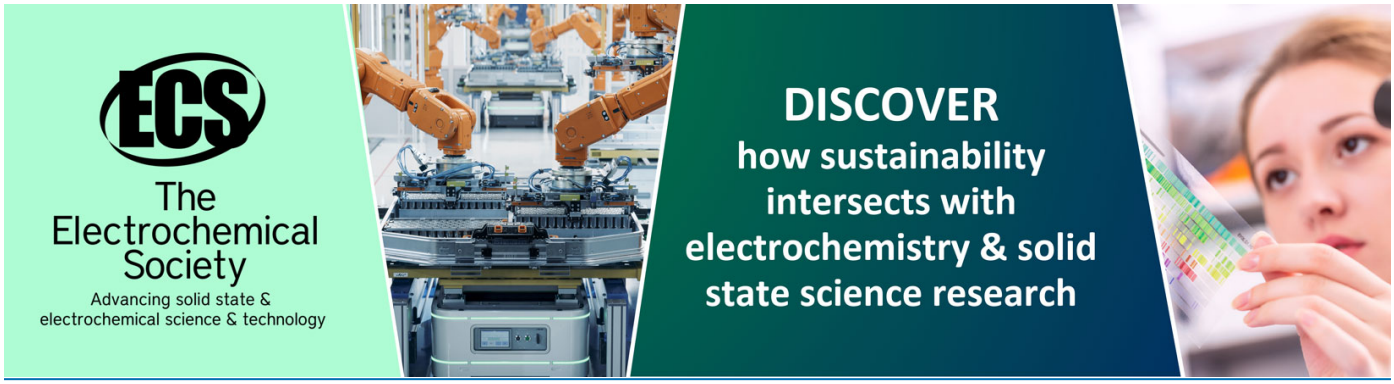
Preliminary sandstone reservoir depth prediction with pre-processing data using principle component analysis (PCA) and partial least square (PLS) based on well logging data attribute



Preliminary sandstone reservoir depth prediction with pre-processing data using principle component analysis (PCA) and partial least square (PLS) based on well logging data attribute

E Utami [1], M H Purnomo [2], R F Rizki[3] and T R Biyanto[3]

[1]Politeknik Energi dan Mineral Akamigas, Ministry of Energy and Mineral Resources, Cepu 58315, Indonesia
[2]Dept of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3]Dept of Physics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

*Email: ernautami@esdm.go.id

Abstract. Sandstones containing 60% of oil or gas because of porosity and permeability. The conventional method required sandstones predicted with complete data. This study used neural networks to determine the depth of sandstones and predict incomplete variables well site at the Sunda Strait-south area. Data pre-processing used PCA-PLS to find the most important variables that affect the output thus improving the prediction results. Multicollinearity analysis is used to determine the data compression needed. Raw data multicollinearity result showed that the multicollinearity occurs indicated with VIF over 10 and tolerance under 0.1 at CALI and SP variables. The PCA-PLS analysis uses to reduce data from 13 variables into six important variables namely DEPT GR RHOB NPHI ILD P_{eff}, these results do not experience multicollinearity. The variable that predicted is ILD with the best ANN multilayer perceptron showed small standard deviation and standard error results 2.85 and 0.03. The best ANN model to predict the depth of radial basis sandstone is due to produce a regression test of 0.8 based on the results of the validation of the log image.

1. Introduction

In general, hydrocarbon fluids (oil or gas) contain at reservoir rocks which have porosity and good permeability, including sandstone with 80% of reservoir in the world discovered in these rocks, and fluid contained in it almost 60%. The existence of sandstone (sandstone) is an indicator of good reserves of hydrocarbon fluid [1]. Previous studies have used neural networks to predict problem of rock layers (lithology), the results show the correlation coefficient average is 70%. The correlation coefficient indicates proximity of actual output data and prediction data. This study targeted to improve the correlation above 70% by data optimizing with pre-process data before it predicts by artificial neural network (ANN). Data pre-process include variables selection that it has influence each other between input variables (X) using PCA (principal component analysis), or the effect of variable X to variable output (Y) using the PLS (partial least squares) [7]. Besides pre-processing data in this

study also compared the architecture of neural networks (NN) to obtain optimal result are radial basis function (RBF) and type multilayer perceptron (MLP). Both compared the missing data prediction and classification to determine the position of a layer of sandstone.

2. Related Work

2.1. Well Data Log

The log is a graph of measured parameters in wells with depth or time. This study uses a digital log, it is a log that was converted from graph into a numerical value. Digital well log data used in this study are 3 wells located in Sunda Strait South area with the same earth formation layer so that the data be trained to become better and robust when tested for prediction.

