Data analytics in asset valuation under uncertainty: A case study of unexplored oilfields

F R Pratikto1,2,*, S W Indratno2, K D Suryad2 and D Santoso2

1 Parahyangan Catholic University, Jl. Ciiumbuleuit 94, Bandung, Indonesia
2 Bandung Institute of Technology, Jl. Ganesha 10, Bandung, Indonesia

Email: *frianp@unpar.ac.id

Abstract. Valuing unexplored oilfields is challenging since it deals with multiple sources of uncertainty and covers a lengthy period. Previous research suggests that the major source of uncertainty in the valuation of unexplored oilfields is reservoir condition. It is usually represented by a bunch of parameters based on which reserve volume and production rates can be estimated. The values of these parameters are revealed only after exploration well is drilled; we refer them as post-discovery parameters. Prior to exploration drilling, data on some reservoir characteristics, called the pre-discovery parameters, are available. This research aims to develop a model to estimate the probability distribution of post-discovery parameters based on pre-discovery data. The model is data-driven, built using the database of proven reservoirs. We use the Bayesian network to develop the model and apply the k-fold cross validation to test the results. By adding a window parameter to the target variable, the model provides information regarding the trade-off between accuracy and confidence. In general, compared to the initial model that uses a priori clustering based on the reservoir’s lithology and depth, our Bayesian network model produces lower variances.

1. Introduction

This work is part of our research to develop a valuation model of unexplored oilfields. The valuation model is to serve situation where an investor is participating in an auction to place a bid on the right to operate an unexplored oilfield. In doing so, the investor needs information about the oilfield value. Valuing unexplored oilfields is very complex since is deals with multiple sources of uncertainty and covers a lengthy period. The uncertainty may be of market-related like oil prices, exchange rates, and investment climate, technical-related like exploration outcome and reservoir condition, and even social and geopolitical factors like war, political upheaval, etc.

Past research on oilfield valuation usually focus on the market and technical-related factors. Oil prices are the most common source of uncertainty and usually represented as a geometric Brownian motion (GBM) as in Pickles and Smith [1], Sunnevag [2], and Dias [3], or mean-reverting stochastic process as in Cortazar and Schwartz [4], Park et al. [5], and Haque et al. [6]. Exchange rates and investment climate are incorporated in the valuation model by Fan and Zhu [7]. Previous research by One the major sources of uncertainty in the valuation of unexplored oilfields is reservoir condition. Reservoir condition affects the oilfield value by determining reserve volume and production rates. Some research in the past represents reservoir condition by directly making assumption about reserve and/or production rates, whether it follows a certain probability distribution as in Zettl [8], Dias [3] and Park et al. [5], or stochastic process as in Smith and McCardle [9]. The only research that represents reservoir condition by its parameters and accordingly estimates the reserve volume and production rates is Hultzsch et al. [10].

Hultzsch et al. [10] uses a bunch of parameters based on which reserve volume and production rates can be estimated. The values of these parameters are revealed only after exploration well is drilled; we refer them as post-discovery parameters. In Hultzsch et al. [10], the probability distributions of post-discovery parameters are estimated from data of proven reservoirs that are considered geologically similar, where similarity is inferred by lithology and depth. In this sense, lithology and depth are the
pre-discovery parameters. Pre-discovery parameters are those whose values are known prior to exploration drilling. Based on these two parameters the proven reservoirs are a priori clustered, and the probability distribution in each cluster is estimated. Hultzsch et al. [10] assumes that all post-discovery parameters follow lognormal distribution and allows associations between them which are represented using simple correlation. We adopt the same approach in our initial model but modify the way associations between post-discovery parameters are represented. Instead of using simple correlation, we use the Gaussian copula to represent the associations between parameters.

