A time series analysis on prediction of oil production rate

Authors: Maryam Bagheri, Li Huang, Manyang Sun, Haoran Zhao

University of Houston

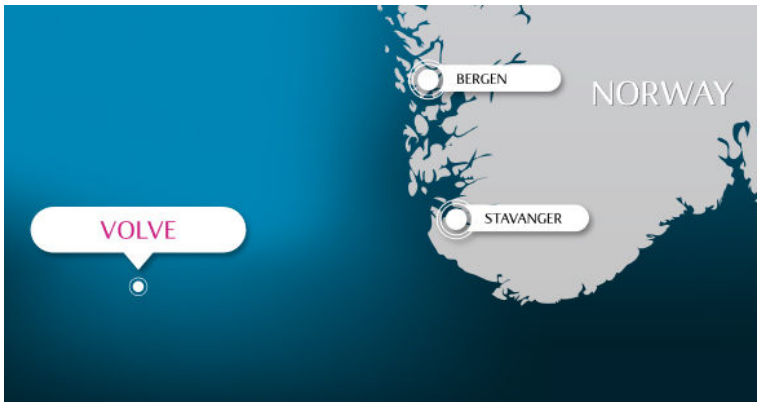
Supervisers:
Srinath Madasu, Peggy Linder, Giulia Toti

April 25, 2019

Overview

Problem Statement

- *Volve*, being operated for approximately 8 years till its shutdown in 2016, 63 million bbl of oil
- The most comprehensive and complete dataset ever gathered on the NCS
- Data ranges from 2008 to 2016, 100000 observations from six production wells and two water injection wells
- Exploring the complex nonlinearity of features and make a prediction on bore oil production



Why it is important?

A well-trained model as a powerful tool to support future oil drilling system design and optimization

Problem Statement

- 23 types of information, covering more than 15000 operation days
- 12 of them are directly related to our target
- Prediction on oil production rate based on these features

Features
Date of production
On-stream hours
Average downhole pressure
Average downhole temperature
Average differential pressure tubing
Average annulus pressure
Average choke size
Average wellhead pressure
Average wellhead temperature
Differential pressure at choke
Production/injection
Bore water injection volume

Literature Review

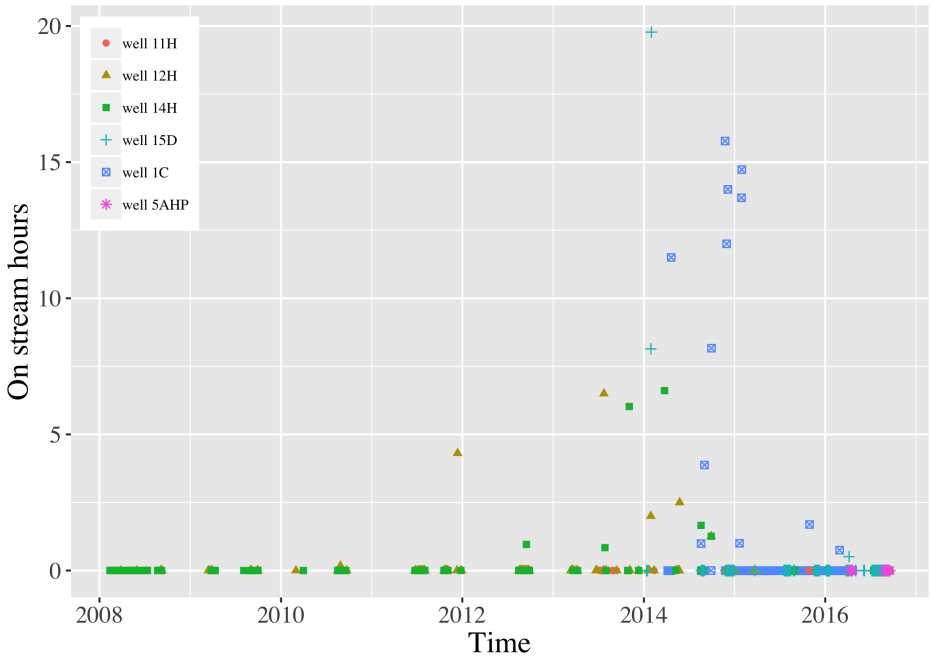
- **Petroleum production prediction**(Thomas, R.S. *et.al.*):
  1. by analogy
  2. volumetric
  3. material balance
  4. decline curve fitting
  5. reservoir simulation
- Artificial Neural Network (ANN): the most recent method is to estimate production values (Wong, P.M.*et al.*).
- Higher accuracy when compared to other correlation methods and curve estimation methods (Gharbi, R.B.,*et al.*).
- Data pre-processing as the most important step in applying the ANN