Table 1. Well Data

Well Name	Depth (ft)	Initial Variable
FARIDA A-11_MAXUSU (Well Test)	7550-9999	DEPT(ft) , GR (API), SP (mv), CALI (inch), SG (ohmm), MED (ohmm), MSFL (ohmm), DEEP (ohmm), DT (us/f), RHOB (g/cc), NPHI (%), DPHI (%), PEF (B/E), DRHO (g/cc)
FARIDA B-3_PIACO (Well Train)	4750 – 8590	
ZELDA E-2_IIA (Well Train)	1459 – 8580	

Table 2. Rock density in formation

Rock	Real density (gr/cc)	Density when logging (gr/cc)
Sandstone	2,650	2,684
Limestone	2,710	2,710
Dolomite	2,870	2,876

2.2. Zoning Process

This step calculates effective porosity based on the log Gamma Ray (GR) value as representative of the permeability properties of depth. Value of log GR which has the same average trend at certain depth be divided to get an accurate result called the zoning process. Gamma-ray measurements at well can generate vary range of values due to differences in the condition of the drill holes and the tools of the respective service company and so we need a wells reference [4]. This study used well - 3_PIACO FARIDA B because has complete data and represents other wells.

2.3. Calculate Shale Volume

Shale volume calculated from GR value. The principle of the gamma ray log is natural earth radioactivity recorded, in which gamma rays are able to penetrate rock and detected by a sensor causing decrease in wave intensity with API unit. Sandstone has low radioactive absorbance. Along with the increase in shale content in the rock, the radioactive material content increases and GR value will increase [4]. With creating a boundary line between the shale base line with a sand base line. Gamma rays can be used as a determinant of permeability properties of layers and also determine the volume of shale by the following equation:

$$V_{sh} = \frac{GR_{\log} - GR_{\min}}{GR_{\max} - GR_{\min}} \quad (1)$$

where:
GR_{log} = GR readings at each depth interval (API)
GR_{min} = GR readings non shale layer (API)
GR_{max} = GR readings shale layer (API)
V_{sh} = Shale volume

2.4. Density Log (RHOB)

Gamma rays emitted from radioactive source by collisions with electrons in the soil formation causing energy loses and equivalent with the mass density of formation in a borehole. The measured density is the overall density of the rock matrix and pore fluid contained in the formation. Rock porosity is also influenced by the fluid content of rock lithology [5]. Density log porosity denoted by with the following equation:

$$\Phi_D = \left(\frac{\rho_{ma} - \rho_b}{\rho_{ma} - \rho_f} \right) - \left(V_{sh} \times \Phi_{Dsh} \right) \quad (2)$$

Where:

V_{sh} = volume shale (from equation 1)
Φ_{Dsh} = Porisity value from density log in shale layer
ρ_{ma} = rock density, gr/cc
ρ_b = bulk density in density log , gr/cc
ρ_f = fluid density (water), gr/cc

2.5. Neutron Log (NPHI)

Neutron log is used for the calculation of rock porosity, lithology evaluations, and the detection of the presence of gas. The principle is measuring rock porous percentage based on hydrogen atoms in it, which it is assumed that the hydrogen will be hydrocarbons or water.

$$\Phi_N = \left(\left(1.02 \times \Phi_{NLog} \right) + 0.0425 \right) - \left(V_{sh} \times \Phi_{Nsh} \right) \quad (3)$$

Where:

Φ_{Nlog} = curve neutron log reads
0.0425 = correction value for limestone
V_{sh} = volume shale (from GR log)
Φ_{Nsh} = shale porosity neutron

2.6. Induction Log Deep (ILD)

The principle of ILD is alternating current with a high frequency (± 20,000 cps) with constant intensity sent through coil sender that produces an electromagnetic field which will produce currents induced the formation. ILD is intended to determine where the water saturation values were prospectively layer is in the range of 0 - 0.5 [5].

