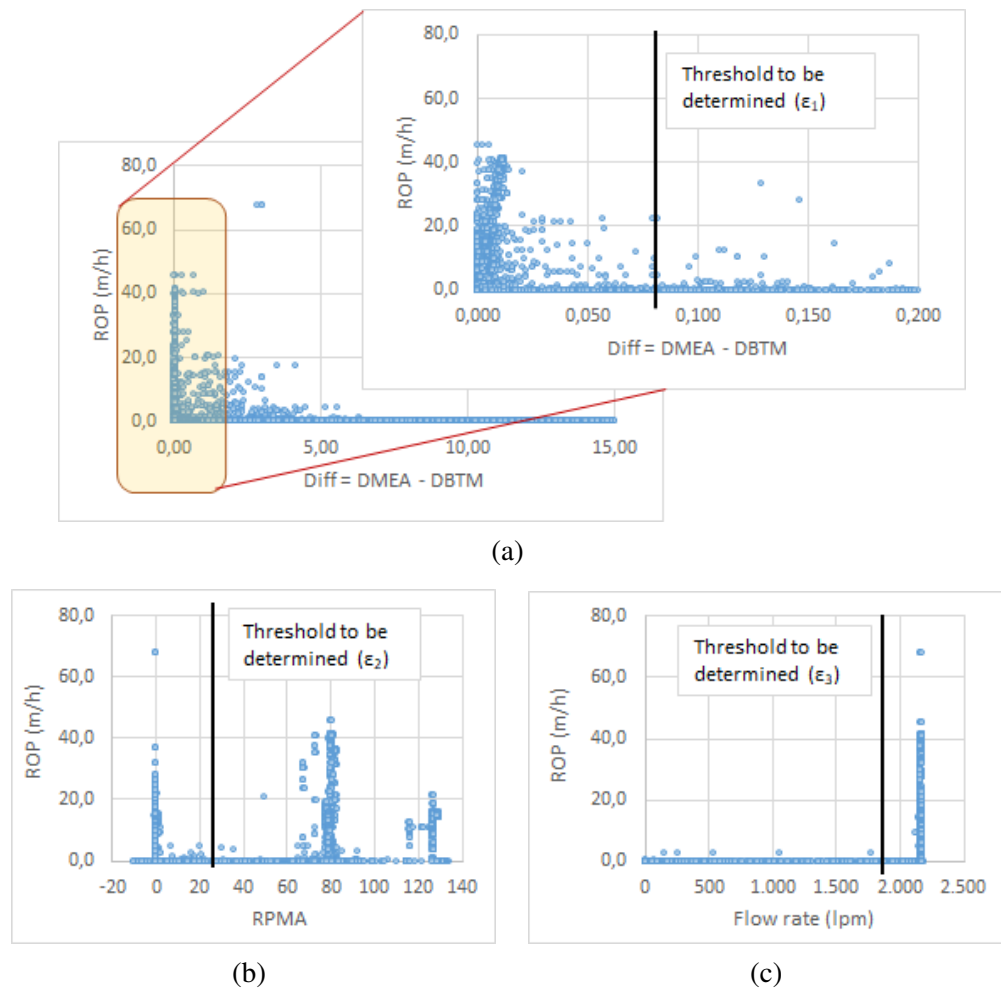
Figure C.3 – Relative amount of outliers for different levels of window half-width and thresholds (TH) - all drill curves from Norway, excluding the well 2.



APPENDIX D – DATA PRE-TREATMENT ANALYSIS OF OPERATION RECOGNITION

The key element for the automated operation recognition is the definition of the threshold ε for each condition: (i) bit on bottom hole, (ii) circulation is present, and (iii) drillstring is rotating. The Figure D.1 illustrates the visual analysis performed to determine the thresholds for each condition. With a little trial and error, the following values for the thresholds were set: $\varepsilon_1 = 0.15$ (m), $\varepsilon_2 = 100$ (lpm) and $\varepsilon_3 = 10$ (rpm) for the first condition, second condition and third condition respectively. For Norwegian drill curves from the second group (without hydraulic parameters recorded), it is not possible to determine the presence of circulation, so that the drill states are determined by the first and third conditions.

Figure D.1 – Visual approach to determine the thresholds for the automated operation recognition. For illustration purpose, drill curves from well 8 are plotted. Similar analysis for all other wells was carried out.