Dealing with NA's and zeros considering the physical meaning of features

Data Cleansing

most data scientists spend almost 60% of their time and effort on data cleansing

Oil production rates less than 10:



After removing the oil production rates less than or equal to 10:

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

Data Cleansing - Missing Data Ratios

Missing data ratios of 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%	—	—	—	—

For injection wells:
14.14% of water injection volume for well 5AHI
10.13% of water injection volume for well 4AH

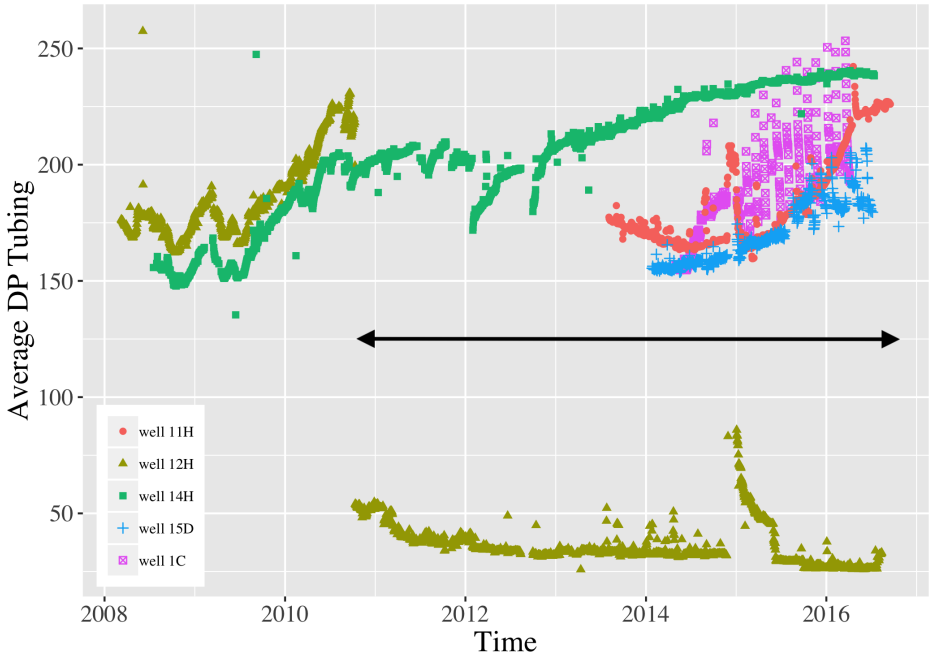
Data Cleansing - Removing Outliers

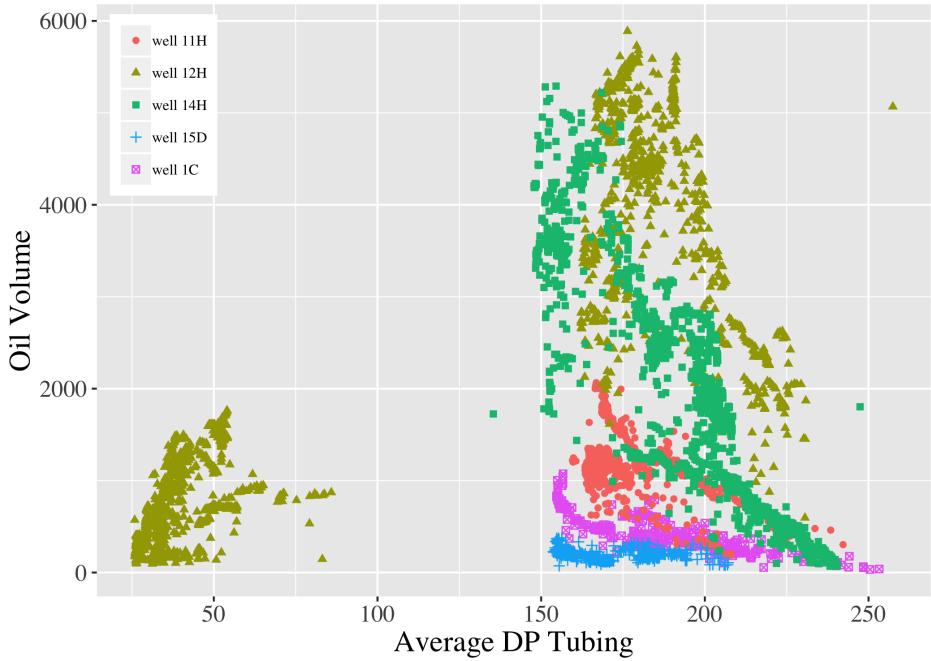
After using z-score with a threshold equals to 3 to drop the outliers:

Well code	**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

Data Cleansing - Missing Data Imputation

Some visual inspections on features...

