Prediction of Well Production Rate using Time Series Analysis

Data Exploration, Strategy and Results

Maryam Bagheri
Li Huang
Manyang Sun
Haoran Zhao
Mentor: Srinath Madasu

June 5, 2019

Contents

1 Problem Statement

Equinor recently disclosed subsurface and production data for the oil field *'Volve'* shown in Figure 1, which is located in the southern area of the Norwegian North Sea. Starting in February 2008, Volve had been operated for approximately 8 years till its shutdown in 2016, during which a total of 63 million bbl of oil were produced.

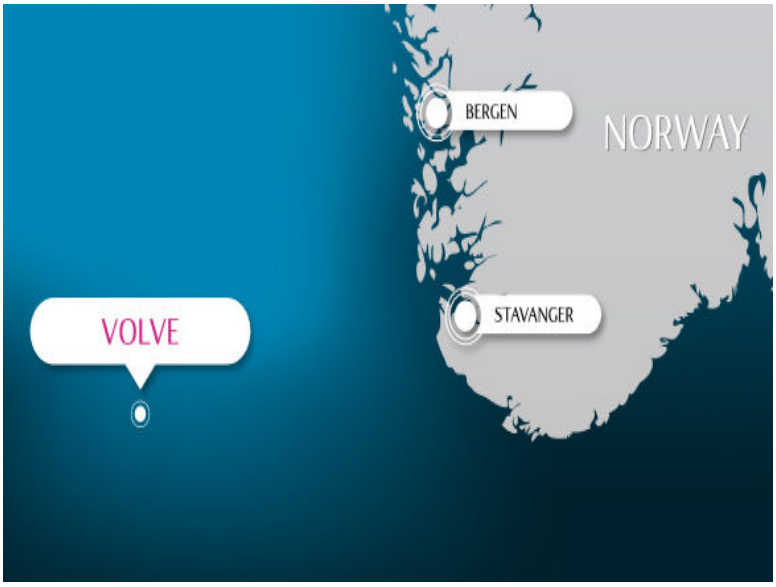


Figure 1: Volve's location near Norway

Given these historical record of subsurface and production data, accurate algorithms/statistical models contribute to extending the field life are helpful for optimal budget and resource allocation, efficient operations, and production maximization. According to Equinor, "this is the most comprehensive and complete data set ever gathered on the NCS". Such time-series data prediction is challenging because the target (bore oil production) depends on large-scale and high-dimensional datasets with unknown distributions, a large amount of missing data and outliers.

The goal of this project is to perform time-series forecasting based on large-scale and high-dimensional datasets. Specifically, given the production with 12 related variables such as dates, hours, temperature, pressure, etc., the task is to explore the complex nonlinearity of features and make a prediction on bore oil production with machine learning and deep learning modeling techniques. Once a well-trained model has been developed, it becomes a powerful tool to support future oil drilling system design and optimization. For example, in the early-stage research and development, an optimal design can be easily and fast obtained by testing different combinations of well properties in simulation, such as temperature, pressure, choke size, water injection, etc. The data ranges from 2008 to 2016, or eight years, with more than 100000 observations from six production wells and two water injection wells. Basically, we want to use the values of water injection volume given in water injection wells dataset as a feature to build our final predictive model for oil production.

Table 1: Related features

Variables	Definition
DATEPRD (string)	Date of production
ON_STREAM_HRS (int)	Production hours
AVG_DOWNHOLE_PRESSURE (int)	Average downhole pressure
AVG_DOWNHOLE_TEMPERATURE (int)	Average downhole temperature
AVG_DP_TUBING (int)	Average differential pressure tubing
AVG_ANNULUS_PRESS (int)	Average annulus pressure
AVG_CHOKE_SIZE (int)	Average choke size
AVG_WHP_P (int)	Average well head pressure percentage
AVG_WHT_P (int)	Average well head temperature
DP_CHOKE_SIZE (int)	Differential pressure at choke
FLOW_KIND (string)	Production/injection
BORE_WI_VOL (int)	Bore water injection volume

The dataset provides 23 types of information, covering more than 15000 operation days. 12 of them are directly related to our target. Our objective is to make a prediction to the bore oil volume based on the features in the Volve datasets shown in Table 1.

The research details include data prediction strategy, data cleaning, data conditioning, exploratory analysis, feature extraction, machine learning algorithms experimentation, prediction metrics, and providing deliverable results (R Squared, Root Mean Square Error, Mean Absolute Error).

2 Literature Review

For decades, petroleum engineers and researchers are looking for a reliable and straightforward 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 is 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.

To estimate the petroleum production in an oil well, the traditional methodologies include: (1) by analogy, (2) volumetric, (3) material balance, (4) decline curve fitting, and (5) reservoir simulation, Thomas, R.S. *et. al*[1]. 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 useful for estimation before drilling, but lacks accuracy. "Material balance" determines original oil-in-place which base on the law of conservation of mass. Each of those methods has limitations but can be used to cross-validate the prediction results generated by other methods.

Curve estimation is a decline curve analysis technique based on exponential, hyperbolic, and harmonic equations. El-Banbi, A.H.*et al.*[2] proves that to fit production data and predicting the results with a decline curve is an insufficient and unreliable method if the historical production data is inaccurate and missing. John, E.G.[3] and Li, K.*et al.*[4] propose several applications of fluid flow mechanism and petroleum production prediction using curve analysis.

The most recent method is to estimate production values using the artificial neural network (ANN). Wong, P.M.*et al.*[5] proves that the Neural Network gave lower errors such as root mean square error (RMSE), and the author also believes that the data pre-processing is the most important step in applying the ANN approach to geological problems. Gelman, A.*et al.*[6], Brownlee, J.[7] and Swalin, A.[8] discuss the data pre-processing of missing values and nan values. Moreover, Gharbi, R.B.,*et al.*[9] indicates that the Neural network model shows higher accuracy when compared to other correlation methods. ANN models are 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.

In this project, the first step of data cleaning and conditioning starts with the real meaning of features. For example, to deal with the NA's and zeros, the physical meaning plays an important role to verify if the zeros are the real data or just NA's; this is more innovative than using missing data imputation methods without considering the real meaning of features. Instead of using the curve estimation, we will focus on the artificial neural network due to the significance and potential to improve the performance of the prediction mentioned in [5].

3 Data Cleansing

Data cleansing is the process of detecting, removing and/or correcting corrupt or inaccurate records from a data set, which plays an important role in getting the data sets ready for analytics processes. It is worth mentioning that most data scientists spend almost 60% of their time and effort on data cleansing. As for the data cleansing process, we first inspect the missing data, then, we remove the outliers, and finally, we impute the missing data.

3.1 Missing data ratios

After some visual inspection on the data set, we found that when the oil production rate is zero or less than 10, it means the well is closed. In fact, wells are sometimes closed in for a period of time to allow stabilization prior to beginning a drawdown-buildup test sequence. Figure 2 shows the on stream hours versus time for production wells with oil production rates less than 10. It can be seen from the figure that for oil production rates less than 10, the on stream hours are mostly 0. This proves the fact that the wells were closed for that time period. Therefore, for the data cleansing process, we keep the data with oil production rate greater than 10, which reduces the size of production wells data set by 14.9%. Table 2 shows the size of the data set for each well after removing the oil production values less than or equal to 10.