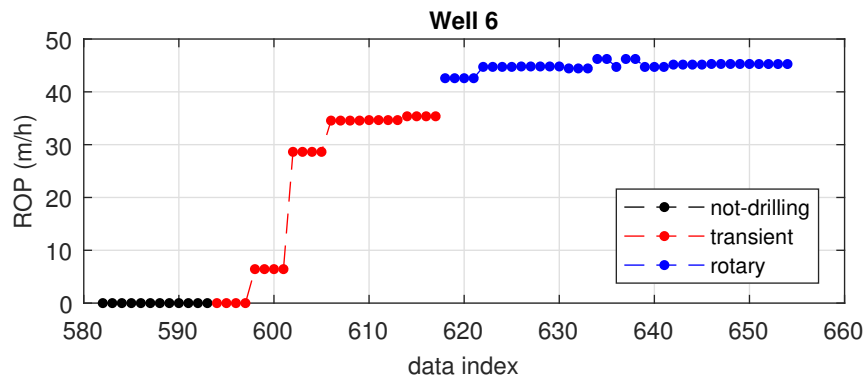


Source: own authorship.

The following plots illustrate some common results obtained with the transient recognition

code. Such plots were extensively analyzed in order to validate the rules employed to identify transient states. The Figure D.2 shows one of many examples when a change from not-drilling to rotary drilling was correctly recognized.

Figure D.2 – An example for a transient state correctly identified - well 6 - first change identified.

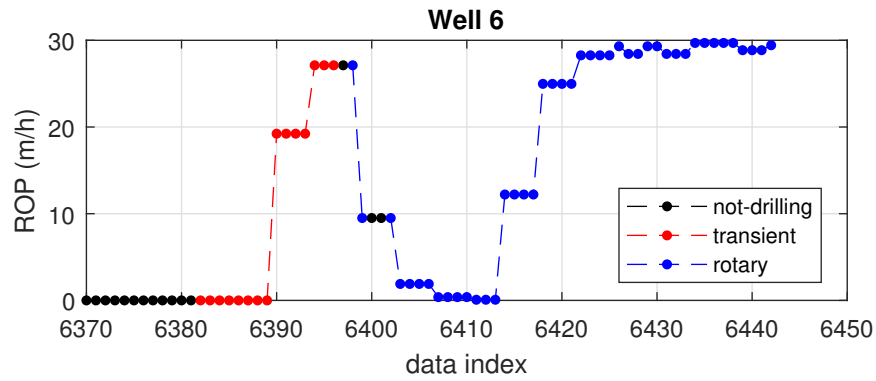


Source: own authorship.

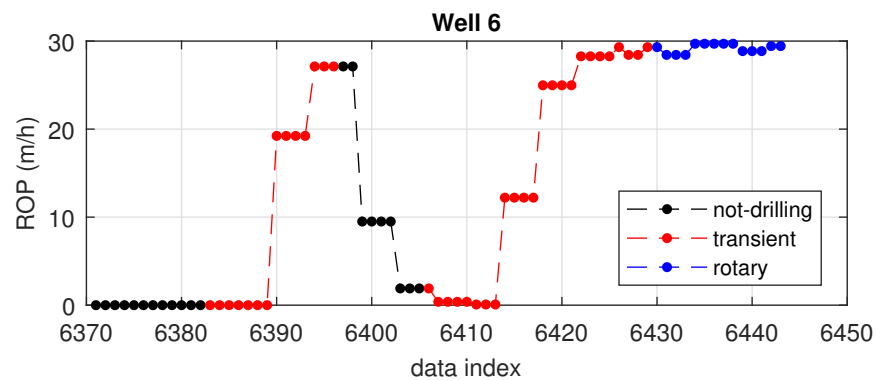
In the Figure D.3, the first change was correctly recognized, but the second transient state could not be recognized when $\varepsilon_1 = 0.2m$. After reducing this threshold to $\varepsilon_1 = 0.15m$, the second change was recognized as a transient state.

The Figure D.4 illustrates an interesting aspect of the automatic transient recognition. The code was developed to identify a change only in one way, i.e. from not-drilling to rotary drilling. Even though, a change in another direction, from rotary-drilling to not-drilling, was identified.

Figure D.3 – Reducing the ε_1 from 0.2m to 0.15m could improve the transient state recognition: a) the second change was not recognized when ε_1 was 0.2 m; b) the second change was recognized after reducing the tolerance to 0.15 m - well 6 - eighth change identified.