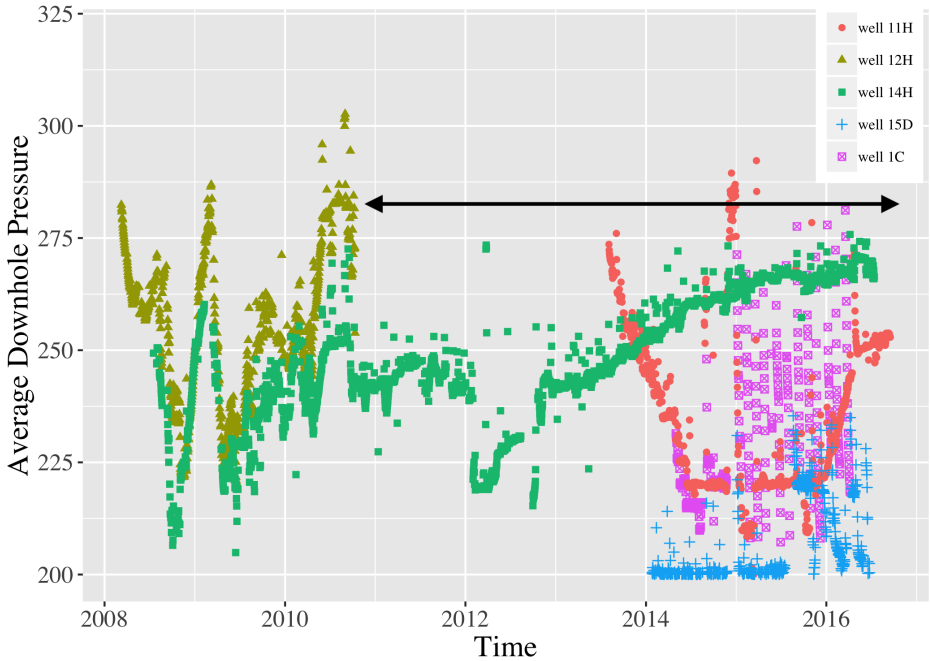


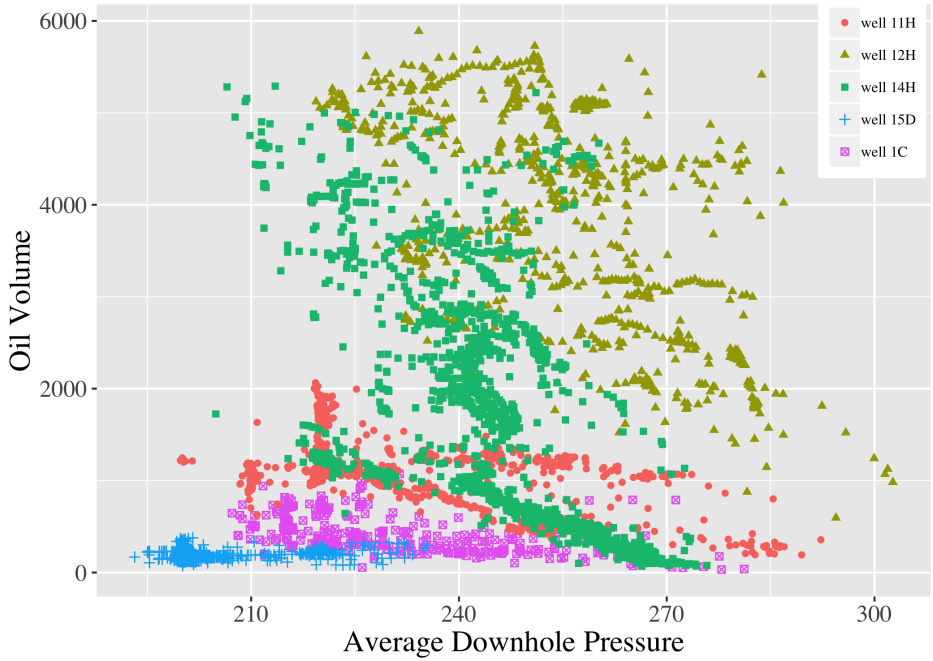


The values less than 100 will be replaced with NA's

Data Cleansing - Missing Data Imputation

Some visual inspections on features...