This research aims to develop a data-driven model to estimate the probability distributions of post-discovery parameters based on pre-discovery data. This empirical approach is chosen because there has not been a solid analytical model representing the relationship between those two groups of parameters. In doing so, the proven reservoirs are posteriori clustered with respect to each post-discovery parameter. The cluster numbers are ordered such that greater number corresponds to reservoirs with greater parameter value. We use the Bayesian network to develop the model to predict the cluster of a reservoir of interest and apply the k-fold cross validation to test the results. The Bayesian network is chosen since we are interested in learning the causal relationship between the pre and post discovery parameters. We use the Tertiary Oil Recovery Information System (TORIS) database which was developed by the US government and contains data of proven reservoirs that have undergone secondary recovery.

2. Methods

Our valuation model is a multistage decision model with stochastic state variables, represented as a dynamic programming problem (see Figure 1).

\[
F_t(x_t) = \max_{u_t} \left\{ \Omega_t(x_t), \max_{\pi_t} \left[ \pi_t(x_t, u_t) + e^{-r \Delta t} E_t[F_{t+1}(x_{t+1})] \right] \right\}
\]

\[
x_t = (H_t, e_0, \Theta_t | e_0 = 1)
\]

**Figure 1.** Relationship between reservoir condition and other sub models in our valuation model.

\(F_t(x_t)\) is the oilfield value at time \(t\) represented as a function of vector of stochastic state variables, \(x_t\), and optimal choice of vector of decision variables, \(u_t\). At any time \(t\), the investor chooses between abandoning the field which results in the payoff of \(\Omega_t(x_t)\), and continuing the operation. The expression of the payoff for continuing the operation is a typical approximate dynamic programming representation which consists of the immediate profit, \(\pi_t(x_t, u_t)\), and the expected oilfield value in the next period, continuously discounted to time \(t\), \(e^{-r \Delta t} E_t[F_{t+1}(x_{t+1})]\). The stochastic state variables are oil prices (\(H\)), exploration outcome (\(e_0\)), and the vector of reservoir parameters (\(\Theta\)) given that the exploration is successful. This paper focuses on the sub model of the reservoir condition. In our valuation model, reserve volume and production rates are estimated using the compressible-liquid tank model. Table 1 lists the parameters needed as inputs to run the tank model, where the data of the first eight are available in the TORIS database and hence the Bayesian network model of those can be developed.
relationships between variables (denoted by nodes which represent the random variables (denoted by domain of random variables). The vector of post-discovery parameters will be estimated. We cluster the reservoirs based on their magnitude of the parameter of interest using the k-means clustering algorithm [11]. The clusters are labelled such that greater-numbered cluster corresponds to group of reservoirs with greater parameter value. The Bayesian network model aims to predict the cluster in which a reservoir belongs with respect to each post-discovery parameter. By doing so, we wish to come up with estimated probability distribution for each parameter with lower variance. Let \( \mathbf{r} = (r_1, r_2, r_3, r_4, r_5, r_6, r_7) \) be the vector of pre-discovery parameters where each element represents parameter in the order of that in Table 2, and let \( \Theta = (\Theta_1, \Theta_2, \Theta_3, \Theta_4, \Theta_5, \Theta_6, \Theta_7, \Theta_8) \) be the vector of post-discovery parameters where \( \Theta_q \) corresponds to the \( q \)-th parameter in Table 1.

Bayesian network (denoted by \( B \)) is a representation of joint probability distributions of \( n \) interrelated random variables (denoted by \( \mathbf{X} = (X_1, \ldots, X_n) \)) in a directed acyclic graph or DAG (denoted by \( G \)) with parameters representing the conditional probability of each variable in \( G \) (denoted by \( \Pi \)). If \( D_i \) is the domain of \( X_i \) then the joint domain of all random variables is \( D = \prod_{i=1}^{n} D_i \). The DAG \( G \) consists of nodes which represent the random variables (denoted by \( V = \{X_1, \ldots, X_n\} \)) and arcs which represent relationships between variables (denoted by \( E \)). Relationships between any two variables can be