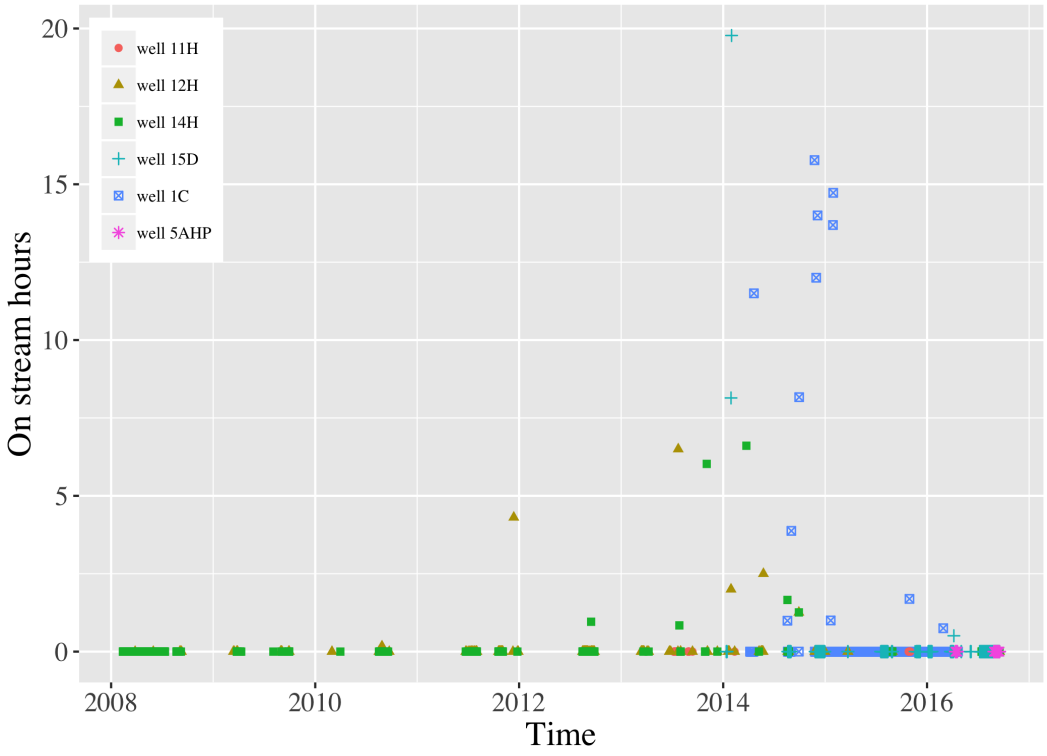


Figure 2: On-stream hours versus time for wells with oil production rate less than 10

Table 2: Size of wells data set

Well name	Well type	Size of data
Well 12H	production	2832
Well 14H	production	2718
Well 11H	production	1123
Well 15D	production	766
Well 1C	production	426
Well 5AHP	production	129
Well 4AH	injection	3327
Well 5AHI	injection	3146

Since we are now working on the data with oil production rate greater than 10, which basically means the wells are not closed for the operation, we can conclude that the zero values in all features are actually missing data at random. So, to get a better sense of missing data ratios, we converted these zeros to NA's. Table 3 shows the missing data ratios for production wells.

Table 3: Missing data ratios for production wells

Feature	Well 1C	Well 11H	Well 12H	Well 14H	Well 15D	Well 5AHP
Avg. annulus pressure	100%	0.53%	0.46%	49.6%	—	—
Avg. downhole pressure	—	0.45%	67.4 %	0.77%	—	100%
Avg. downhole temperature	—	0.45%	67.4%	0.77%	—	100%
Avg. ΔP tubing	—	0.45%	0.21%	0.22%	—	100%
Avg. well head pressure	—	0.45%	0.04%	0.04%	—	—
Avg. well head temperature	—	0.45%	—	—	—	—
ΔP chock size	—	0.45%	—	—	—	—
On-stream hours	—	0.09%	—	—	—	—

The data with missing ratios less than 5% will be simply removed. For well 5AHP, the highlighted features including average downhole pressure, average downhole temperature and Average ΔP tubing are missing. Since this well is the least important production well based on the size of data set, i.e., about 1% of all data, we consider removing this well for now. The highlighted missing ratios for other wells, including annulus pressure for well 1C and 14H, average downhole temperature and pressure for well 12H need to be imputed.

As we discussed earlier, water injection values given in 4AH and 5AHI wells dataset will be later used as a feature to build the predictive model, so, it is important to find the missing data ratios for the injection wells too. Like production wells, missing data ratios greater than 5% need to be imputed, which are 14.14% of water injection volume for well 5AHI and 10.13% of water injection volume for well 4AH. Since for imputation of water injection volume, we are going to use the feature itself, there is no need to remove missing data of on-stream hours, which are actually less than 5%; this will avoid unnecessary decreasing size of the data. We will discuss the details of the imputation in section 3.3.

3.2 Removing outliers

Outliers in all data are removed using the z-score method and a threshold equals to 3. Table 4 shows the outlier percentage for all wells. For injection wells, there is only water injection volume feature which we remove the outlier data from.

Table 4: Outliers percentage for each well

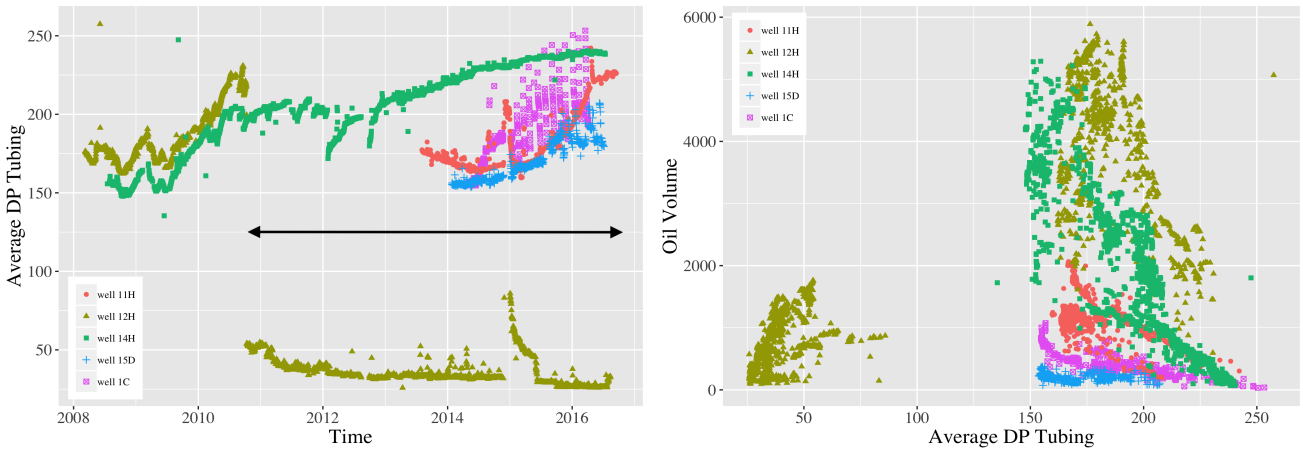
Well name	Well type	Outlier percentage
Well 12H	production	5.1%
Well 14H	production	6.5%
Well 11H	production	1.8%
Well 15D	production	8.3%
Well 1C	production	7%
Well 4AH	injection	no outlier
Well 5AHI	injection	no outlier

3.3 Data imputation

In this section we will discuss the details of imputation. In section 3.3.1 we will discuss the results of water injection volume imputation for well 4AH and well 5AHI using the Auto-Regressive Integrated Moving Average (ARIMA) model.

In section 3.3.3 and 3.3.4, we will discuss the results of average downhole pressure and average downhole temperature missing values imputation for well 12H and average annulus pressure missing values imputation for well 1C and well 14H. Basically, we will train a model on all available non-missing data to predict the missing values. The procedure starts with predicting average downhole pressure and average downhole temperature missing values and then using these two features to predict average annulus pressure missing values. The two approaches implemented and compared in this report for prediction of missing data are Multi-Layer Perceptron (MLP) and Support Vector Regression (SVR). However, the basic idea is to use all available features of all the wells when training the model. It is always useful to do a visual inspection on features to see if they give us any additional information before prediction so that it can be excluded when training the model.

Figure 3a shows average ΔP tubing versus time and figure 3b shows the oil production versus average ΔP tubing. From figure 3a, it can be seen that there is a sharp drop in average ΔP tubing for well 12H in a specific time period shown by the black arrow in the figure. Figure 3b confirms this abnormal behavior of the average ΔP tubing data for well 12H. So, we will replace the average ΔP tubing values less than 100 with NA's and then we will build a model to predict these missing values using average ΔP tubing non-missing values of all other wells. We discuss the results of the imputation in section 3.3.2.