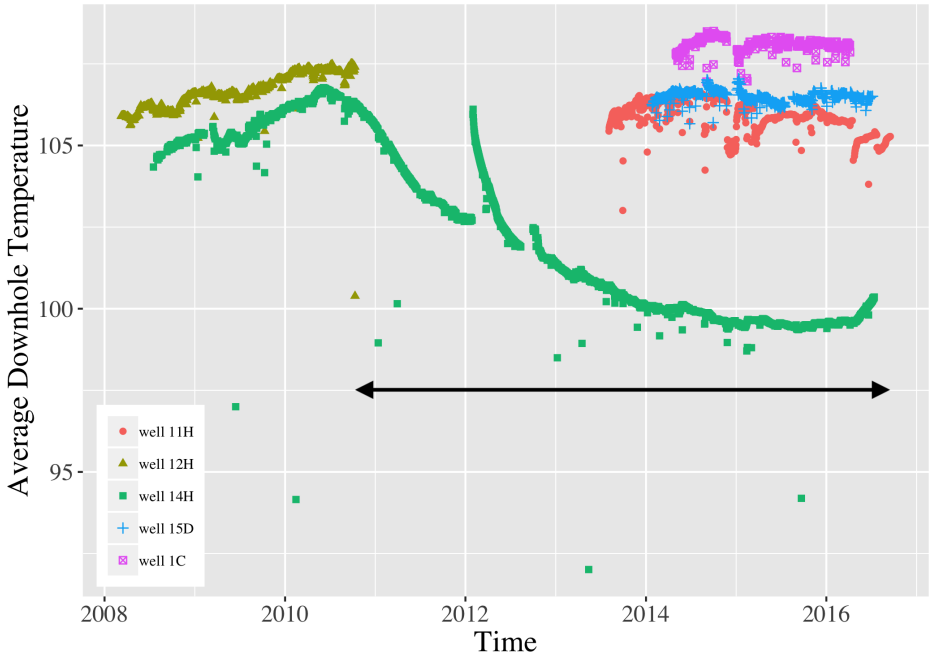


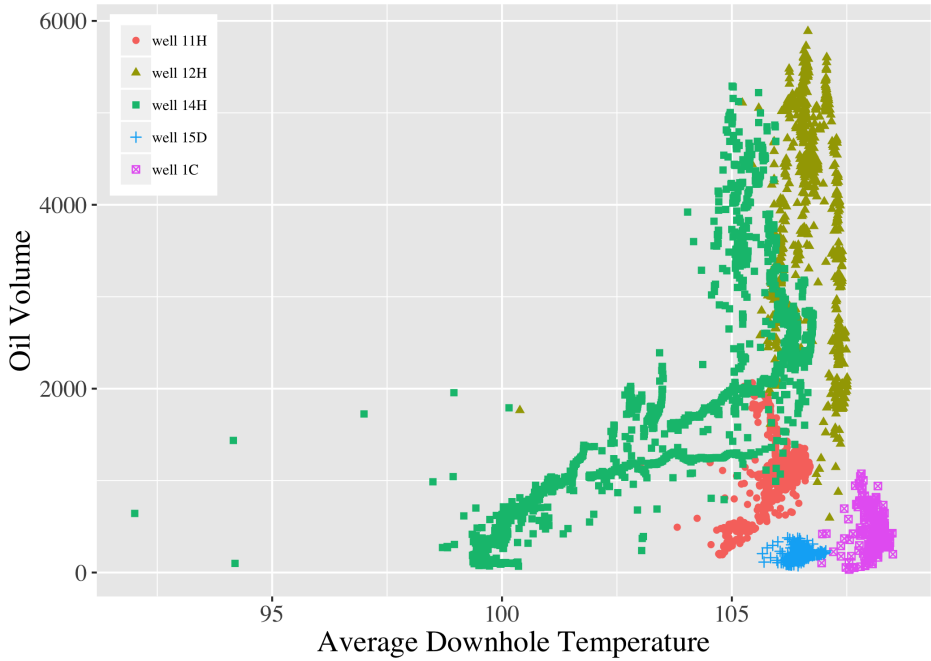


The values in the shown time range will be imputed

Data Cleansing - Missing Data Imputation

Some visual inspections on features...