### Table 1. Inputs to the compressible-liquid tank model.

| No | Parameters                              | Remarks          |
|----|----------------------------------------|------------------|
| 1  | Net pay                                | Estimated empirically |
| 2  | Porosity                               | Estimated empirically |
| 3  | Initial oil saturation                 | Estimated empirically |
| 4  | Formation volume factor                | Estimated empirically |
| 5  | Oil permeability                       | Estimated empirically |
| 6  | Oil viscosity                          | Estimated empirically |
| 7  | Initial pressure                       | Estimated empirically |
| 8  | Bottomhole pressure                    | Estimated empirically |
| 9  | Shape factor                           | Estimated using expert judgment |
| 10 | Skin factor                            | Estimated using expert judgment |
| 11 | Water compressibility                  | Estimated theoretically |
| 12 | Oil compressibility                    | Estimated theoretically |
| 13 | Formation compressibility              | Estimated theoretically |
| 14 | Well radius                            | Assumed |

The data of the last six parameters are not available and hence the data-driven model of those cannot be built. Instead, we use their theoretical values (for water compressibility, oil compressibility, and formation compressibility), common practice value (for well radius), and expert judgment based (for shape factor and skin factor). The pre-discovery parameters are chosen based on data availability in the TORIS. The list of the parameters and their corresponding measurement scale are presented in Table 2.

### Table 2. Pre-discovery parameters and their corresponding values.

| No | Parameters                              | Scales          |
|----|----------------------------------------|------------------|
| 1  | Lithology                              | Nominal, 4-point scale |
| 2  | Depth                                  | Ratio            |
| 3  | Depositional system and its degree of confidence | Nominal, 165-point scale |
| 4  | Digenetic overprint and its degree of confidence | Nominal, 30-point scale |
| 5  | Structural compartmentalization and its degree of confidence | Nominal, 36-point scale |
| 6  | Predominant element of heterogeneity   | Nominal, 3-point scale |
| 7  | Trap type                               | Nominal, 3-point scale |

All post-discovery parameters are stochastic in nature. The idea of predicting the values of such parameters does not make sense. Instead, our model aims to predict the cluster of a reservoir with respect to a post-discovery parameter, based on which the probability distribution of that reservoir’s parameter will be estimated.
determined using the Pearson chi-square or the likelihood ratio test.

We use the Markov blanket algorithm to develop our classification model. Markov blanket of a node representing \( X_i \) is the set of nodes comprising its parent nodes (denoted by \( R_{X_i} \)), its children nodes (denoted by \( C_{X_i} \)), and the nodes sharing a child with this node (denoted by \( M_{X_i} \)).

Conditional probability \( \mathbf{II} \) in Bayesian network \( B \) is represented by \( \{ \pi_{ijk} \}_{i \in 1...n,j \in D_{P_{X_i}},k \in D_i} \) where \( \pi_{ijk} = P_B(X_i = x_{ik} | R_{X_i} = w_{ij}) \), in which \( P_B(\cdot) \) denotes the probability in Bayesian network \( B \), \( x_{ik} \) is the \( k \)-th value of \( X_i \), and \( w_{ij} \) is the \( j \)-th configuration of \( R_{X_i} \). Then, the joint probability of all elements of \( \mathbf{X} \) in \( B \) can be calculated as follows

\[
P_B(X_1, ..., X_n) = \prod_{i=1}^{n} P_B(X_i | P_{X_i})
\]

This algorithm aims to obtain posterior joint probability of target variable by learning from the data. This requires the set of variables \( \mathbf{V} \) and the data of their corresponding values as training dataset. In this research, \( \mathbf{V} = \{ t_1, t_2, t_3, t_4, t_5, t_6, t_7, \Theta_1 \} \). There are two learning processes in this method, structural learning to determine nodes and arcs and parameter learning to estimate conditional probability in each node. The objective of the Markov blanket algorithm is to obtain a Bayesian network that can predict the target variable and fit with the dataset. In this research, the fitness is the Markov blanket algorithm of target variable and fit with the dataset. In this research, the fitness is

\[
P_B(X_i, ..., X_n) = \prod_{i=1}^{n} P_B(X_i | P_{X_i})
\]