(a) Average ΔP tubing vs. time

(b) Oil production vs. average ΔP tubing

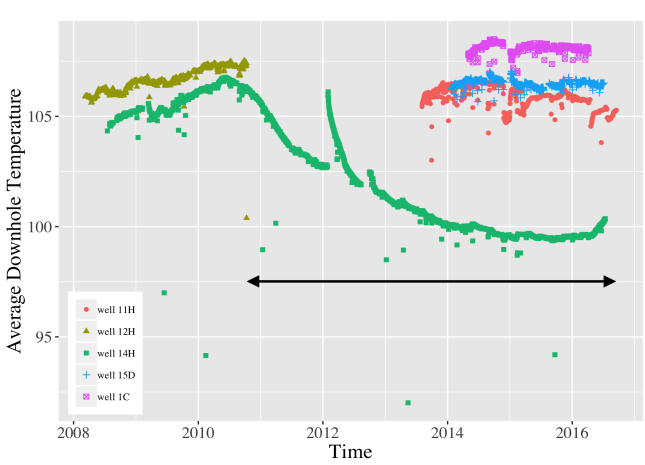
Figure 3: Visual inspection of average ΔP tubing for production wells

Moreover, since we want to use all wells' data to predict the missing values of a feature for a specific well, we need to see the scatter plots of the feature for all wells. Figure 4 to 6 show the time history and oil production plots for average downhole pressure, average downhole temperature and average annulus pressure, respectively. The black arrows in figure 4a and 5a show the missing values time range for average downhole pressure and temperature of well 12H. These figures also show that average downhole pressure and temperature data for well 1C, 11H and 15D have almost the same value ranges for time, oil production and average downhole pressure and temperature, and also the same for well 12H and 14H. Therefore, we decided to use only well 12H non-missing values and well 14H data to predict the missing values of average downhole pressure and temperature of well 12H, in addition to use the available data from all wells. We will compare both results in section 3.3.3.

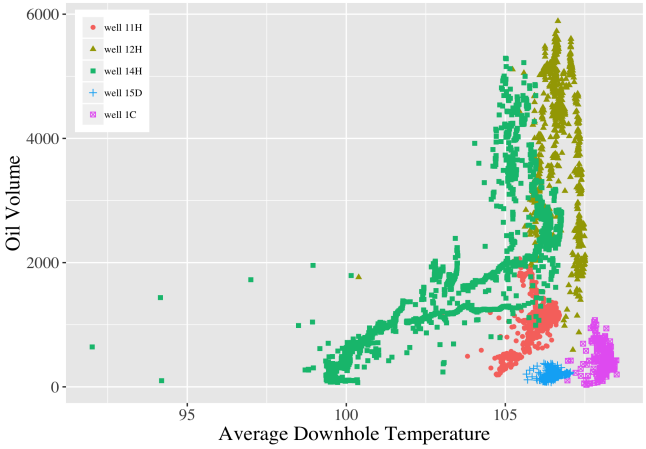
(a) Average downhole pressure vs. time

(b) Oil production vs. average downhole pressure

Figure 4: Visual inspection of average downhole pressure for production wells



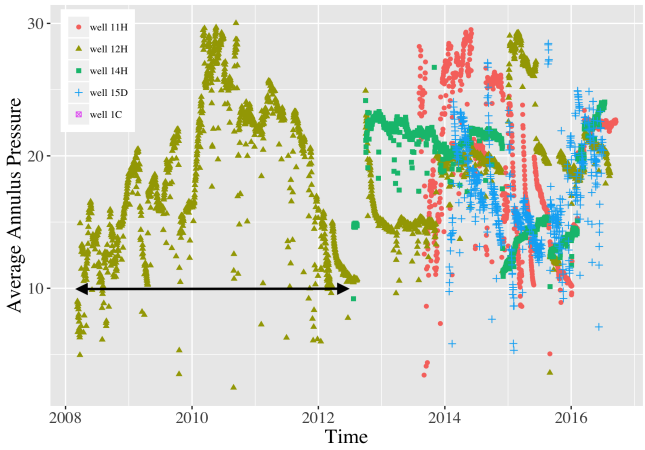
(a) Average downhole temperature vs. time



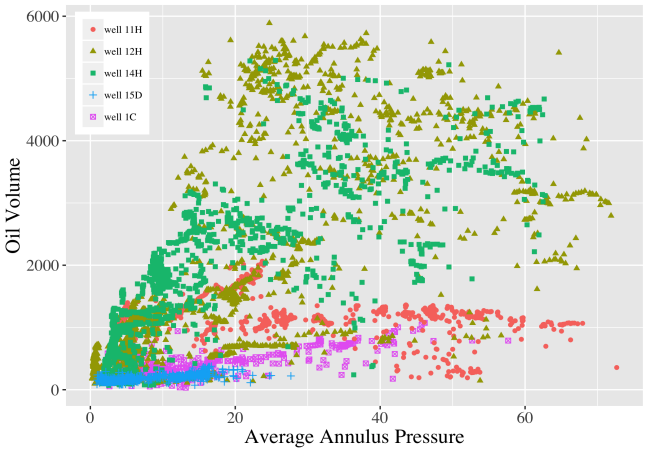
(b) Oil production vs. average downhole temperature

Figure 5: Visual inspection of average downhole temperature for production wells

The black arrows in figure 6a show the time range of the missing values for average annulus pressure for well 14H. We use non-missing average annulus pressure values of all four wells to predict missing values of average annulus pressure for well 1C and well 14H. We will discuss the results in section 3.3.4.



(a) Average annulus pressure vs. time

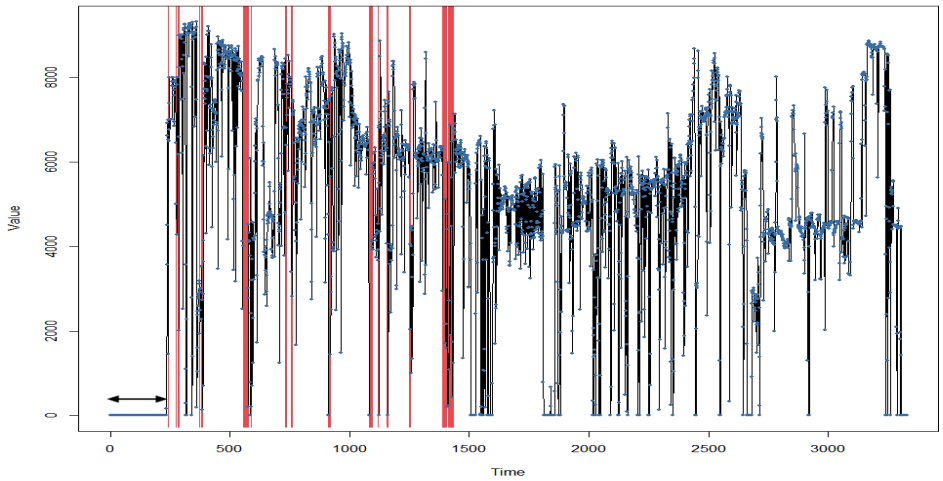


(b) Oil production vs. average annulus pressure

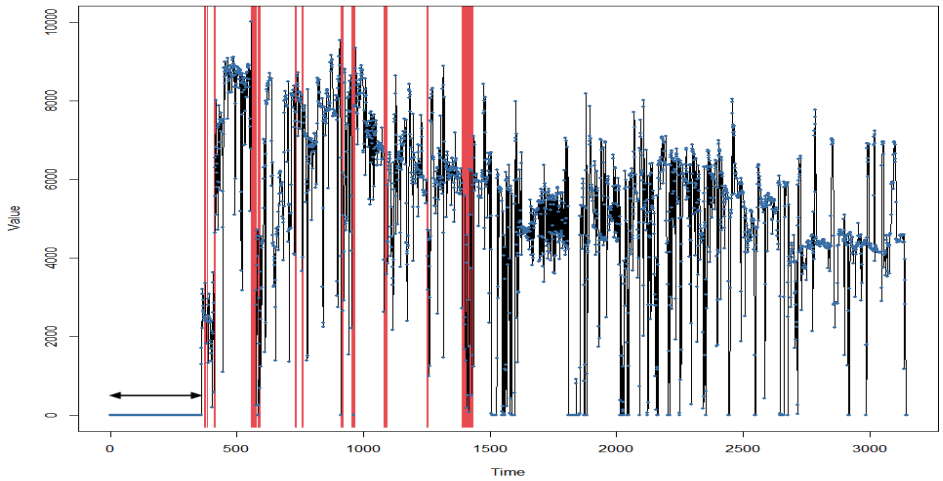
Figure 6: Visual inspection of average annulus pressure for production wells

3.3.1 Water injection volume

As discussed, both water injection wells, i.e., wells 4AH and 5AHI, have over 10% missing values for water injection volume. The missing values of water injection for wells 4AH and 5AHI are shown by red shading in figure 7. As shown in this figure, a part of data at the very beginning, which is shown by a black arrow, is zero. The zero values mean that no water has been injected to the other wells for this period of time. Thus, since zero values of water injection are meaningful, we keep the zero values and only impute the missing values shown by red shading in the figures.