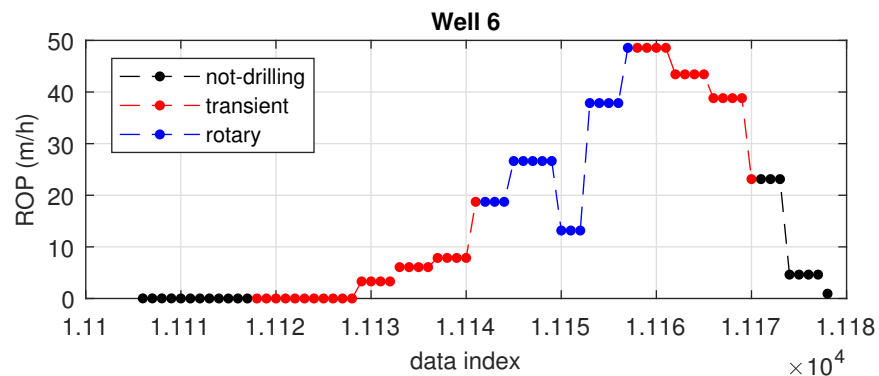
(a) $\varepsilon_1 = 0.2m$



(b) $\varepsilon_1 = 0.15m$

Source: own authorship.

Figure D.4 – An example for a transient states identified from not-drilling to rotary drilling, and vice-verse - well 6 - sixteenth change identified.

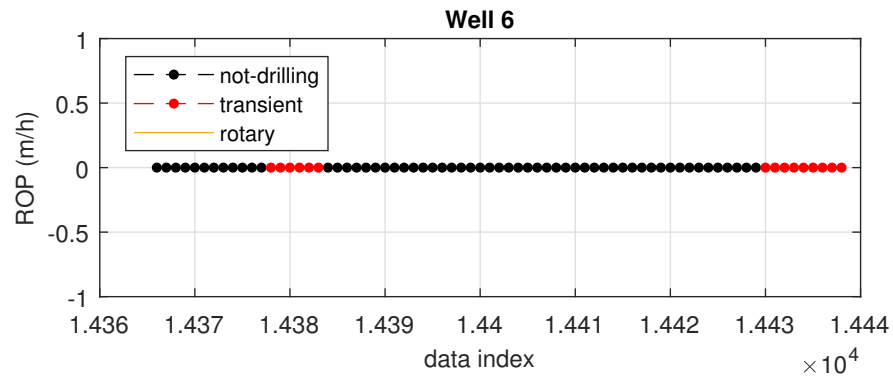


Source: own authorship.

The rules to identify the drilling modes are not free of errors. In Figure D.5, a long period of clearly not-drilling state had some observations considered as rotary-drilling. However, the

transient recognition code identified them as a change from not-drilling to drilling, and then changed the index to transient state. In this case, the transient state is acting to correct rotary drilling state wrongly identified.

Figure D.5 – The transient state detection could avoid considering some observations as rotary drilling - well 6 - twenty-first change identified.



Source: own authorship.

Despite several transient states could be correctly recognized, the developed code was not able to identify all possible transient states, especially when often changes from drilling to not-drilling occurred. However, we believe that the rotary drilling state can be systematically obtained, based on rules related to drilling engineering discipline, which is the main purpose of the data pre-treatment process presented in the current work.

