Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

SIMULATION OF OIL DRILLING TIME SERIES USING MONTE CARLO AND BAYESIAN NETWORKS

Mariana Dehon Costa e Lima
Silvia Modesto Nassar
Paulo José de Freitas Filho

Departamento de Informática e Estatística - PPGCC
Universidade Federal de Santa Catarina
Florianópolis, SC 88040-900, BRAZIL

ABSTRACT

During oil drilling the main goal is to minimize the total cost of the process. There are two major costs: the cost of the drills and the operation cost. It is necessary to find the best combination of bits in order to maximize the Rate of Penetration (ROP). Many environmental and operational variables influence ROP, but the relationship between them is not always clear. In addition, the lack of historical data makes this problem an even bigger challenge. This paper proposes an approach using Bayesian Networks with the Monte Carlo simulation for generating data for an oil drilling process and compares it with the historical data.

1 INTRODUCTION

The prediction of Rate of Penetration (ROP) is characteristic of environments with high complexity and risk that seek to optimize the cost of the drilling process. The minimization of those costs are directly related with the maximization of the ROP.

In order to reduce costs, it is necessary to optimize the project of the drilling. The necessary time of the drilling process needs to be estimated with a high precision, since most part of the costs are related to the rent of the necessary equipments for this operation (Gandelman 2012).

However, each operation has its unique properties which make this an extremely hard task. Many of theses properties vary during the drilling, like the type of rock, the presence of gas, the pressure, the bit wear rate, among others. All those properties affect not only the ROP but also the operation parameters.

An oil drilling process have variables direct related to the Rate of Penetration, that can be either controlled by the drill operator or are related to the geology. In this problem, the ROP is the output variable and the set of input variables is: UCS (Unconfined Compressive Strength), WOB (weight on bit), RPM (revolutions per minute), HSI (hydraulic horsepower per square inch) and the bit wear rate.

In the perforation process the ROP decreases with the bit wear rate. In this paper we use the variable "perforated meters" to express this relation. The variable has linear and accumulative behavior that increases its value by one for each perforated meter.

During the data collection some sensors can present failures and by consequence, the data can have inconsistencies. Those sensors can not work for a period or transmit wrong and unreliable data. When a sensor failure happens it's necessary to make a decision: eliminate the data or modify them.

A possible approach to fill the spaces on the database or modify the inconsistent data is using a simulation methodology such as Monte Carlo (MC) (Hammersley and Handscomb 1964). Using those methods, it's possible to create a simulated temporal series with similar features of the historical series.

This paper uses the simulation technique of MC combined with a Bayesian Network (BN) approach (Pearl 1988). The MC technique was used for simulate scenarios considering the input variables and the BN was used to estimated the output variable (ROP).

The proposed hybrid method can be used when the function of the output variable isn't clear or known. The BN is created using the PVD method of discretization which was developed in a previous work (Lima, Nassar, de Freitas Filho, et al. 2014) and has optimistic results on the field.

The mainly goal of the proposed method is finding a simulation model based on the MC technique capable of replicate the behavior of the historical data using Bayesian Networks to estimate the output variable, in this case, the ROP.

This paper is organized as follow: Section 2 shows the related work for this study. Section 3 presents a brief overview over the Bayesian Networks and Monte Carlo techniques. Section 4 describes the proposed method. Section 5 shows the results for this approach and finally on Section 6 we show the discussion about the results and conclude the study.

2 RELATED WORK

(Reed 1972) proposes an approach an optimization method based on field experience using Monte Carlo. The optimization consists in finding the minimum value of cost per foot (CPF) of a drilling process subject to certain constraints. The Monte Carlo approach is used here to find parameters for the CPF function. The method was applied on several databases. The authors conclude that the developed MC method was proved valuable because of its rigor and the remarkable ease of constraint inclusion.