For Bayesian network \( B \) using dataset \( T \), the BDe score can be calculated as follows:

\[
BDe(B, T) = \log(P(B)) + \sum_{i=1}^{n} \sum_{j=1}^{q_i} \left( \log \left( \frac{r(N_{ij})}{r(N_{ij}+N_{ijk})} \right) + \sum_{i=1}^{r_i} \log \left( \frac{r(N_{ij})}{r(N_{ij}+N_{ijk})} \right) \right)
\]

In the above equation, \( N'_{ij} = N' \times P(X_i = x_{ik}, R_{X_i} = w_{ij} | \mathbf{G}) \) and \( N'_{ij} = N' \times P(R_{X_i} = w_{ij} | \mathbf{G}) \) where \( N' \) is the equivalent sample size indicating the confidence of the prior probability, and \( \Gamma \) is the gamma function. We use the following algorithm to train our Bayesian network model:

Step 1 Test the significance of each arc between any two nodes in \( \mathbf{V} \) using the Pearson chi-square or the likelihood ratio. For these significant arcs, find the smallest set \( S \) such that the relationship between the nodes connected by those arcs conditioned on \( S \), are significant. If \( S \) is empty, remove the arc.

Step 2 Determine the direction of each arc.

Step 3 Identify the Markov blanket of target variable \( \Theta_1 \) (denoted by \( MB_{\Theta_1} \)).

Step 4 Determine the value of \( \{ \pi_{ijk} \}_{i \in 1...n,j \in D_{P_{X_i}},k \in D_i} \) that maximize \( BDe(MB_{\Theta_1}, T) \) using Equation 2.

Once we get the values of \( \pi_{ijk} = P_B(X_i = x_{ik} | R_{X_i} = w_{ij}) \), the conditional probability for each target variable \( \Theta_1 \) can be determined.

Since the number of records is limited, we use the k-fold cross validation to test our model. Figure 2 depicts the mechanism of the validation process. In this validation process, the data are divided equally into \( k \) folds. Subsequently, for \( i = 1, ..., k \) we take the \( i \)-th group as the validation dataset and and the remaining data as the training dataset. The training dataset are then used to train the classification model,
and the predicted values of target variable are compared to those of actual to measure the fold’s performance. The fold’s performance is measured by the proportion of record whose cluster are correctly predicted. The model performance is measured as the average performance of all folds.

![Figure 2. The k-fold cross validation process.](image)

When the classification model assigns a reservoir to cluster $i$ with respect to parameter $\Theta_q$, a nonnegative integer $w \leq 3$ is added such that when all records are assign to cluster $[\max(1,i-w) \cup ... \cup \min(i+w,k)]$ the proportion of records whose clusters are correctly predicted is at least 95%, where $k$ represents the number of clusters. For example, if $w = 1$ and $C_i = \{1, 2, 3, 4, 5\}$ denotes the initial cluster resulted from the classification model, then the final cluster would be $C_f = \{1_p, 2_p, 3_p, 4_p, 5_p\} = \{1 \cup 2, 1 \cup 2 \cup 3, 2 \cup 3 \cup 4, 3 \cup 4 \cup 5\}$. The performance of all folds and models will be evaluated by considering the final cluster after the window parameter is added.

3. Result and Discussion

The TORIS database has about 2,500 records with about 1,400 records are available for public, out of which 498 records are complete and used to train our classification model. There are two $k$’s in the methods that we use, the first one refers to the k-means clustering algorithm, and the second refers to the k-fold cross validation. We pick $k = 5$ for the k-means, while for the cross-validation we pick $k = 10$ since it is considered the best practice. The distribution of reservoirs resulted from the k-means clustering is presented in Table 3.

The target variable of the post-discovery parameter in our Bayesian network model is the cluster number. Figure 3 depicts the structure of our Bayesian network model for each target variable resulted from the training using all data in the training dataset.
Table 3. Distributions of reservoirs in each cluster.