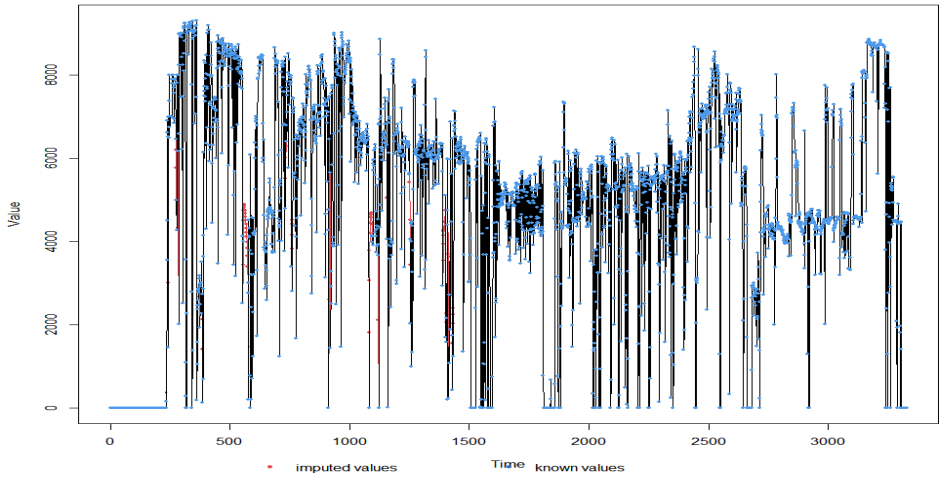
(a) Water injection missing values for well 4AH



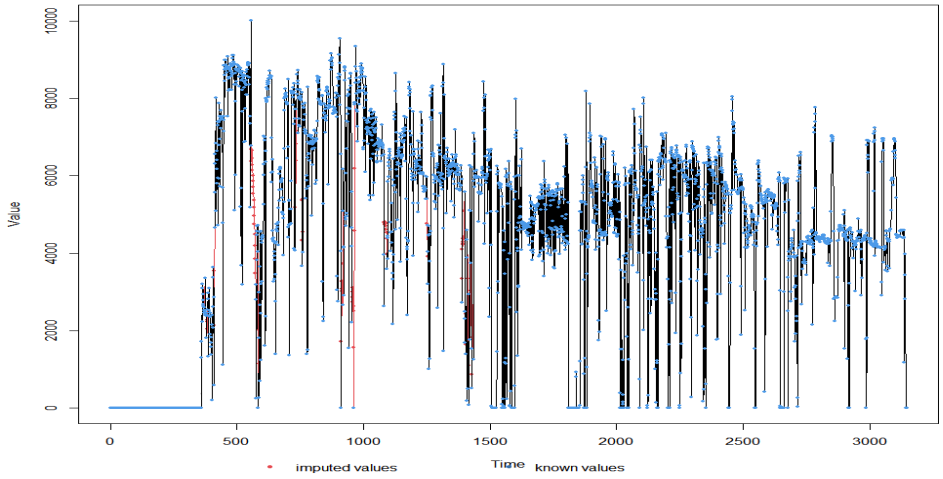
(b) Water injection missing values for well 5AHI

Figure 7: Visualization of the distribution of missing values of water injection volume. The time series is plotted and the background is colored in red whenever a value is missing

We used an ARIMA model [9] to impute the missing values. To apply this methodology on the water injection data, we use the package *"imputeTS"* [10], which automatically determines the best ARIMA model based on the given time series and then imputes the missing data. Figure 8 shows that the imputed values are following the trend of the original data. Here, imputation by ARIMA is more reasonable than merely using the mean value, due to the fact that the water injection is a univariate time series data.



(a) Water injection imputed for well 4AH



(b) Water injection imputed values for well 5AHI

Figure 8: Visualization of the imputed values of water injection volume. The imputed values (filled missing data gaps) are shown in red color

3.3.2 Average ΔP tubing

We used Support Vector Regression (SVR) with the radial basis function kernel [11,12] to impute missing values of average ΔP tubing well 12H. We used *"caret"* package [13] to train our model. We trained the model over several values for *"cost"* and *"sigma"* as tuning parameters. Table 5 shows the *"cost"* and *"sigma"* values that we used to tune the model to find the model with largest R-squared.

Table 5: Tuning parameters for average ΔP tubing SVR model

Cost	Sigma
1	2
8	3
64	5
256	6

The model has been validated using cross validation with 10 folds. We used on-stream hours, average chock size, average wellhead pressure, average wellhead temperature and ΔP chock size as the predictors to predict missing values for average ΔP tubing of well 12H. In fact, average downhole pressure, average downhole temperature and average annulus pressure have been excluded due to q large number of missing values. We used the datasets from wells 1C, 11H, 14H, 15D and 12H (non-missing values) to train the model on. We split the dataset into two training and test sets, which consisted of 80% and 20% of the dataset. The best SVR model picked with largest R-squared was tuned with $cost = 8$ and $sigma = 3$.

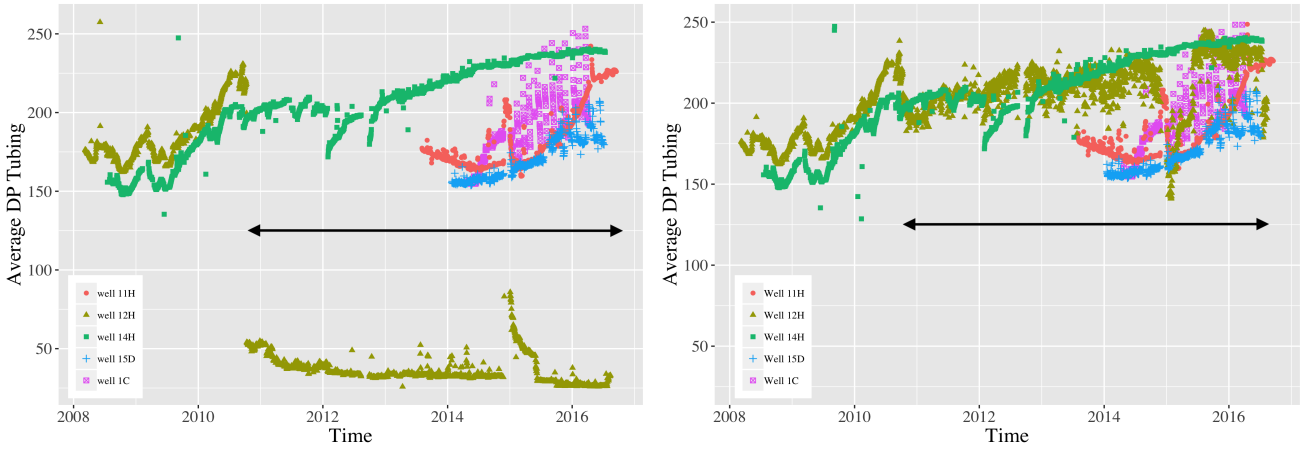
As said before, another approach used for predicting the missing values of average ΔP tubing for well 12H is MLP. For MLP, we split the whole dataset, which is data from wells 1C, 11H, 14H, 15D and 12H (non-missing values), into two training and test sets, which consisted of 80% and 20% of the whole dataset. In order to compare the results of SVR and MLP, the training and testing sets in both training models should be the same. After tuning the parameter and the structure of the MLP, the final structure of the NN has three hidden layers, and each layer has 20 neurons.

Table 6 shows the result of training models that has been tested for imputation of average ΔP tubing.

Table 6: Results of average ΔP tubing imputation

Imputed feature	Model	R-squared	RMSE	Sigma	Cost	Layer & Neurons	Epochs
average ΔP tubing	SVR	0.90	8.00	3	8	—	—
average ΔP tubing	MLP	0.91	0.308	—	—	[32, 16]	100

Based on the results of table 6, the SVR model has been used to predict the missing values of average ΔP tubing for further analysis. Figure 9a and 9b show the average ΔP tubing plot for all wells before and after missing data imputation. The black arrow in figure 9b shows the time period where average ΔP tubing abnormal data of well 12H has been imputed.



(a) Before imputation

(b) After imputation

Figure 9: Average ΔP tubing before and after imputation, the black arrow on the right figure shows the time period where average ΔP tubing abnormal data of well 12H have been imputed

3.3.3 Average downhole pressure and temperature

In this section, we discuss the imputation of average downhole pressure and temperature missing values of well 12H using MLP and SVR models.

For both training models, we used on-stream hours, average chock size, average wellhead pressure, average wellhead temperature and ΔP chock size as the predictors to predict missing values for average downhole pressure and temperature of well 12H. In fact, average annulus pressure has been excluded due to large number of missing values as well as average ΔP tubing due to previously mentioned abnormal behavior. The same training and testing sets have been used as described in section 3.3.2 for all models. Also, the models have been validated using cross validation with 10 folds.