(Engelhardt and Macdonald 2008) presents an approach using MC simulation methods for describing the development of localizes corrosion damaged. The method was proposed for predicting the pitting corrosion damage and was effectively describing the progression of damage for either when several or even one pit is alive. According the authors, with the proposed method is possible to reduce the number of unknown parameters of the interaction between pits.

(Komlosi and Komlosi 2009) affirms that the fundamental issue in reserver estimation is the volume of hydrocarbon that can be economically recovered from the reservoir. It's a complex task, mainly because of the limited amount of the information. That's the main reason why they use the MC procedures on this study. The mainly focus on this article is in estimate the probability density function used on the MC, for the variables in three different fields. With the result the authors conclude that MC simulation is a good tool for (technical) reserve assessment.

(Mireault 2013) presents a methodology to prepare an aggregate production forecast for unconventional resource drilling using MC simulation and quantify the variance for a small group of locations that is undrilled. The approach was tested on a 10-location drilling example that illustrate the variance in forecasts over a 20-year project life. With the results, the authors conclude that a better management of uncertainty and financial risk should be lead by the resulting confidence interval forecasts.

(Steinmann and de Freitas Filho 2013) presents a study of a Monte Carlo simulation used to generate data to evaluate incoming call forecasting algorithms. The study was motivated because of the lack of available records for call center data series. The simulator used three elements: an incoming call generator, a random event generator and a call logger. To evaluate the study, a data series from a call center was used. With the results the authors concludes that generating synthetic data with proposed method could be an interesting alternative for the field and countless different scenarios can be tested and random factor can be included with varying impacts.

3 BACKGROUND

3.1 Bayesian Networks

A Bayesian Network (BN) (Pearl 1988) is a model of representation and reasoning of uncertainty that uses the conditional probability between variables of a specific domain, expressed by Directed Acyclic

Graphs (DAG). Its graphical structure can tackle correlations between variables effectively, with appropriate language and efficient resources to represent the joint probability distribution over a set of random variables (Friedman and Goldszmidt 1996).

Defining formally, a BN is a pair (G,P), where $G=(V,E)$ is a DAG. The nodes $V=\{v_1,,v_n\}$ represent the variables and edges $E=\{e_1,,e_m\}$ represent a direct correlation between each node in V.

P is defined as a set of probabilistic parameters expressed through tables. Given a particular variable, a conditional probability distribution is made for each of their classes/values $X=\{x_1,,x_z\}$ joining each classes/value of their parents.

With that configuration, the network establishes that a variable is independent of all other variables except their descendants in the graph, given the state of its parents. The inference inside the network is done by the Bayes theorem:

$$P(V=v|X=x)=\frac{P(X=x|V=v)*P(V=v)}{P(X=x)}$$

The joint probability is determined by the called chain rule and assumes the conditional independence between the variables:

$$P(v_1,\ldots,v_n)=\prod_{i=1}^{n}P(V_i|parent(V_i))$$

where $parent(V_i)$ determines the set of parent nodes from V_i.

The BN reasoning is established in two distinct scenarios:

$$\begin{cases}\textbf{if } \text{“input”} \textbf{ then } \text{“output”}\\ \textbf{if } \text{“output”} \textbf{ then } \text{“input”}\end{cases}$$

Considering all the possible network topologies for a Bayesian network the well-known structure Naïve Bayes is the simplest one. It assumes that all variables are mutually independent given the class context. Although this model does not reflect the reality in most real-world tasks it is very effective, because the parameters of each attribute can be learned separately, facilitating the learning process (McCallum and Nigam 1998). The naïve Bayes topology is therefore a set of mutually independent variables that works as the input which collectively has a single parent (output node).

3.1.1 Peak-Valley Discretization Method

The Bayesian inference in cases where there are quantitative variables on the domain is not always exact and is strongly associated with the application domain. Discretization can handle this problem converting each quantitative value (x_i) into a qualitative value (x_i^*) under some predefined criteria, but information loss may become an issue (Yang and Webb 2009).