2.7. Effective Porosity (Peff)

Porosity value in oil reservoir indicates available space to be occupied by a liquid or gas. Porosity can be defined as the ratio between total volumes of rock pores to a total volume of rock whole volume. Rock porosity can be classified:

According to the following equation:

$$\phi = \frac{\phi_N + \phi_D}{2} \, x \, 100\,\% \quad (4)$$

Where:

ϕ_N = Neutron Log Porosity
ϕ_D = Density Log Porosity

3. Multicollinearity

Multicollinearity is a condition of a linear relationship or a high correlation between each independent variable in the regression model. Multicollinearity occurs when input variables related to each other in the model. Therefore, multicollinearity problem does not occur in the simple linear regression involving only one independent variable. Multicollinearity indicated by VIF and tolerance values. VIF

determines data distribution that the variables influence other variables while tolerance value is the value of the magnitude of the error that occurred. Multicollinearity the VIF value is> 10 and tolerance <0.1, following equation [7]:

$$r^2 = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \quad (5)$$

$$VIF_n = \frac{1}{1-r^2} \quad (6)$$

$$tolerance = 1/VIF \quad (7)$$

4. Pre-processing Data

PCA method can reduce the dimensionality of the dataset without important information dismiss from the dataset. The number of components is equal to the number of original variables so that no information is lost in the process [13]. New variables that formed as a result of PCA called principle component values and the formation of the variable is referred to as the principal component score. All new variable p indicates the maximum variance is not calculated on the p-1 variables before, and the whole new variable p is not correlated with each other [8].

Partial least squares (PLS) is one of the methods that can be used to overcome the problem of multicollinearity [10]. The method function is connecting two matrix data x and y, with linear model multivariate, but beyond regression traditional, because the model also has structure x and y. PLS can analyse a very large amount of data, noise, collinear even variables incomplete in both x and y [9].

5. Results and discussions

In large amounts of data, the correlation between inputs is undesirable because multicollinearity resulting in multiple regression. Results Table 3 found that VIF in CALI variable and SP exceeds 10 which is equal to 26.95 and 19.95 as well as collinearity of less than 0.1, namely 0037 and 12:05.

Table 3. Raw variable multicollinearity analysis data result

Model	Unit	Collinearity Statitics	
		Tolerance	VIF
DEPT	Ft	0.879	1.138
CALI	Inch	0.037	26.95
SP	Mv	0.05	19.955
ILD	Ohmm	0.423	2.362
SFLU	Ohmm	0.516	1.939
MSFL	Ohmm	0.769	1.3
DT	Us/f	0.115	8.664
RHOB	g/cc	0.425	2.354
NPHI	v/v	0.366	2.729
DRHO	g/cc	0.699	1.431
PEF	B/E	0.948	1.054
Dependent Variable: GR			

Table 4. PCA result

Zone 1	MSFL - ILD
Zone 2	ILD-DEPT-Peff-NPHI
Zone 3	NPHI-DT-Peff-ILD-MSFL
Zone 5	DT – NPHI - Peff

The PCA analysis establishes important variables in the proximity of the well data by observing major component values (p) were visualized with coordinates. The location of the proximity of the coordinates between variables in table 4 as much as 3 of the 4-zoning observed there are three variables that are not multicollinear including Peff, NPHI, ILD. MSFL variable has no value or empty

some 4665 data from 5940 MSFL data is processed so that 78.5% MSFL data is not there and it is very less to be taken as training data neural network. In this study the desired output is data that correlates permeability because it will be used as inputs to the ANN. The VIP value of more than 1 indicates that an important variable x, the value of less than 0.5 indicates a variable x is not critical and values between 0.5 - 1 is a variable in the gray zone, which means the effect of variable x to variable y depends on the dataset used. Table 5 shows for the output i.e. GR highly correlated variable is RHOB-ILD-Peff-DEPT.

Peff or effective porosity calculated based GR, RHOB and NPHI, it can be concluded that for input for ANN. DEPT variable as an output for the purpose of this study is to predict the depth positions, therefore need to be included as a training DEPT. Overall input ANN used is ILD, RHOB, NPHI, GR, Peff, and DEPT.

Table 5. PLS result

Variable	Unit	Total Correlation ≥1
RHOB	g/cc	4
DT	us/f	1
ILD	ohmm	3
NPHI	v/v	1
MSFL	ohmm	1
CALI	inch	1
DEPT	ft	3
SP	mv	2
DRHO	g/cc	1
Peff	%	4

Table 6. Multicollinearity result with variable after PCA-PLS

Model	Unit	Collinearity Statistics	
		Tolerance	VIF
DEPT	ft	0.928	1.078
ILD	ohmm	0.858	1.165
RHOB	g/cc	0.568	1.761
NPHI	v/v	0.629	1.591
Peff	%	0.948	1.07
a. Dependent Variable: GR			

Table 7. Data training neural network result

	Multilayer perceptron	Radial Basis
Regresi	0.99702	0.9993
MSE	0.000009	0
Epoch	947	10000

5.1. Multicollinearity Result and Analysis with Variable after PCA-PLS

This analysis aims to determine the data still experiencing multicollinarization or not after treatment of pre-processing data using PCA-PLS, here their iterations when VIF and tolerance still not reaching the targets. Table 6 the results obtained that the VIF in all the variables of less than 10 as well as all the variables collinearity tolerance exceeds 0.1. Therefore, this data is not experiencing the condition multicollinarization so no need for pre-treatment process the returned data and these results can be used as inputs to the neural network shown in table 6.