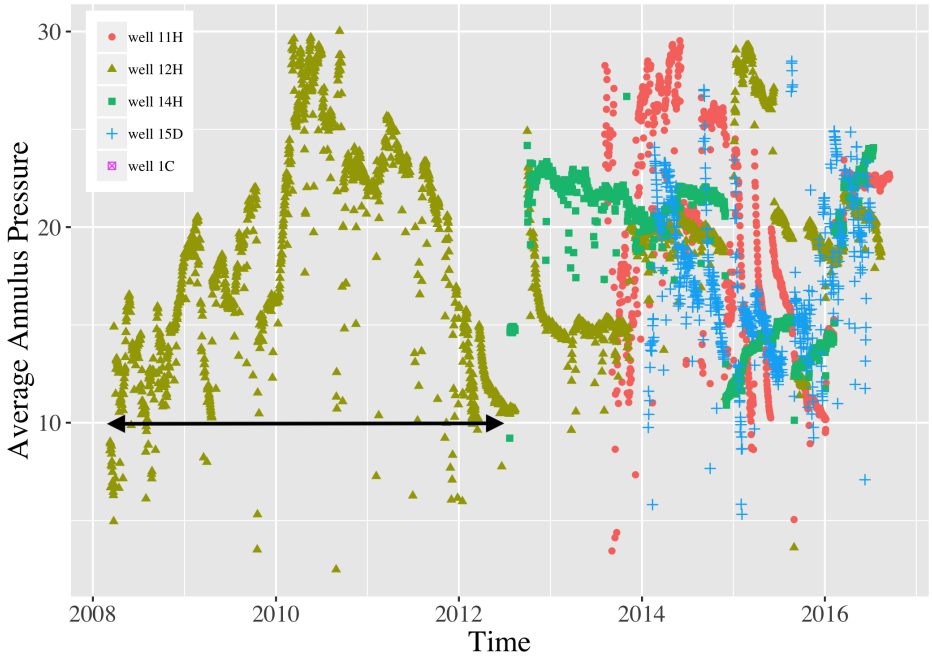


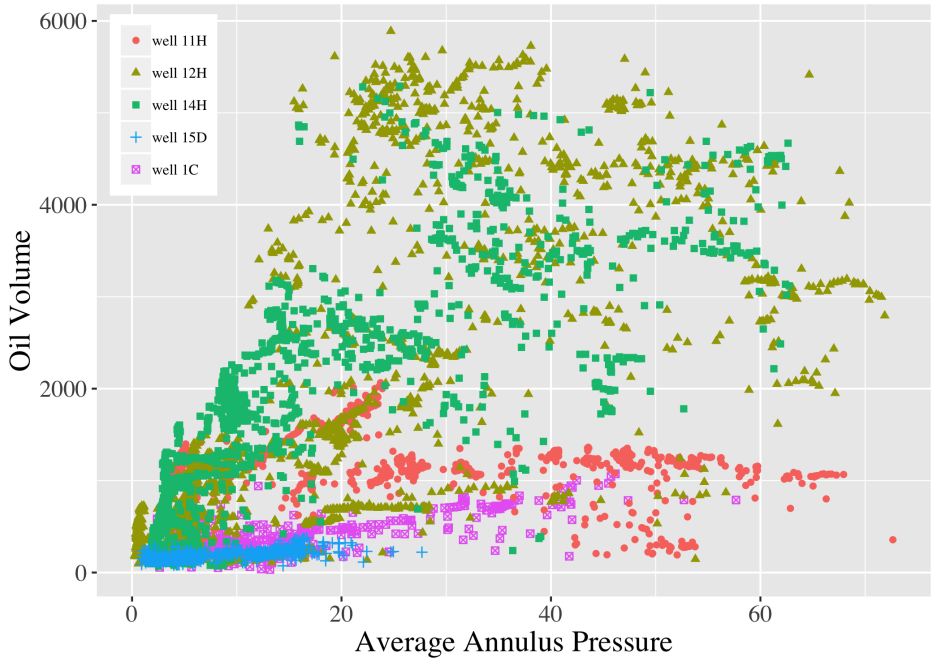


The values in the shown time range will be imputed

Data Cleansing - Missing Data Imputation