The Peak-Valley Discretization (PVD) (Lima, Nassar, de Freitas Filho, et al. 2014) is a parametric method and assumes that quantitative variables $v_i \in V$ have values that can be either on an extreme or in an intermediate range. Analyzing those ranges it's possible tod establish the probability conditionals of the BN.

The range delimitation on PVD is expressed with two cut-points: the first one (**peak**) selects the area considered “high” and the second one (**valley**) selects the area considered “low”. The intermediate area is considered “moderate”. The cut-points are expressed in percentile in order to include the frequency concept on the algorithm.

The PVD method finds the cut-points using a Genetic Algorithm (GA) (Goldberg and Holland 1988) where the fitness function seeks the accuracy maximization or the error minimization.

During the GA search its possible to find a cut-point that doesn't improve the result and is not necessary for the model. It happens if the range limited by the cut-points is close to the boundaries or too close to each other. In that case the PVD can merge or eliminate the cut-point.

Considering the possibility of merging or ignoring a cut-point the PVD can have two or three classes in each variable. The topology adopted on PVD is the Naïve Bayes.

3.2 Monte Carlo Simulation

This technique involves the random sampling of each probability distribution within the model to produce hundreds or even thousands of scenarios. Each probability distribution is sampled in a manner that reproduces the distribution's shape.

The distribution's shape can be reproduced using a function called PDF (probability distribution function). Distribution fitting is a process that has the objective of find the probability distribution that has the best fit to a series of data points. The methods of distribution fitting are divided in two categories: parametric and regression methods.

The distribution of the values calculated for the model outcome therefore reflects the probability of the values that could occur.

Monte Carlo simulation offers many advantages over the other techniques presented above:

- The distributions of the model's variables do not have to be approximated in any way;
- Correlation and other interdependencies can be modeled;
- The level of mathematics required to perform a Monte Carlo simulation is quite basic;
- The computer does all of the work required in determining the outcome distribution;
- Software is commercially available to automate the tasks involved in the simulation;
- Complex mathematics can be included (e.g. power functions, logs, IF statements, etc.) with no extra difficulty;
- Monte Carlo simulation is widely recognized as a valid technique, so its results are more likely to be accepted;
- The behavior of the model can be investigated with great ease;
- Changes to the model can be made very quickly and the results compared with previous models.

Monte Carlo simulation is often criticized as being an approximate technique. However, in theory at least, any required level of precision can be achieved by simply increasing the number of iterations in a simulation. The limitations are in the number of random numbers that can be produced from a random number generating algorithm and, more commonly, the time a computer needs to generate the iterations.

3.3 Random Sampling from Input Distributions

Consider the distribution of an uncertain input variable x. The cumulative distribution function $F(x)$ gives the probability P that the variable X will be less than or equal to x, i.e. $F(x) = P(X \leq x)$. $F(x)$ obviously ranges from $0 to 1$. Now, looking at this equation in the reverse direction: what is it is possible to find the value of $F(x)$ for a given value of x. This inverse function $G(F(x))$ is written as $G(F(x)) = x$. It is this concept of the inverse function $G(F(x))$ that is used in the generation of random samples from each distribution in the model. Figure 1 provides a graphical representation of the relationship between $F(x)$ and $G(F(x))$.

To generate a random sample for a probability distribution, a random number r is generated between 0 and 1. This value is then fed into the equation to determine the value to be generated for the distribution:

$$G(r) = x$$

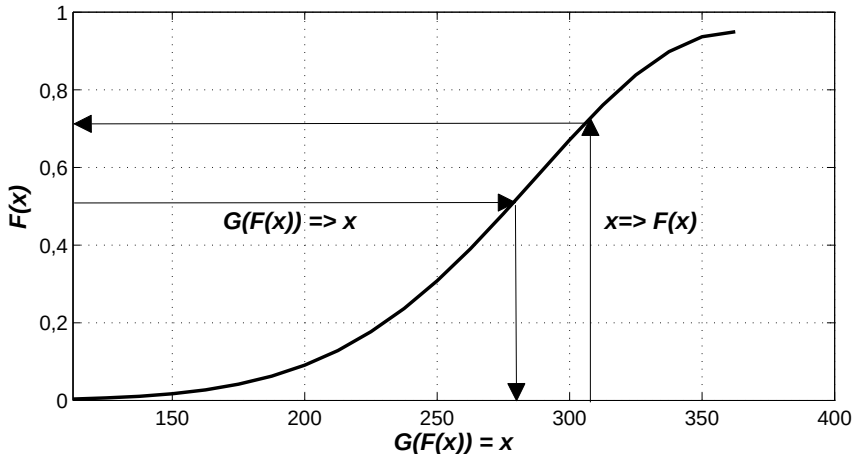

Figure 1: Representation of the relationship between $F(x)$ and $G(F(x))$.

The random number r is generated from a $Uniform(0,1)$ distribution to provide equal opportunity of an x value being generated in any percentile range. The inverse function concept is employed in a number of sampling methods like Monte Carlo.

4 OUR APPROACH

In our approach we simulate an oil drilling data using a hybrid methodology between Bayesian Networks and the MC. The variables used on the problem are shown on the Table 1.

Table 1: Input and Output variables on the drilling time series.

Variable	Type	Metric
WOB	input	klb
RPM	input	rpm
HSI	input	hp/pol^2
UCS	input	psi
perforated meters	input	m
ROP	output	m/h

The MC method is used to simulate the input variables of this problem. In order to simulate those values it's necessary to define the probability density function (PDF) of each variable. In this work we use a fitting algorithm on the historical data to define the PDFs.

Since the function of the output variable (ROP) is unknown, the BN method is used to estimate its value with the Naïve Bayes reasoning. However, some preparation is necessary because the traditional BN only uses discretized variables on its structure. In this paper we used a discretization method called PVD (Lima, Nassar, de Freitas Filho, et al. 2014) as described on Section 3.1.1.

After training the BN with historical data, it's possible to get a numeric value for the output node considering the probability distribution of its classes. The follow function returns the expected quantitative value of the output node in BN, based on current beliefs and a list of real numbers that represent each class:

$$\sum_{i=1}^{n} belief_i * value_i$$

where n is the number of classes, the $belief_i$ is the output belief of the class and the $value$ is the value that represents the class.

For example, consider an output node with 3 classes and each class represents a numeric value and has a related belief (Table 2).

Table 2: Example of values that express each class of an output variable.

Class	Value	Belief
class1	10	0.7
class2	20	0.1
class3	30	0.2

In that case the output value is expressed by:

$$(0.7 * 10) + (0.1 * 20) + (0.2 * 30) = 15$$

We consider the midpoint of each variable on the historical data as the "value" that represents the variable.

The entire process is shown on the flowchart expressed at the Figure 2.

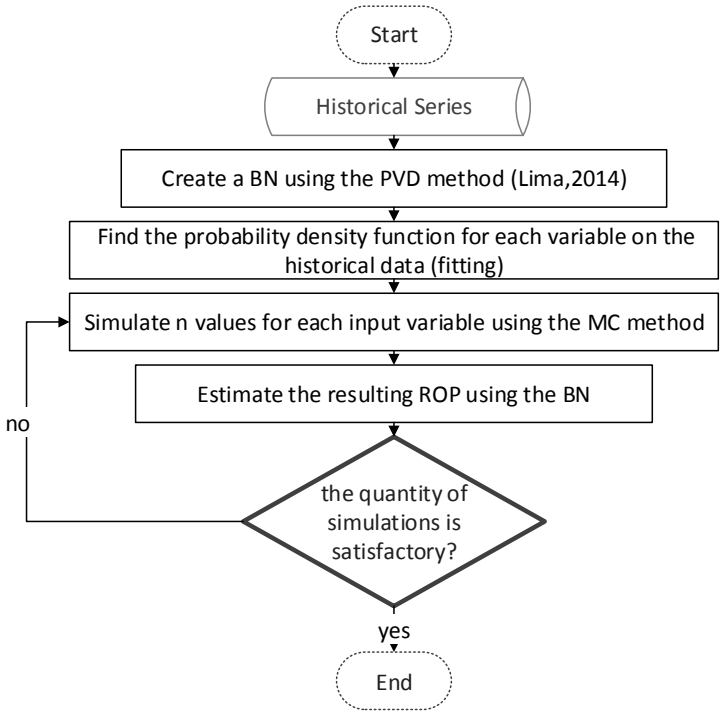

Figure 2: Flowchart of the presented methodology.

