Literature review

For decades, petroleum engineers and researchers are looking for a simple and reliable way to predict oil production of petroleum wells. The production prediction model can help and forecast in numerical and physical ways. Technic engineers and researchers exploration mainly divided into three parts: 1. Petroleum production prediction which is the traditional method which concludes five subcategories. 2. Curve estimations, and 3. Neural networks.

For the petroleum production prediction in an oil well, the traditional methodologies include: (1) by analogy, (2) volumetric, (3) material balance, (4) decline curve fitting, and (5) reservoir simulation. Each method could be used for prediction but with different data requirements. For example, “by analogy” performs the prediction of the target well based on similar wells. This method is efficient, inexpensive, and good for estimation before drilling, but lack of accuracy. “Material balance” determines original oil-in-place which based on the law of conservation of mass. Moreover, each of those methods has limitations but can be used to cross-validation the prediction results of the prediction results generated by other methods.

Curve estimation is a decline curve analysis technique based on exponential, hyperbolic, and harmonic equations. [1] proves that fitting production data and predict the results with a decline curve is an insufficient and unreliable way if the historical production data is unreliable and missing. Several applications of fluid flow mechanism and petroleum production prediction using curve analysis are proposed such as [2] and [3].

The recent methods are to estimate production values using Artificial neural network. [4] proves that the Neural Network gave lower errors such as root mean square error (RMSE), and author also believes that the data pre-processing to be the most important steps in applying the ANN approach to geological problems. Data preprocessing dealing with missing values and nan values discussed in [5], [6], [7]. Moreover, [8] Indicates that the Neural network model shows higher accuracy when compared to other correlation methods. ANN models trained with more advanced, non-linear, deep & wide NN structures than the polynomial fitting equations implemented in the curve estimations methods. Instead of solving a bunch of mathematical equations to obtain the best coefficients, the neural network model updates weights to reduce the error at each step in each training epoch with objective functions and the back-propagation algorithms.

For our project, we should start with the physics meaning of features for the first step of data cleaning and conditioning. For example, to deal with the zeros in the ‘AVG\_DOWNHOLE\_PRESSURE’ and ‘AVG\_DOWNHOLE\_TEMPERATURE’, physics meaning plays an important role to verify if the zeros are the type II outliers or not. Moreover, to predict the missing data and zeros in features, the prediction based on similar wells. Above two cases are both based on the traditional methods for petroleum production prediction in an oil well. Instead of using the curve estimation, we are more focus on the artificial neural network. We have spent much time on the data preprocessing to prepare the datasets, due to the significance and potential to improve the performance of the prediction mentioned as [4]. During the preprocessing process, we have implemented a mean value (traditional method), multilayer perceptron, and super vector regression to predict the missing value and type II outliers.

[1] “El-Banbi, A.H., Wattenbarger, R.A., 1996. Analysis of commingled tight gas reservoirs. SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA, 6–9 October 1996 (SPE 36736).”

[2] “John, E.G., 1998. Simplified curve fitting using spreadsheet add-ins. International Journal of Engineering Education 14 (5), 375–380.”

[3] “Li, K., Horne, R.N., 2003. A decline curve analysis model based on Fluid flow mechanisms, SPE western regional/AAPG Pacific section joint meeting held in long beach, California, USA, 19–24 May 2003 (SPE 83470)”

[4] “Wong, P.M., Taggart, I.J., 1995. Use of neural network methods to predict porosity and permeability of a petroleum reservoir. AI Appl. 9 (2), 27–37.”

[5] Gelman, A., & Hill, J. (2006). Missing-data imputation. In Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research, pp. 529-544). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511790942.031

[6] Jason Brownlee, How to Handle Missing Data with Python <https://machinelearningmastery.com/handle-missing-data-python/>

[7] Alvira Swalin, How to Handle Missing Data <https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>

[8] “Gharbi, R.B., Elsharkawy, A.M., Karkoub, M., 1999. Universal neural-network-based model for estimating the pressure–volume–temperature (PVT) properties of crude oil systems. Energy & Fuels 13, 454–458.”