Some visual inspections on features...





The values in the shown time range will be imputed

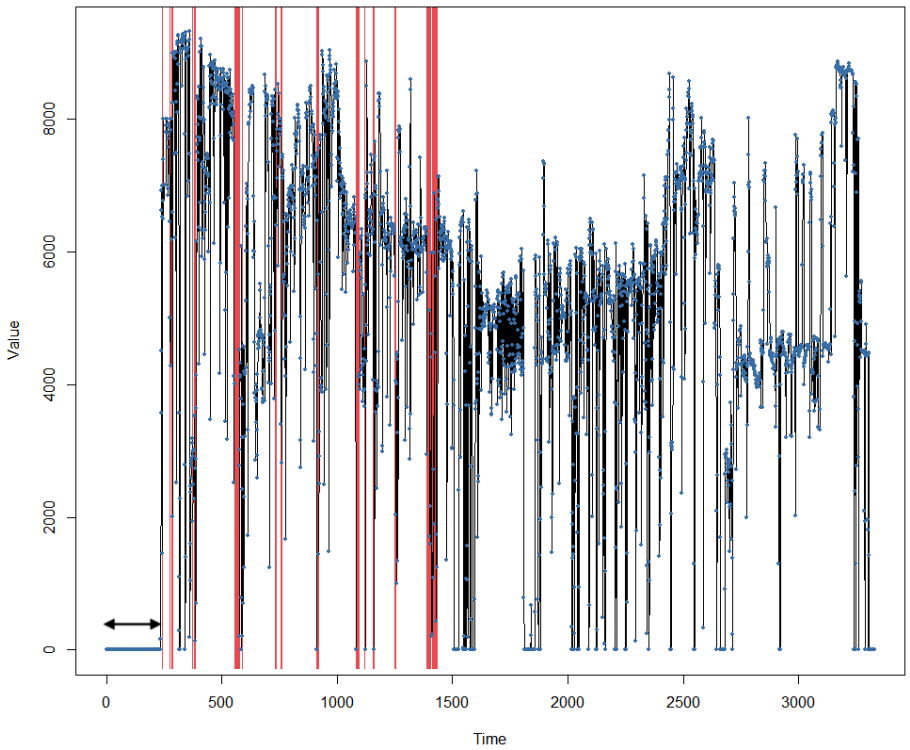
Data Cleansing - Missing Data Imputation