For the MLP model, we used the datasets from wells 1C, 11H, 14H, 15D and 12H (non-missing values) to train our model on, and we kept the same number of hidden layers and neurons as section 3.3.2. For the SVR model, we used two sets of data, the first using data from wells 1C, 11H, 14H, 15D and 12H (non-missing values), so called SVR1, the second using data of well 12H (non-missing values) and 14H, so called SVR2. Table 9 shows the *"cost"* and *"sigma"* values that we used to tune the model to find the model with largest R-squared.

Table 7: Tuning parameters for average downhole pressure and temperature SVR model

Cost	Sigma
3	5
5	6
6	8
7	10
8	12
9	
10	

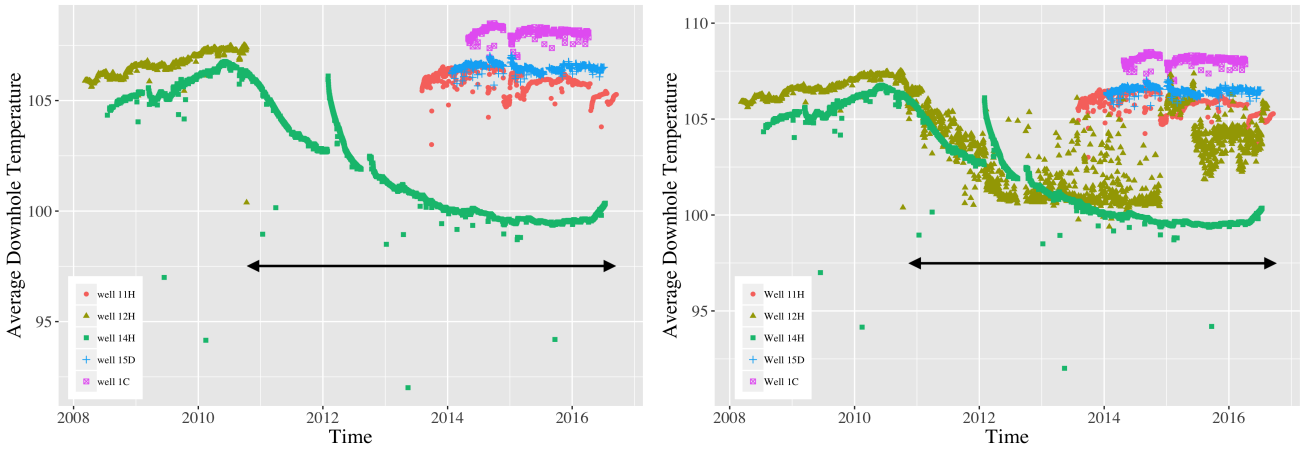
Table 8 shows the result of training models that have been tested for imputation of average downhole pressure and temperature of well 12H. The *"cost"* and *"sigma"* values shown in the table are the values which give us the SVR model with the largest R-squared.

Table 8: Results of average downhole pressure and temperature imputation

Imputed feature	Model	R-squared	RMSE	Sigma	Cost
average downhole pressure	SVR1	0.89	7.44	5	6
average downhole temerature	SVR1	0.94	0.63	5	3
average downhole pressure	SVR2	0.95	4.44	5	3
average downhole temperature	SVR2	0.93	0.16	5	3
Imputed feature	Model	R-squared	RMSE	Layer & Neurons	Epochs
average downhole pressure	MLP	0.86	0.37	[32, 16]	100
average downhole temperature	MLP	0.95	0.23	[32, 16]	100

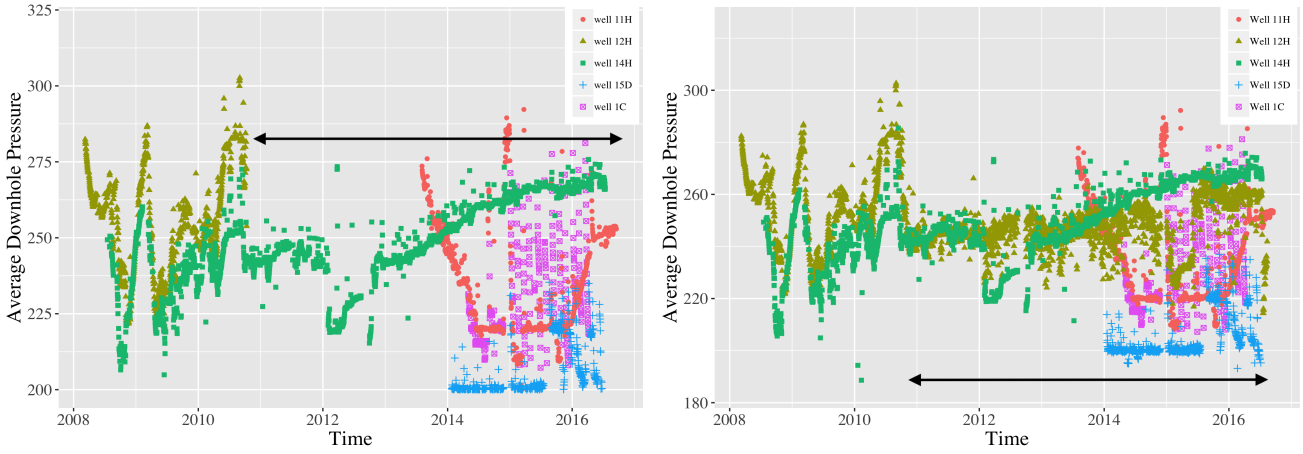
Based on the results of table 8, the SVR1 model resulted in a larger R-squared and smaller RSME, but this result can not be reliable due to the small size of training set used to build the model. So, the SVR2 model which has been trained on all wells' datasets has been used to predict the missing values of average downhole pressure and temperature for further analysis .

Figures 10 and 11 show the average downhole pressure and temperature plots of all wells versus time before and after missing data imputation. The black arrows in figures 10b and 11b show the time period where average downhole temperature and pressure missing data of well 12H have been imputed.



(a) Before imputation (b) After imputation

Figure 10: Average downhole temperature before and after imputation, the black arrow on the right figure shows the time period where average downhole temperature missing data of well 12H have been imputed



(a) Before imputation (b) After imputation

Figure 11: Average downhole pressure before and after imputation, the black arrow on the right figure shows the time period where average downhole pressure missing data of well 12H have been imputed

3.3.4 Average annulus pressure

In this section, we discuss the imputation of average annulus pressure missing values of well 1C and 14H using MLP and SVR models. For both training models everything is kept the same as what we have done in section 3.3.2 and 3.3.3, except that now we use new values of average ΔP tubing, average downhole pressure and average downhole temperature along with on-stream hours, average chock size, average wellhead pressure, average wellhead temperature and ΔP chock size as predictors. We trained both models using the datasets of wells 11H, 12H,

14H (non-missing values) and 15D. Table 9 shows the *"cost"* and *"sigma"* values that we used to tune the model to find the model with the largest R-squared.

Table 9: Tuning parameters for average annulus pressure SVR model

Cost	Sigma
5	20
8	25
10	30
	40

Table 10 shows the results of training models that have been tested for imputation of average downhole pressure and temperature of well 12H. The *"cost"* and *"sigma"* values shown in the table are the values which give us the SVR model with the largest R-squared. Based on the results of table 10, the SVR model has been used to predict the missing values of average annulus pressure of well 1C and Well 14H for further analysis. Figures 12a and 12b show the average annulus pressure of well 1C and 14H versus time before and after missing data imputation. The black arrow in figure 12b shows the time period where average annulus pressure missing data of well 14H has been imputed. Now that all the missing values have been imputed, we would like to see the correlation between features and target as well as the level of importance of features. In the next section, we will discuss the correlation and the PCA plots of the cleaned data.

Table 10: Results of average annulus pressure imputation

Imputed feature	Model	R-squared	RMSE	Sigma	Cost
average annulus pressure	SVR	0.77	2.30	5	20
Imputed feature	**Model**	**R-squared**	**RMSE**	**Layer & Neurons**	**Epochs**
average annulus temperature	MLP	0.74	0.50	[64, 32]	200

(a) Before imputation

(b) After imputation

Figure 12: Average annulus pressure before and after imputation, the black arrow on the right figure shows the time period where average annulus pressure missing data of well 14H have been imputed

4 Correlation and PCA plots

In this section we will show the correlation and PCA plots after data cleaning and imputation discussed in previous sections. Figure 13 shows the correlation plot for all of the production wells. The figure shows that the most correlated features with our target, which is oil volume, are average well head temperature and ΔP chock size, yet the values are not large enough to conclude that they are highly correlated with oil volume. Also, it can be seen that oil volume is not correlated with average downhole pressure and average annulus pressure. Between the features themselves, ΔP chock size and average water head pressure are highly correlated, also average water head temperature and average chock size are relatively correlated.