5.2. Recovery Data Lost Analysis and Result.
In FARIDA test wells A-11_MAXUSU, ILD are important variables that do not exist, while the other two wells there. So, it can be predicted using neural networks. In this research, use neural network architecture are multilayer perceptron and radial basis.

5.2.1. Data Training Result
At the training was good or not based on the value of the correlation or regression between outputs and targets as well as the mean squared error (MSE) is generated, shown in table 7.On the results on table 7 very good training for the regression close to unity means that the predicted value has a small MSE value against the target value. MSE is the squared prediction error value the smaller the MSE then the model is the best. In this training process JST radial basis the best for the regression error close to zero and that comes closest to the value of 1.

5.2.2. Data Test Result
In this study used 3 wells adjacent and in the same rock formations. The standard error is the standard deviation of error of itself and as an indicator that the data is representative shown in table 8. In training better model of radial basis but when tested the model produces a value that is bad then there has been a overfitting [12]. In principle, radial basis vector measuring the distance between the input and output to produce a predictive value, so the architecture is not suitable for predicting the data in this case.

Table 8. Data training neural network result

	Multilayer perceptron	Radial Basis
Deviation standard	2.85	664538.80
Standard Error	0.03	6680
Range value of ILD prediction (ohmm)	0.75 – 55.58	(-3055847) - 34397088
Range value of ILD train (ohmm)	0.32 – 59.47	

6. Conclusions
Data get multicollinearity because the two variables have a value in the variable VIF CALI and SP exceeds 10 that is equal to 26.95 and 19.95 as well as collinearity less than 0.1 so that the necessary analysis of PCA-PLS. PCA-PLS result for neural network variable input are ILD, RHOB, NPHI, GR, Peff, DEPT. These results reduce the dimension variables from 13 to 6 variables. Multilayer perceptron architecture is the best neural network to predict the missing data in data set and radial basis architecture is the best neural network for predicts the position of sandstone in this case.

Acknowledgment
The writer would like to thank to Politeknik Energi dan Mineral Akamigas where the author works. The writer would also like to thank to Ministry of Energy and Mineral Resources providing the opportunity to write on the data provided to the author as research material

References
[1] Bora, “Pradyut, Formation Evaluation Based On Well logging Data”, Canada : Petrofed, 2011.
[2] Crain, Ross, Crain’s Petrophysical Handbook. Canada : Rocky Mountain House, 2010.
[3] H. Mohseni, M. Esfandyari, and E. Habibi Asl, “Application of artificial neural networks for prediction of Sarvak Formation lithofacies based on well log data, Marun oil field, SW Iran.,” Geopersia, vol. 2, pp. 111–123, 2015.
[4] Rolon, Luisa. “Using artificial neural networks to generate syhintetic log”. USA : West Virginia University, 2009

[5] Pinar, Yilmaz . Gary, Isaksen. ST55 - Oil and Gas of the Greater Caspian Area. USA : AAPG Geology, 2007.
[6] Asquith, G., and D. Krygowski, Basic Well Log Analysis: AAPG Methods in Exploration 16, p.31-35, 2014.
[7] Silvey, Multicollinearity and Imprecise estimation. UK : University of glasgow, 2005.
[8] Shlen, Jonathon, "A Tutorial on Principal component analysis". New york : Cornell University, 2014.
[9] Zayuman, Hidayat, "Pengenalan Wajah Manusia Menggunakan Analisis Komponen Utama (PCA) Dan Jaringan Syaraf Tiruan Perambatan-Balik". Semarang : Universitas Diponegoro, 2011.
[10] Monecke, A. & Leisch, F. SEM PLS: Structural Equation Modeling Using Partial Least Square. Journal of Statistic Software, 2012.
[11] Grag. Comparison of regression analysis, Artificial Neural Network and genetic programming in Handling the multicollinearity problem. Singapore : Nanyang technological University, 2012.
[12] Biyanto, Totok. 2009. Overfitting pada neural network. https://totokbiyanto.wordpress.com/2009/08/05/over-fitting-pada-neural-network/. Diakses tanggal 14 Mei 2016.
[13] Utami, E., "Reservoir Zone Prediction Using Logging Data - Multi Well Based On Levenberg-Marquardt Method", IEEE, 2017.