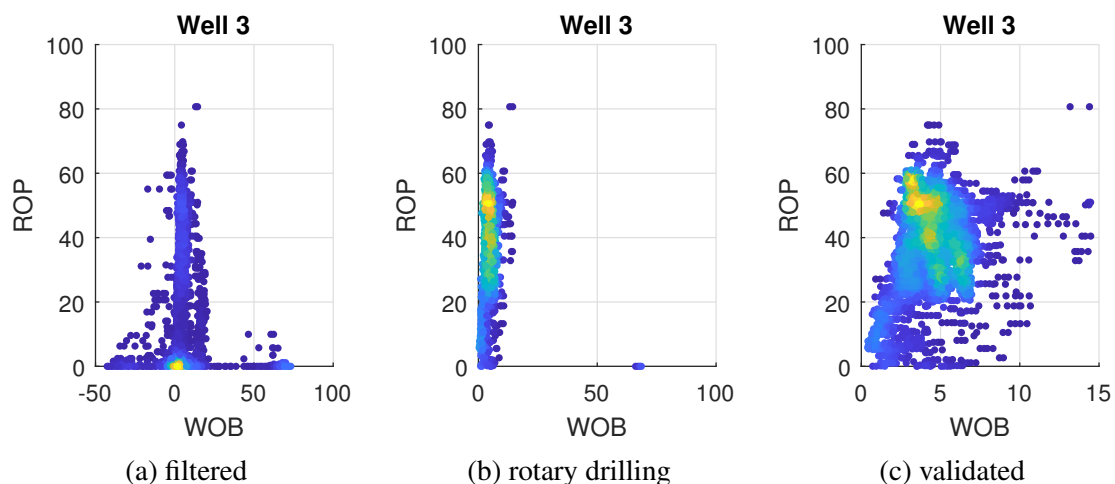
APPENDIX E – DATA PRE-TREATMENT ANALYSIS OF VALIDATION DRILL CURVES

The validation is the last step of data pre-treatment. After hampel filter and, for Norway dataset, automated operation recognition, some observations may still be invalid due to measurement errors. Two rules are employed to check whether an observation is valid or not:

- first rule: an observation is invalid if ROP value is above zero and WOB is less or equal than zero;
- second rule: an observation is invalid if WOB values is too high. In the case of Norway, the WOB is considered too high if WOB is above 45 ton. For the drill curves from pre-salt region, the threshold is 100 ton.

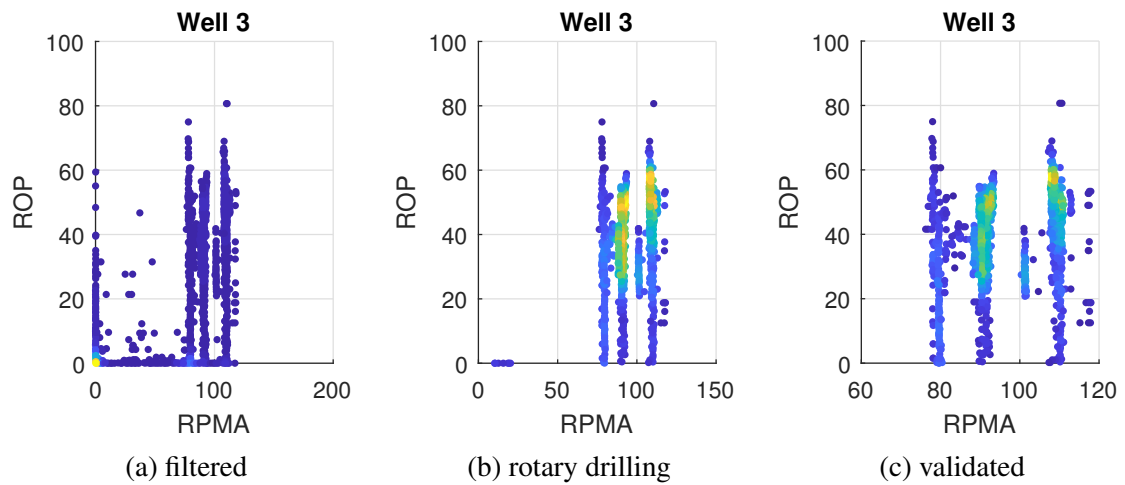
The validation step has a great impact on the drill curves from Norway. To illustrate it, some plots (two-dimensional histograms) are generated for the wells 3 and 4. Both drill curves has not the flow rate measured, so that the automated operation recognition can only be partially performed (the condition for presence of circulation cannot be checked). The plots in the Figure E.1 show two-dimensional histogram of ROP against WOB for the Well 3. After getting the instances considered as rotary drilling, it is possible to observe a small cluster with high values of WOB and ROP near to zero. After the validation step, this small cluster is not considered anymore as rotary drilling.

Figure E.1 – Two dimensional histogram of ROP against WOB for the Well 3: a) filtered data after step ③ with all drilling states, b) after step ④ with only rotary drilling state, c) after validation step ⑤.



Source: own authorship.