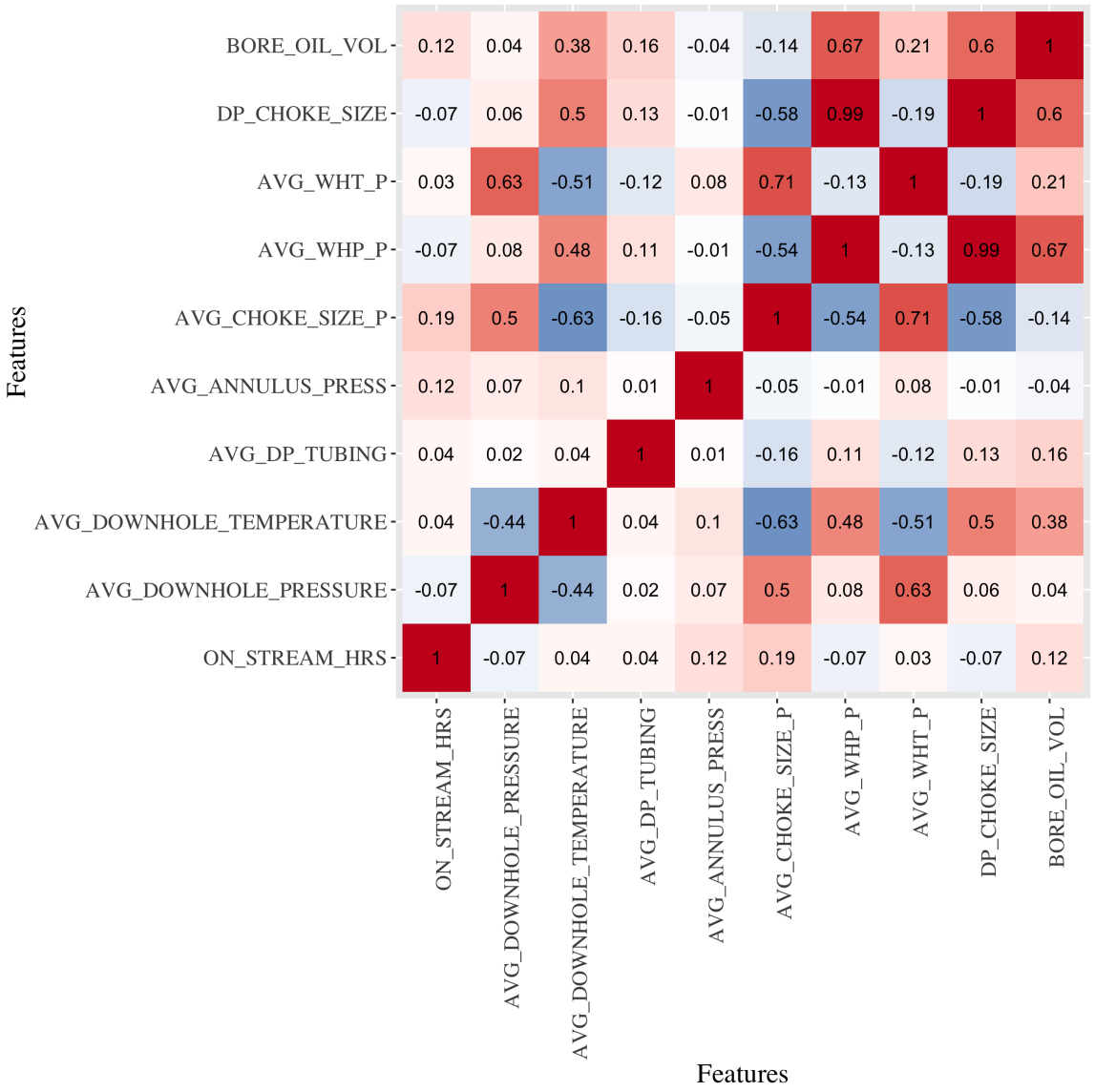


Figure 13: Correlation plot for all production wells

Figure 14 shows the PCA plot for all of the production wells. As we can see, the least important features are average annulus pressure and on-stream hours. Since, there are only 9 features and since the correlation and PCA plots give us ambiguous information about the features level of importance, we decided to use all features for final analysis. In the next section, we will discuss the results of oil production prediction using all the cleaned data from all the wells along with the water injection values.

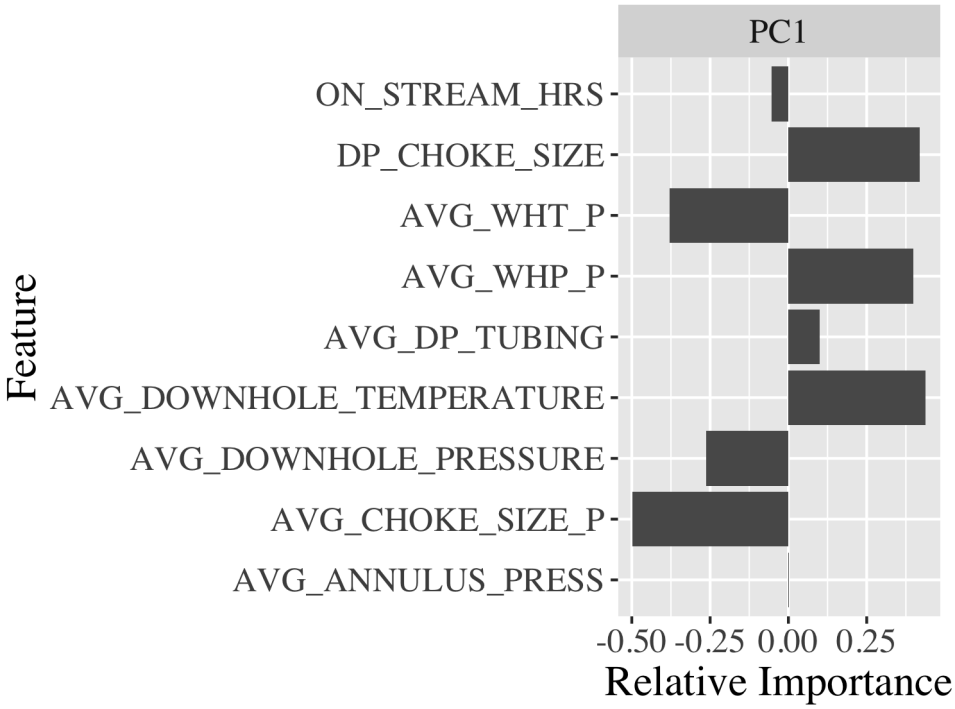


Figure 14: PCA plot for all production wells

5 Time-series prediction

In this section, we discuss the results of SVR and Long Short-Term Memory (LSTM) models used to predict the oil production rates. To build both models, we used all the features including on-stream hours, average downhole pressure, average downhole temperature, average ΔP tubing, average choke size, average annulus pressure, wellhead pressure, wellhead temperature and ΔP choke size as well as water injection values.