Multi-layer perceptron (MLP)

Another approach to impute the missing data and zeros in ‘AVG\_ANNULUS\_PRESS’ feature is the multi-layer perceptron (MLP). For the dataset, we used to train the model is the data from all the production including well 11H, well 12H, well 15D and well 14H. The reason we did not use the same data of wells to train and predict missing data is that first, the mainly missing data of ‘AVG\_ANNULUS\_PRESS’ exists in the well 1C whose missing ratio is 97%, that means no useful dataset for training. The second reason is that there is no strong relationships and consistent characteristics between each well. So, to use one of the wells to predict the well 1C is not reliable. Thus, we trained the model based on all production data which may conclude all the relationships and the features. The steps of data preprocessing are the following: 1. Drop Nans if the missing values are less than 5% of the whole dataset of each feature. 2. Drop outliers using Z-scores with threshold = 3 in all nine input features and the output ‘BORE\_OIL\_VOL.’ 3. Construct the input of the NN with ten input features including ‘BORE\_OIL\_VOL’ and the output is ‘AVG\_ANNULUS\_PRESS.’ We split the whole dataset as training and test sets, which consist of 75% and 25% of the whole dataset. Moreover, for the training set, we split 20% as the validation set to do the cross-validation. After tuning the parameter and the structure of the MLP, the final structure of NN is which has three hidden layers, and each layer has 20 neural as **(Figure. NN structure.png caption: Neural network structure).** The training and validation results are shown in (Figure. Modelloss.png), and the mean square error (MSE) is 6.93 and r squared is 70.87%. Then, we implemented it with all the missing data, and zeros data of ‘AVG\_ANNULUS\_PRESS’ in the whole dataset and impute the data.

To analyze the data of each well, the box plot of each well show in the following. The boxplot of the well 1C show in **(Figure. well1box.png, captions: Box plot of well 1C, (A) the missing data ratio of well 1C, (B) the boxplot of well 1C with raw data, (C) the boxplot of well 1C without zeros (D) the boxplot after imputed data.)**. As shown in **Fig. ?** (A), the missing ratio of ‘AVG\_ANNULUS\_PRESS’ is 97.72%, so there is no box plot for ‘AVG\_ANNULUS\_PRESS’ in (B) and (C), and the imputed data box plot of ‘AVG\_ANNULUS\_PRESS’ is shown in (D), which has the same range as other wells which have the real values of ‘AVG\_ANNULUS\_PRESS’. The zeros are mainly existing in well 12H shown in **(Figure. well3box.png, captions: Box plot of well 12H, (A) the ‘AVG\_DOWNHOLE\_PRESSURE’ scatter plot of well 12H, (B) the boxplot with raw data, (C) the boxplot without zeros (D) the boxplot after imputed data.).** As shown in **Fig.? (A)**, there are around 2000 data points when ‘AVG\_DOWNHOLE\_PRESSURE’ is equal to zero, and only less than 1000 non-zero points. Moreover, the non-zero data points falling in the concentration range from 200 to 350, so we decide to use mean value to replace the zeros because there is no physical meaning and no reason for it equals zeros. The reason for these zeros may be due to the mistake during the data recording. After replacing the zeros with mean values, from the **FIG.?** (B) and (C) we can see that the box plot of ‘AVG\_DOWNHOLE\_PRESSURE’ is mainly equal to the mean value. From all the box plot of each well, we can see the significant change of each feature before and after the data preprocessing, especially in well 1C and well 12H. The pair plot of well 1C and 12H are shown in **(Fig.well1pairplot)**, and **(Fig.well3pairplot)**, and the correlation plots are shown as **(Fig.well1cor)**,and **(Fig.well3cor)**.