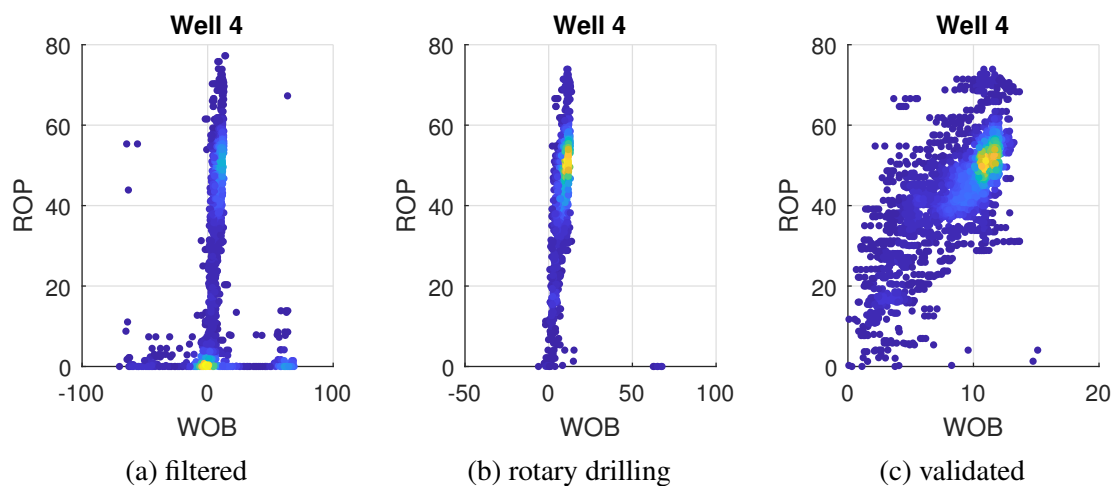
Figure E.2 – Two dimensional histogram of ROP against RPM for the Well 3: a) filtered data after step ③ with all drilling states, b) after step ④ with only rotary drilling state, c) after validation step ⑤.



Source: own authorship.

In the Figure E.2, it is possible to observe a small cluster with low values of RPM and ROP. After the validation step, this cluster was eliminated. This fact is interesting because the validation rules based on other parameters (WOB and ROP) could identify this small cluster. The Figure E.3 illustrates the validation step applying both rules. All plots in this Appendix were generated with a function, called `dscatter`, developed by Eilers and Goeman (2004).

Figure E.3 – Two dimensional histogram of ROP against WOB for the Well 4: a) filtered data after step ③ with all drilling states, b) after step ④ with only rotary drilling state, c) after validation step ⑤.



Source: own authorship.

APPENDIX F – TORQUE PREDICTION

We investigated the estimate accuracy of torque predictive models based on random forest regression, using only four variables as inputs: depth, surface weight on bit, drillstring rotary speed and total flow rate of all active pumps (if available). We followed the same train and test procedure employed to obtain other predictive models in the current works, such as the ROP. The number of variables selected at random for each decision split was set to two. After leaving out a random 20% of observations, we trained random forests with 100 trees on the remaining 80%, and used the left-out 20% as a test set. We repeated this procedure 100 times, and at each time with a different data partition, but maintaining the 80-20 data-partition. We computed the average of training and testing errors for each well, as well as the standard deviation (values shown in parentheses). For clarity, only the average of evaluation metrics are shown in the Table F.1.

Table F.1 – Torque prediction based on random forests: average of evaluation metrics on train and test dataset.

Well	Train dataset				Test dataset			
	MAPE _a	MAE	MSE	R	MAPE _a	MAE	MSE	R
<i>Pre-Salt</i>								
A	-	-	-	-	-	-	-	-
B	0.023	0.093	0.018	0.912	0.035	0.140	0.038	0.790
D	-	-	-	-	-	-	-	-
H	0.294	1.038	3.072	0.905	0.437	1.519	6.340	0.759
<i>Norway</i>								
3	0.008	127.573	5.193E+04	0.994	0.013	194.994	1.153E+05	0.987
4	0.005	96.046	5.051E+04	0.978	0.008	146.604	1.127E+05	0.949
5	0.007	95.025	6.205E+04	0.953	0.011	143.052	1.311E+05	0.899
6	0.008	68.931	2.420E+04	0.973	0.011	102.686	4.318E+04	0.953
7	0.009	132.091	5.269E+04	0.985	0.013	198.930	1.149E+05	0.967
8	0.007	99.877	2.865E+04	0.987	0.011	148.646	5.986E+04	0.973
9	0.014	338.704	2.561E+05	0.983	0.021	522.945	5.924E+05	0.959
10	0.023	334.674	2.537E+05	0.970	0.034	505.713	5.529E+05	0.932

Source: own authorship.