In this project, the dataset consists of historical features. Therefore, any prediction of the oil production will be dependent on the historical feature. Then, to prepare the dataset for SVR, we first convert the raw dataset to a time series structure. The input X_{SVR} of SVR should include n time steps, the input features and the labels. The converted input features are of the following form:

$$\begin{aligned} X_{SVR} = [&Var_1(t-n), Var_2(t-n), \cdots, Var_m(t-n), label(t-n), \cdots, Var_1(t-1), \\ &Var_2(t-1), \cdots, Var_m(t-1), label(t-1), \cdots, Var_1(t) Var_2(t) \cdots Var_m(t)], \end{aligned} \tag{1}$$

where Var_i is the ith feature, and $label(t-j)$ is the oil production value of time step j. The output Y_{SVR} of SVR is:

$$Y_{SVR} = [label(t)]. \tag{2}$$

Taking advantage of the LSTM's insensitivity to the time interval length, the Recurrent Neural Network (RNN) structure may help provide better predictions on the dataset. Accordingly, the LSTM method is implemented to forecast the oil production. The LSTM model will learn a function that maps a sequence of the past observations as the input to the output observations. The sequence of the input observations must be transformed into the multiple examples from which the LSTM can learn. The input X_{LSTM} of LSTM should include all the input features of each time step, so the input data is constructed as the follows:

$$X_{LSTM} = \begin{bmatrix} Var_1 t-n & Var_2 t-n & \cdots & Var_m t-n \\ \vdots & \vdots & \cdots & \vdots \\ Var_1 t-1 & Var_2 t-1 & \cdots & Var_m t-1 \\ Var_1 t & Var_2 t & \cdots & Var_m t \end{bmatrix}, \tag{3}$$

And the output Y_{LSTM} is:

$$Y_{LSTM} = [label(t)]. \tag{4}$$

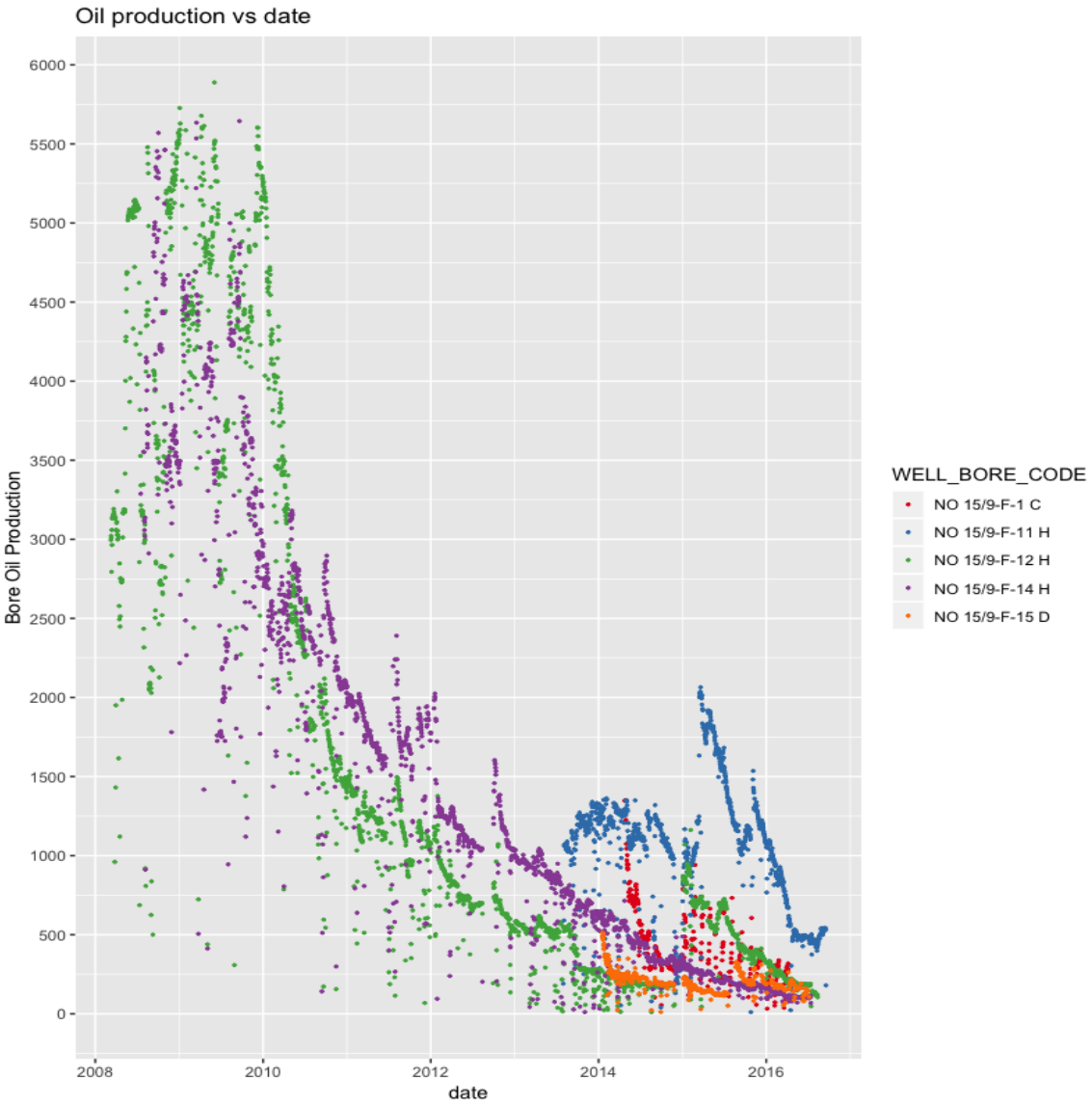


Figure 15: Oil production of each well VS time

It can be seen from Figure 15, the range of oil production rates of wells is different. The range of production from wells 12H and 14H is $[0, 6000]$, the range of well 1C is $[250, 1250]$, the range of well 11H is $[350, 2100]$, and the range of well 15D is $[0, 500]$. Normalization or standardization of training data is a crucial step to for model development and accurate results. The normalization process adjusts the dataset based on the minimum and maximum values ? this can make training less sensitive to the scale of features and reduces variance. Standardization normalizes the data about the mean and standard deviation, which can bring the training dataset to a consistent scale. As the range of the production is varying from well to well, after normalizing, the data of wells 1C, 11H, and 15D will approach to zero. Therefore, we standardized data to prepare the training set. To minimize the effect of the randomness when splitting the datasets into training and test sets, we assign the same random seed of 7 in both methods. The data preparation steps can be summarized in the below steps:

Algorithm 1 Standardize data

1: Setting the random seed equal to 7
2: Constructing the time series data
3: Shuffling the time series data
4: Splitting first 80% data as training set and rest 20% data as test set
5: Calculating mean and standard deviation of the training data
6: Normalizing all the data based on mean and standard deviation

5.1 Preliminary results

Preliminary prediction using SVR with the radial basis function kernel for oil volume is conventional supervised learning without constructing time series dataset. We used all the features of cleaned data to train our model. The dataset was shuffled and split into training and test sets with the ratio as 80% and 20%, respectively. A grid search was implemented on the hyper-parameters *"cost"* and *"sigma"* to find the model with the largest R-squared, i.e. 1, 2, 4, 8, 16 and 32 for *"cost"* and 0.0625, 0.25, 1, 4, 16 and 64 for *"sigma"*. The model has been validated using cross-validation method with 10 folds. The average R-squared of this prediction model is 0.97 with $cost = 16$ and $sigma = 1$. With the same input features, the average R-squared of prediction using MLP is 0.98. The MLP has two layers, the first layer has 32 neurons and the second layer has 16 neurons. The overall results are shown in Table 11.

Table 11: Results of oil production prediction without time series using SVR

Imputed feature	Model	R-squared	RMSE	Sigma	Cost
oil production	SVR	0.97	2.30	1	16
Imputed feature	Model	R-squared	RMSE	Layer & Neurons	Epochs
oil production	MLP	0.98	0.15	[32, 16]	150

5.2 SVR results

For the preliminary studies using SVR and MLP, the historical feature and water injection features were not utilized. For further studies the results of SVR and LSTM models considered both the aforementioned features as well. As we previously mentioned, the oil production value range of wells are different. Thus, we use the previously mentioned approach for normalization process. The dataset is divided into 7 combinations including, 11H, 1C, 15D, 11H, 1C, 11H, 1C, 15D, 12H, 14H and 11H, 1C, 15D, 12H, 14H. A hyper-parameter has been performed using different number of time-steps, which is 3, 5, 10, 20, *cost* and the *sigma*. The result of hyper-parameter tuning is shown in table associated with the best results are shown in Table The prediction results of SVR is also shown in the Table 13. We can see from the table that the best results has been achieved when training the model on datasets 11H, 1C, 15D and 12H, 14H.