5 RESULTS

The proposed method was tested in a dataset of Bit's Rate of Penetration Problem with 548 data points about the drilling process of a specific type of bit. The domain variables and how they are related to the process are expressed on Figure 3.

Following the flowchart expressed on the Figure 2 a BN was generated using the PVD method (Figure 4).

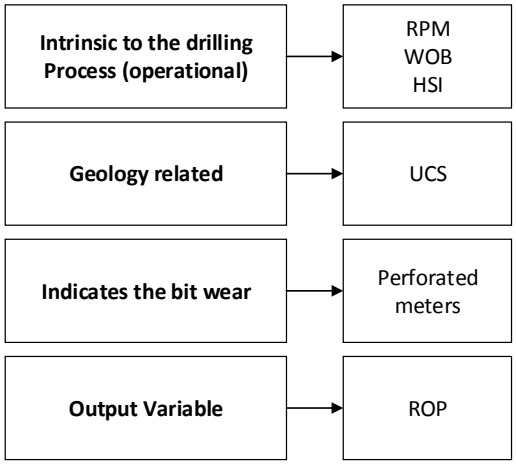

Figure 3: Domain variables and how they relate to the Oil Drilling Problem.

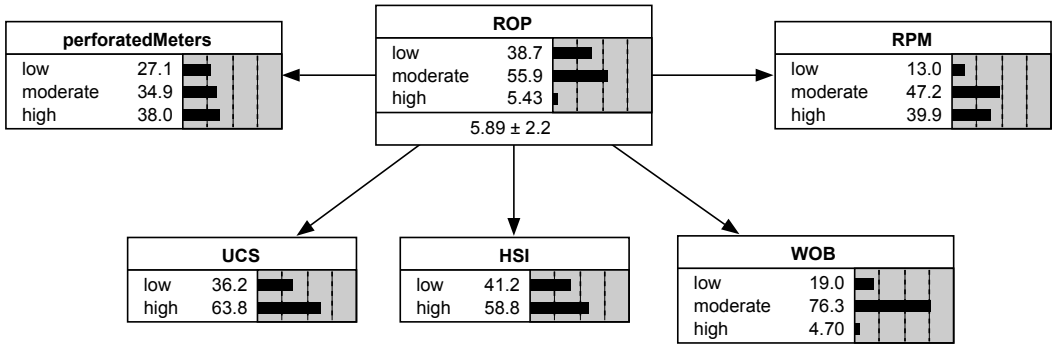

Figure 4: Generated BN using the PVD method. The expected value considering the probability distribution of the ROP is 5.89 and its standard deviation is 2.2.

A fitting algorithm was applied to find probability density function for each variable on the historical data. The obtained functions are shown on the Table 3.

Table 3: Probability Density Functions for each variable on the Drilling Time Series.

Variable	PDF
WOB	Weibull
RPM	Beta
HSI	Gaussian
UCS	Minimal Extreme
perforated meters	Integer - Uniform

In order to establish the confidence interval for mean considering the significance level of 0.05 (95% confidence) and 0.01 (99% confidence), 15 independent simulations $S = \{s_1, \cdots, s_{n_initial}\}$ were made where $n_{initial} = 15$ is the initial number of simulations.

In order to properly validate the generated data in comparison with the historical data, we simulated 578 data points (same number of the historical series) for each variable on the dataset. The ROP graph (output function) of the mean for each simulation is on Figure 5, and it shows a low variation on the values. We can observe the same pattern on the others variables.