| Post-discovery parameters | Number of reservoirs in cluster |
|---------------------------|---------------------------------|
|                           | 1     | 2     | 3     | 4     | 5     |
| Net pay ratio (\(\Theta_1\)) | 77    | 117   | 115   | 153   | 36    |
| Porosity (\(\Theta_2\))   | 126   | 161   | 99    | 110   | 2     |
| Initial oil saturation (\(\Theta_3\)) | 7     | 51    | 162   | 229   | 49    |
| Formation volume factor (\(\Theta_4\)) | 303   | 131   | 53    | 9     | 2     |
| Permeability (\(\Theta_5\)) | 340   | 100   | 42    | 14    | 2     |
| Viscosity (\(\Theta_6\))   | 480   | 11    | 2     | 2     | 3     |
| Initial pressure (\(\Theta_7\)) | 182   | 167   | 106   | 36    | 7     |
| Bottomhole pressure ratio (\(\Theta_8\)) | 51    | 71    | 98    | 130   | 148   |

Table 4 shows the result of the k-fold cross validation and the model’s performance in term of the proportion of records that are correctly classified.

Table 4. The proportion of records that are correctly classified by the model.

| Folds | Post-discovery parameters |
|-------|---------------------------|
|       | \(\Theta_1\) | \(\Theta_2\) | \(\Theta_3\) | \(\Theta_4\) | \(\Theta_5\) | \(\Theta_6\) | \(\Theta_7\) | \(\Theta_8\) |
|       | \(w = 3\) | \(w = 2\) | \(w = 2\) | \(w = 2\) | \(w = 2\) | \(w = 1\) | \(w = 1\) | \(w = 3\) |
| 1     | 0.980 | 0.900 | 0.940 | 0.960 | 0.980 | 0.980 | 0.900 | 0.980 |
| 2     | 1.000 | 1.000 | 0.960 | 0.960 | 0.900 | 0.980 | 0.960 | 0.940 |
| 3     | 0.980 | 1.000 | 0.900 | 1.000 | 0.960 | 0.980 | 0.980 | 1.000 |
| 4     | 0.980 | 0.980 | 0.860 | 0.940 | 1.000 | 1.000 | 0.900 | 0.860 |
| 5     | 0.980 | 0.980 | 0.900 | 0.960 | 0.960 | 0.980 | 0.920 | 0.820 |
| 6     | 0.940 | 0.980 | 0.840 | 0.960 | 1.000 | 0.980 | 0.900 | 0.920 |
| 7     | 0.940 | 0.960 | 0.980 | 0.980 | 1.000 | 0.960 | 0.960 | 0.900 |
| 8     | 0.960 | 0.960 | 0.900 | 0.960 | 0.980 | 0.980 | 0.980 | 0.900 |
| 9     | 0.939 | 1.000 | 0.816 | 0.980 | 0.980 | 0.980 | 0.939 | 1.000 |
| 10    | 0.959 | 1.000 | 0.959 | 0.959 | 1.000 | 0.959 | 0.959 | 1.000 |
| Folds average | 0.966 | 0.976 | 0.906 | 0.966 | 0.976 | 0.978 | 0.940 | 0.932 |
| All data | 0.996 | 0.994 | 0.998 | 1.000 | 0.990 | 0.996 | 1.000 | 0.988 |

The window parameters are set such that the proportion of correct classification of both the model and the cross-validation is at least 0.90. It is interesting to see if the classification model produces lower variance compared to our initial model which uses a priori clustering based on lithology and depth. Table 5 shows the comparison between the initial and the Bayesian network model in term of the total weighted variance implied by the reservoir classification suggested by the model.
Figure 3. The structure of our Bayesian network models.
Table 5. Total weighted variance of the initial and the Bayesian network model.

| Models          | \( \Theta_1 \) | \( \Theta_2 \) | \( \Theta_3 \) | \( \Theta_4 \) | \( \Theta_5 \) | \( \Theta_6 \) | \( \Theta_7 \) | \( \Theta_8 \) |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
|                 | \( w = 3 \)    | \( w = 2 \)    | \( w = 2 \)    | \( w = 2 \)    | \( w = 2 \)    | \( w = 1 \)    | \( w = 1 \)    | \( w = 3 \)    |
| Initial         | 0.0637         | 0.0091         | 0.0226         | 343,516        | 124,436        | 383,338        | 0.0717         |
| Bayesian network| 0.0456         | 0.0033         | 0.0212         | 205,220        | 2,990          | 1,088,988      | 0.0443         |
| Variance change | -28.4%         | -35.3%         | -41.8%         | -6.2%          | -40.3%         | -97.6%         | 184.1%         | -38.2%         |