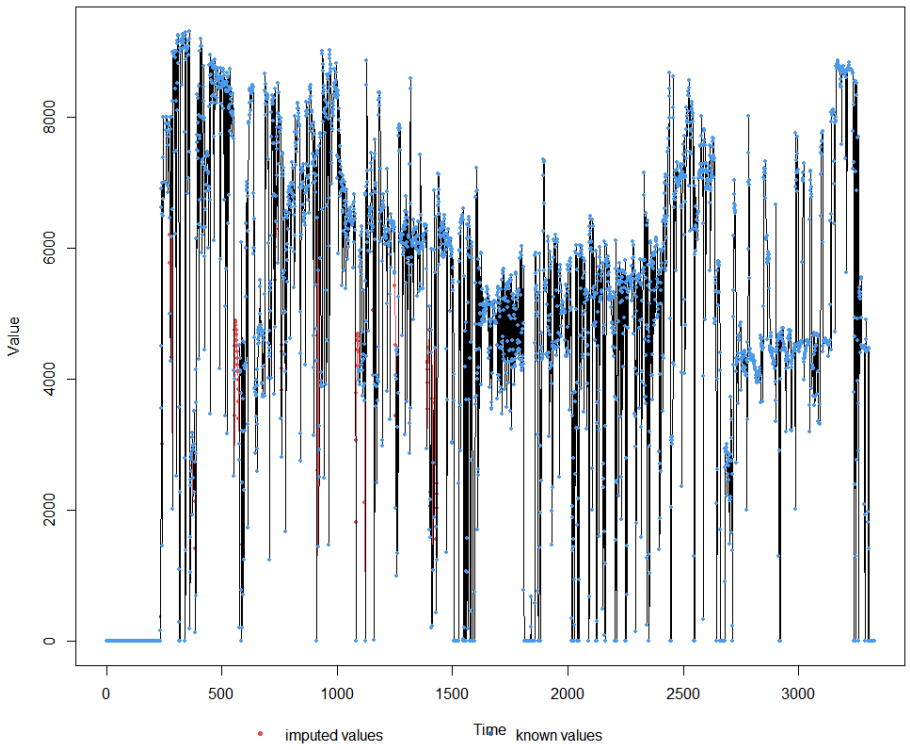
Missing data values which need to be imputed:

1. Water injection volume for well 5AHI and 4AHI using Kalman Filter
2. Avg. ΔP tubing for well 14H using SVR and MLP model
3. Avg. downhole pressure and temperature for well 12H using SVR and MLP model
4. Avg. annulus pressure for well 1C and 14H using SVR and MLP model