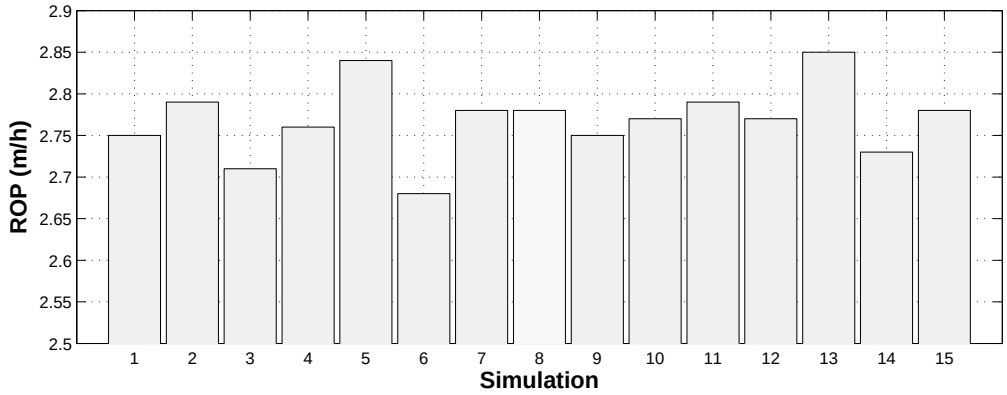

Figure 5: Mean of the ROP for each simulation.

After the initial 15 simulations it was calculated the Confidence Interval (CI) for mean using Student's t-distribution considering 95% and 99% confidence. The resultant CI was satisfactory for the problem and the the final number of simulations didn't change resulting in $n_{final} = n_{initial} = 15$. The descriptive analysis of the simulations are shown at Table 4.

Table 4: Descriptive analysis parameters of the Drilling Time Series simulation.

			CI for mean (95%)		CI for mean (99%)	
	mean	standard deviation	lower	higher	lower	higher
perforated meters	136.776	2.653	135.316	138.236	134.757	138.795
WOB	12.475	0.276	12.323	12.626	12.265	12.685
RPM	75.711	1.908	746.662	76.761	74.260	77.163
UCS	21333.349	252.315	21194.520	21472.178	21141.360	21525.339
HSI	1.175	0,017	1.166	1.184	1.162	1.187
ROP	2.768	0.043	2.744	2.792	2.735	2.801

For most variables, the simulations mean doesn't have a higher variance when compared to the historical data mean. It also happens with others statistics parameters like standard deviation, the midpoint, the coefficient of variation and the minimum and maximum values. In order to illustrate this fact we show the results of a single simulation $s_k \in S$ on Table 5.

Table 5: Descriptive analysis parameters of the Drilling Time Series simulations.

	mean		standard deviation		coefficient of variation		median		minimum		maximum	
	hist.	simu.	hist.	simu.	hist.	simu.	hist.	simu.	hist.	simu.	hist.	simu.
perf.meters	137.50	136.41	79.10	79.60	0.58	0.58	137.5	134.0	1	1	274	274
WOB	12.43	12.49	7.18	7.33	0.58	0.59	11.47	11.36	0.22	0.18	33.30	44.52
RPM	56.01	73.10	48.68	56.81	0.87	0.78	23.10	64.73	12.20	12.20	136.70	136.70
UCS	20993.33	21680.08	72221.87	6348.03	0.34	0.29	21947.69	22252.07	6358.22	1026.38	32225.13	35109.13
HSI	1.17	1.18	0.31	0.32	0.27	0.27	1.15	1.18	0.61	0.27	3.68	2.19
ROP	3.03	2.75	1.87	1.11	0.62	0.40	2.63	3.23	0.29	0.98	9.52	4.16

In the Table 6 we compare the simulations mean shown at Table 4 with the historical ones. Analyzing the relation between them, it's possible to identify a highest discrepancy on the RPM. One of the reasons of this difference is on the fact that the variable has "activity points", that are frequencies of operation with possible values that it can assume (Figure 6).