It can be observed from Table 5 that the Bayesian network model produces classification with lower variance except for parameter \( \Theta_5 \), initial pressure. The result of the k-fold cross validation provides information about the confidence of our Bayesian network model, while the window parameter gives the accuracy of the post-discovery parameter estimates. Greater value of \( w \) refers to greater overlap between two consecutive final cluster, which corresponds to greater variance. Our initial model which is based on Hultsch et al. [10] does not provide this information. In this sense, our Bayesian network model is considered better.

4. Conclusion

We develop a Bayesian network classification model to determine the cluster of a certain reservoir with respect to a post-discovery parameter, based on which the probability distribution of the parameter will be estimated and used in the stochastic valuation model. The classification model aims to specify the cluster based on the value of pre-discovery parameters.

Our Bayesian network classification model is better than our initial one which uses a priori clustering based on lithology and depth. It is not only because the Bayesian network model produces lower variance on almost all parameters, but more than that, it provides information regarding accuracy and confidence. This information would be useful when the model is implemented and the data evolves overtime, and the models are retrained, accordingly.

Acknowledgments

FRP thanks Parahyangan Catholic University for their support in conducting this research.

References

[1] Pickles E, Smith JL. Petroleum Property Valuation: A Binomial Lattice Implementation of Option Pricing Theory. Energy J., 1993 Apr 1;14(2). doi.org/10.5547/issn0195-6574-ej-vol14-no2-1
[2] Sunnevåg K. An option pricing approach to exploration licensing strategy. Resour. Policy. 1998 Mar;24(1):25–38. doi.org/10.1016/s0301-4207(98)00005-1
[3] Dias MAG. Valuation of exploration and production assets: an overview of real options models. J. Pet. Sci. Eng., 2004 Oct;44(1-2):93–114. doi.org/10.1016/j.petrol.2004.02.008
[4] Cortazar G, Schwartz ES. Implementing a Real Option Model for Valuing an Undeveloped Oil Field. Int Trans Oper Res. 1997 Mar;4(2):125–37. doi.org/10.1111/j.1475-3995.1997.tb00069.x
[5] Park C, Kang JM, Min B. Compound Real Options Incorporated With a Stochastic Approach for Evaluating an Uncertainty in Petroleum Exploration. Energ Sources Part B. 2013 Jul;8(3):252–62. doi.org/10.1080/15567240903117617
[6] Haque MA, Topal E, Lilford E. Evaluation of a mining project under the joint effect of commodity price and exchange rate uncertainties using real options valuation. Eng Econ. 2016 Jul 28;62(3):231–53. doi.org/10.1080/0013791X.2016.1217366
[7] Fan Y, Zhu L. A real options based model and its application to China's overseas oil. Energy Economics. 2010;32:627-637. doi.org/10.1016/j.eneco.2009.08.021
[8] Zettl M. Valuing exploration and production projects by means of option pricing theory. *Int J Prod Econ*. 2002 Jul;78(1):109–16. doi.org/10.1016/s0925-5273(01)00215-8

[9] Smith JE, McCardle KF. Valuing Oil Properties: Integrating Option Pricing and Decision Analysis Approaches. *Oper Res*. 1998 Apr;46(2):198–217. doi.org/10.1287/opre.46.2.198

[10] Hultzsch P, Lake LW, Gilbert RB. Decision And Risk Analysis Through The Life Of The Field. *Hydrocarbon Economics and Evaluation Symposium*. 2007. doi.org/10.2118/107704-ms

[11] Strauss DJ, Hartigan JA. Clustering Algorithms. *Biometrics*. 1975 Sep;31(3):793. doi.org/10.2307/2529577

[12] Heckerman D, Geiger D, Chickering DM. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Uncertainty Proceedings 1994*. 1994;293–301. doi.org/10.1016/b978-1-55860-332-5.50042-0