ANNEX A – ABBREVIATIONS FOR REAL-TIME DRILLING DATA FROM PRE-SALT

Table A.1 – Mnemonics employed in drilling data from Pre-Salt

Mnemonic	Description	Unit
AJAM	Anti-Jams Detected	-
BLKA	Block Acceleration	m/s^2
BLKP	Height of block above rig floor	m
CDEPTH	Composite Depth	m
CRPM	Collar Rotational Speed	c/min
DEPT	Depth Index	m
DHAP	Downhole Annulus Pressure	psi
DHAT	Downhole Annulus Temperature	$degC$
DTOR	Downhole Torque (MWD)	$ft.lbf$
DWOB	Downhole Weight on Bit	lbf
ECD	Equivalent Circulating Density	lbm/gal
GR	Gamma Ray	$gAPI$
GR_CAL	Calibrated Gamma Ray	$gAPI$
HKLA	Average Hookload	lbf
PVEL	Pipe Velocity	m/h
ROP	Rate of Penetration	m/h
ROP5	Rate of Penetration Averaged over the last 5 ft	m/h
RPM	Rotational Speed	c/min
SHKL	Total Shock Level	-
SHKR	Shock Rate	$1/s$
SHKRSK	Shock Risk	-
SHOCKS_ACC	Total shocks accumulated over tool life time	-
SPPA	Standpipe Pressure	psi
STICK	Stick Slip Indicator	c/min
STICKRATIO	Stick Ratio	-
STOR	Surface Torque	$ft.lbf$
STUCK	Percent Time Stuck Below 5 RPM Indicator	-
SWOB	Surface Weight on Bit	$klbf$
TFLO	Total Flow Rate of all active pumps	gal/min
TRPM	MWD Turbine Rotation Speed	c/min
TVDE	True Vertical Depth	m
VIB_LAT	Transverse RMS Vibration	gn
VIB_X	RMS Vibration X-Axis	gn

Source: Nascimento (2016).

ANNEX B – ABBREVIATIONS FOR REAL-TIME DRILLING DATA FROM NORWAY

Table B.1 – Mnemonics employed in drilling data from Norway

Mnemonic	Description	Unit
BPOS	Block Position	<i>m</i>
CEPP	Cement Pump Pressure - Time Based	<i>bar</i>
CFI	Cement Flow In - Time Based	<i>l/min</i>
CHP	Choke Pressure - Time Based	<i>bar</i>
CTVL	Cementing Total Volume Pumped - Time Based	<i>m³</i>
DBTM	Bit Depth (MD)	<i>m</i>
DEPT	Bit Depth	<i>m</i>
DMEA	Hole depth (MD)	<i>m</i>
ECDB	Effective Circulating Density at Bit - Time Based	<i>g/cm³</i>
ECDM	Measured Effective Circulating Density at bit	<i>g/cm³</i>
GAS	Total gas in mud - Time Based	<i>%</i>
HKL	HookLoad - Time Based	<i>tonne</i>
KLP	Kill Line Pressure - Time Based	<i>bar</i>
MDI	Mud Density in average - Time Based	<i>g/cm³</i>
MDO	Mud Density out average - Time Based	<i>g/cm³</i>
MFI	Mud Flow in average - Time Based	<i>l/min</i>
MFO	Mud Flow out average - Time Based	<i>l/min</i>
MTI	Mud Temperature in - Time Based	<i>degC</i>
MTO	Mud Temperature Out - Time Based	<i>degC</i>
ROP	Rate of Penetration	<i>m/h</i>
RPM	Average Rotary Speed	<i>rev/min</i>
RPMA	String RPM average	<i>rpm</i>
RPMB	Bit RPM average	<i>rpm</i>
SPP	Stand Pipe Pressure average - Time Based	<i>bar</i>
SPPA	Average Standpipe Pressure	<i>kPa</i>
SWOB	Weight on Bit	<i>1000 kgf</i>
TPVT	Trip pit volume totaliser - Time Based	<i>m³</i>
TRQ	Torque - Time Based	<i>kN.m</i>
TVA	Active Tank Volume	<i>m³</i>
WOB	Weight on bit - Time Based	<i>tonne</i>

Source: Skalle (2018).