Table 6: Historical and Simulated mean for each variable.

	mean	
	historical series	simulated series
perforated meters	137.50	136.776
WOB	12.43	12.475
RPM	56.01	75.741
UCS	20993.33	21333.349
HSI	1.17	1.175
ROP	3.03	2.768

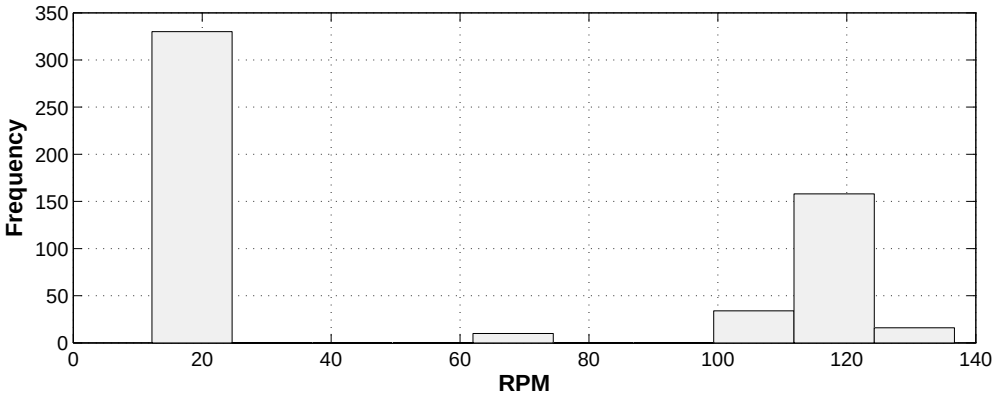

Figure 6: Historical histogram for the variable RPM.

The variable UCS also shows discrepant values, when the minimum values between the historical and the simulated time series are analyzed.

6 DISCUSSION AND CONCLUSION

Analyzing the results it's possible to observe that the simulated data were capable to properly reproduce most of the statistical indicators. The variable with the most discrepant behavior is the RPM. The discrepancy occurs because the probability density function didn't have a high adherence with the historical data like the other variables.

The variable UCS presents a discrepancy on the "minimum value" when we compare with the historical series and it happens for the same reason. The PDF belong to the interval $]0, +\infty]$ and there is a chance, even that it's a small one, that we simulate lower values than the ones on the historical data.

The ROP estimation was made with a BN that utilizes a method of discretization (PVD). The method cover the range of $[-\infty, +\infty]$, so even with the lower values of RPM it's possible to estimate the corresponding ROP since the output function would discretized the lower value as its lowest class on the BN.

During the simulation we assume that a variable is independent of the others. Therefore, the value of a variable x doesn't affect the value of a variable y. However, during the drilling process we can have situations were a specific range of x never occurs at the same time with a specific range of y. Investigate the relations between the input variables and their possible restriction is a possible complement to this study.

Since the ultimate goal of an oil drilling process is to maximize the ROP the prediction of its value is always a concern. However, often there are a lack of historical data and possible inconsistencies on them. The use of synthetic data can also reinforce some aspects there are not sufficient represented on the historical one.

In this work we propose a model to generate synthetic data using a hybrid method that combines the MC techniques with a BN reasoning. Analyzing the results of the simulation and the statistical indicators associated to them we can conclude that the simulated variables could, in most cases, reply the behavior of the historical data.

REFERENCES

Engelhardt, G. R., and D. D. Macdonald. 2008. "Monte Carlo Simulation Of Localized Corrosion". In *CORROSION 2008*. NACE International.

Friedman, N., and M. Goldszmidt. 1996. "Discretizing Continuous Attributes while Learning Bayesian Networks". In *13th International Conference on Machine Learning (ICML)*, 157–165. Morgan Kaufmann Publishers.

Gandelman, R. A. 2012. "ROP Prediction and Real-Time Optimization of Operational Parameters on Drilling Offshore Oil Wells". Master's thesis, Federal University of Rio de Janeiro.