Table 12: Hyper-parameter tuning results for SVR

Model name	Dataset	Time-steps	cost	sigma
SVR	{11H}	5	100	0.001
SVR	{1C}	5	100	0.0001
SVR	{15D}	5	100	0.0005
SVR	{11H, 1C}	5	64	0.004
SVR	{11H, 1C, 15D}	5	100	0.001
SVR	{12H, 14H}	5	16	0.001
SVR	{11H, 1C, 15D, 12H, 14H}	5	100	0.001

Table 13: Average Results of time series oil production prediction using SVR

Datasets	{11H}		{1C}		{15D}		{11H, 1C}		{12H, 14H}		{11H, 1C, 15D}		{All Wells}	
Metrics	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
11H	0.96	0.19	—	—	—	—	0.95	0.18	—	—	0.96	0.14	0.94	0.06
1C	—	—	0.93	0.26	—	—	0.88	0.15	—	—	0.92	0.11	0.78	0.06
15D	—	—	—	—	0.82	0.41	—	—	—	—	0.65	0.06	-0.23	0.04
12H	—	—	—	—	—	—	—	—	0.99	0.09	—	—	0.99	0.09
14H	—	—	—	—	—	—	—	—	0.99	0.07	—	—	0.99	0.08
Overal	0.96	0.19	0.93	0.26	0.82	0.41	0.96	0.17	0.99	0.08	0.98	0.11	0.99	0.08

5.3 LSTM results

Recently, LSTM recurrent neural network (RNN) technique has drawn great attention in time-series modeling because of its capability of handling input data spanning over long sequences. Unlike feedforward neural networks which may stack layers with different number of neurons, RNNs (figure 16) keep a constant structure and recurrently feed its output into the input, and can use their internal cells to process sequential inputs. As a vanilla RNN structure usually brings the vanishing gradient problem, LSTM RNN (figure 17) comes to solve the problem.

Generally, RNN cell takes as input two vectors: observation and a hidden state, and produces the current observation and the hidden state. Besides, LSTM includes four gates that handle memory and forgetting: cell state, forget gate, input gate, and output gate. Multiplied by the forget gate, the cell state only keeps information that passes through the forget gate. The forget gate applies a sigmoid function to previous observation and hidden state to determine to either keep or skip the information. The input gate deploys a sigmoid and a tanh activation function to allow information entering the cell state. The cell state is also known as the long-term memory, and its recursive nature allows historical information to be stored in the state. The cell state is modified by both the forget gate and the input gate, which dynamically combines historical information and recent observation. The output gate deploys a sigmoid activation function to determine which part of information should go to the next hidden state.

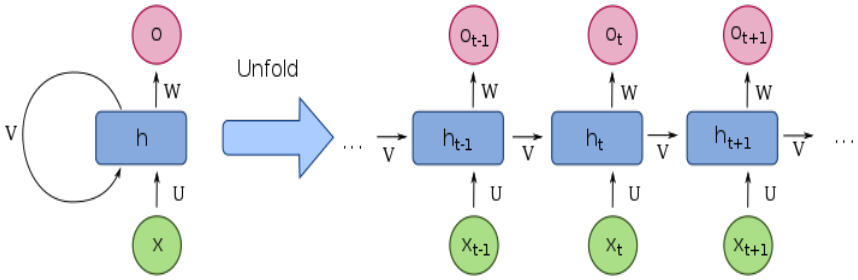


Figure 16: RNN structure (credit by Aamir Nazir)

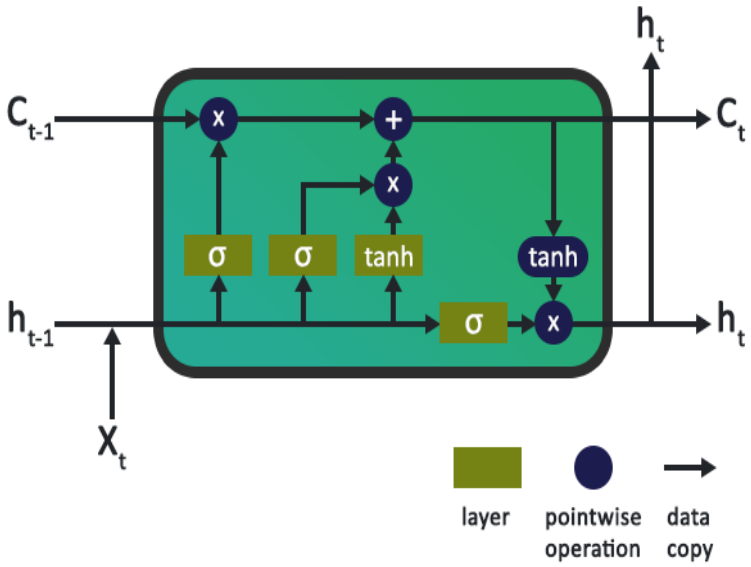


Figure 17: LSTM RNN structure (credit by Jakob Aungiers)

For the LSTM studies, we investigated the same 7 combinations of datasets as SVR. To tune the model, we trained the model over different number of time -steps , which is 3, 5, 10, 20, different number of neurons and epochs. The results of hyper-parameter tuning are shown in Table 14. The model evaluation results of each well and the overall results of oil production prediction using LSTM are shown in Table 15. To minimize the randomness, the results in Table 15 is the average out of 10 trials. From the first 5 cases in Table 15, we can see that the training results is based on how well the dataset is. As mentioned in the previous section, the best combination of dataset is that of 11H, 1C, 15D and12H, 14H. We can conclude that when the oil production rates of wells are in a similar range, better prediction results can be achieved on each well and overall results. In the last case, we can conclude that, although the training results are good for some wells, the range and magnitude of well production rates significantly affects results.

Table 14: Hyper-parameter tuning results for LSTM

Model name	Dataset	Time-steps	Neurons	Epochs
LSTM	{11H}	5	20	60
LSTM	{1C}	5	20	60
LSTM	{15D}	5	20	60
LSTM	{11H, 1C}	10	50	60
LSTM	{11H, 1C, 15D}	10	50	60
LSTM	{12H, 14H}	10	50	60
LSTM	{11H, 1C, 15D, 12H, 14H}	10	50	60

Table 15: Average Results of time series oil production prediction using LSTM

Datasets	{11H}		{1C}		{15D}		{11H, 1C}		{12H, 14H}		{11H, 1C, 15D}		{All Wells}	
Metrics	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
11H	0.96	0.04	—	—	—	—	0.97	0.01	—	—	0.96	0.02	0.91	0.08
1C	—	—	0.91	0.07	—	—	0.92	0.01	—	—	0.92	0.01	0.71	0.74
15D	—	—	—	—	0.84	0.13	—	—	—	—	0.77	0.003	0.87	0.09
12H	—	—	—	—	—	—	—	—	0.99	0.007	—	—	0.11	0.90
14H	—	—	—	—	—	—	—	—	0.99	0.005	—	—	0.83	0.13
Overal	0.96	0.04	0.91	0.07	0.84	0.13	0.98	0.01	0.99	0.006	0.98	0.01	0.87	0.11

6 Future Analysis

As we discussed in previous section, the results of LSTM and SVR when using history as a feature show that the best result comes out when combining wells 12H and 14H (dataset D) which is $R - squared = 0.99$ and when combining wells 1C, 11H and 15D (dataset G) which is $R - squared = 0.98$. However, LSTM and SVR show different results when we train the model on well 1C, 11H and 15D datasets separately (datasets A, B and C respectively). The SVR model shows a very good result when training the model on all datasets. But, this results can not be reliable since the training model is mostly affected by wells 12H and 14H as they consist 70% of the dataset. The LSTM model, on the other hand, does not show a very good result when we train the model on the data from all the wells. The question is do we need to separate the datasets into two sets, i.e., {12H , 14H} and {11H, 1C, 15D} and train two different models on them, combining the data from all the wells together.

To address the above question, we still need to train other models on the current dataset , thus, the next step will be investigating the ARIMA and the Gated Recurrent Unit (GRU) models to predict the oil production rate.

7 References

1. Thompson, R.S., and Wright, J.D.. Oil property evaluation. United States: N. p., 1984. Web.
2. 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).
3. John, E.G., 1998. Simplified curve fitting using spreadsheet add-ins. International Journal of Engineering Education 14 (5), 375?380
4. 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)?
5. 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.
6. 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
7. Jason Brownlee, How to Handle Missing Data with Python (https://machinelearningmastery.com/handle-missing-data-python/)
8. Alvira Swalin, How to Handle Missing Data (https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4)
9. 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.
10. Robert Nau, Introduction to ARIMA models (http://people.duke.edu/∼rnau/411arim.htm#arima010)
11. Steffen Moritz, Time Series Missing Value Imputation (https://cran.r-project.org/web/packages/imputeTS/imputeTS.pdf)
12. Drucker, Harris; Burges, Christopher J. C.; Kaufman, Linda; Smola, Alexander J.; and Vapnik, Vladimir N. (1997); "Support Vector Regression Machines", in Advances in Neural Information Processing Systems 9, NIPS 1996, 155?161, MIT Press.
13. Ryan Kelly , Support Vector Machines (https://rpubs.com/ryankelly/svm)
14. Max Kuhn, The Caret Package (https://topepo.github.io/caret/index.html)