Goldberg, D. E., and J. H. Holland. 1988. "Genetic Algorithms and Machine Learning". *Machine learning* 3 (2): 95–99.

Hammersley, J. M., and D. C. Handscomb. 1964. *Monte carlo methods*, Volume 1. Methuen London.

Komlosi, Z. P., and J. Komlosi. 2009. "Application of the Monte Carlo Simulation in Calculating HC-Reserves". In *EUROPEC/EAGE Conference and Exhibition*. Society of Petroleum Engineers.

Lima, M. D. C., S. M. Nassar, P. J. de Freitas Filho et al. 2014. "Heuristic Discretization Method for Bayesian Networks". *Journal of Computer Science* 10 (5): 869–878.

McCallum, A., and K. Nigam. 1998. "A Comparison of Event Models for Naive Bayes Text Classification". In *AAAI-98 Workshop on Learning for Text Categorization*, 41–48.

Mireault, R. 2013. "Aggregate Production Forecasts for Unconventional Resource Drilling Locations Using Monte Carlo Simulation". In *SPE Unconventional Resources Conference Canada*. Society of Petroleum Engineers.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausble Inference*. Morgan Kaufmann Publishers.

Reed, R. L. 1972. "A Monte Carlo Approach to Optimal Drilling". *Society of Petroleum Engineers Journal* 12 (05): 423–438.

Steinmann, G., and P. J. de Freitas Filho. 2013. "Using Simulation to Evaluate Call Forecasting Algorithms for Inbound Call Center". In *Proceedings of the 2013 Winter Simulation Conference*, 1132–1139. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Yang, Y., and G. I. Webb. 2009. "Discretization for Naive-Bayes Learning: Managing Discretization Bias and Variance". *Machine Learning* 74 (1): 39–74.

AUTHOR BIOGRAPHIES

MARIANA DEHON COSTA E LIMA graduated from the Universidade Federal de Alfenas in 2011 with a degree in Computer Science, completed a Masters in Computer Science at the Universidade Federal de Santa Catarina (UFSC) in 2014 and is currently doing her doctorate in Computer Science at UFSC. Her research interests are in the artificial intelligence field, having experience, primarily, on the following topics: statistical data analysis, Bayesian networks, expert systems, classification methods and regression methods. Her email address is mariana.dehon@posgrad.ufsc.br.

SILVIA MODESTO NASSAR is a professor at Universidade Federal de Santa Catarina. She graduated from the Universidade Federal do Pará in 1975 with a degree in Electrical Engineering, completed a Masters in Production Engineering at the UFSC in 1985 and his doctorate in Electrical Engineering was granted by Universidade Federal de Santa Catarina (UFSC) in 1995. She has experience in Probability and Statistics, mainly in the following topics: statistical data analysis, artificial intelligence, expert systems, Bayesian

networks, fuzzy systems, medical informatics and information technology in education. She is currently coordinator of the project e-TEC for Monitoring and Validation of Instructional Materials. Since 2006 is part of the Projeto Amanhecer / HU / UFSC offering therapeutic services to the community, through complementary integrative techniques. Her email address is silvia.nassar@ufsc.br.

PAULO JOSÉ DE FREITAS FILHO is a professor at Universidade Federal de Santa Catarina (UFSC). He graduated from the Universidade Federal do Rio Grande do Sul in 1978 with a degree in Metallurgical Engineering, completed a Masters in Production Engineering at the UFSC in 1985 and his doctorate in Systems Engineering was granted by Universidade Federal de Santa Catarina (UFSC) and the University of South Florida in 1994. His experience is centered on Computer Science with an emphasis on Systems Performance and his teaching and research is primarily in the following areas: modeling and simulation, performance assessment and systems capacity planning. Over the past five years he devoted much of his time to research on modeling, simulation and data analysis of service engineering mainly in the call center area. His email address is freitas@inf.ufsc.br.