Data Cleansing - Missing Data Imputation

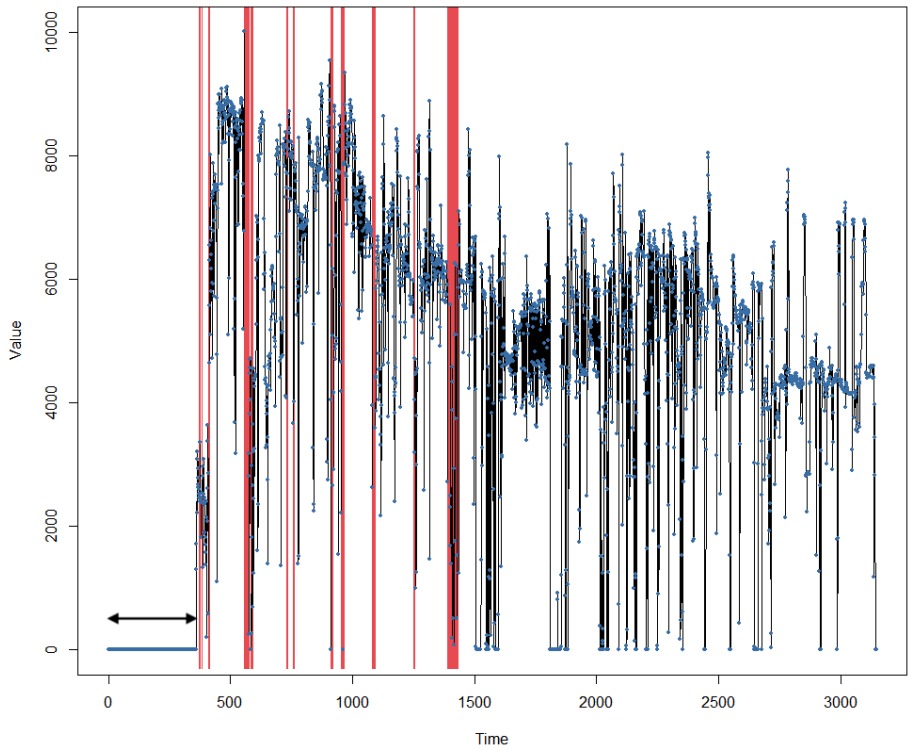
Imputation of water injection missing values of well 4AH using Kalman Filter:

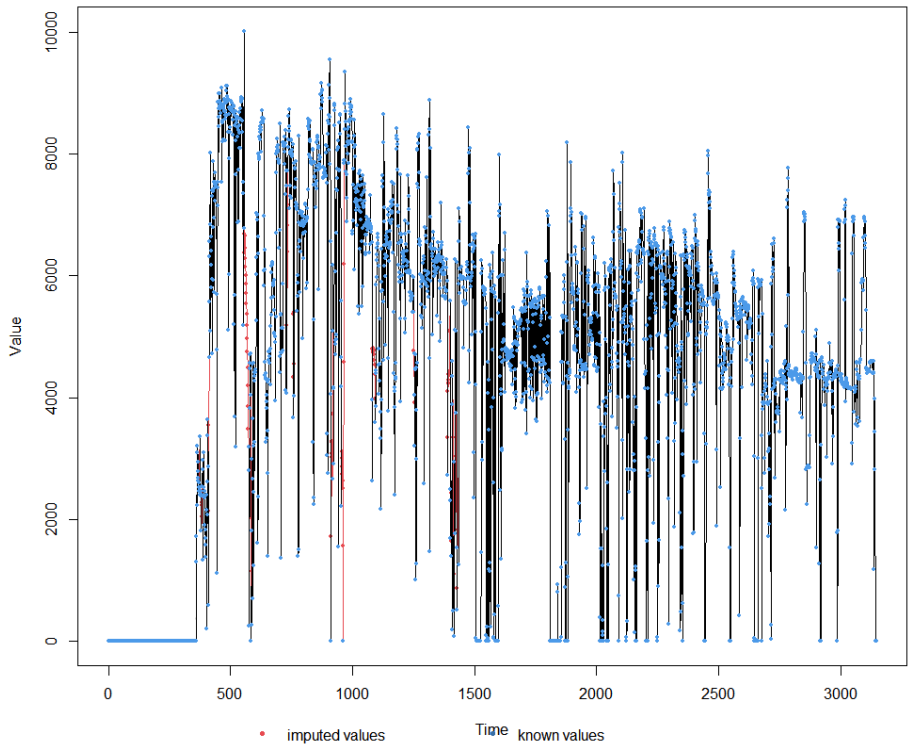




Data Cleansing - Missing Data Imputation

Imputation of water injection missing values of well 5AHI using Kalman Filter:





Data Cleansing - Missing Data Imputation

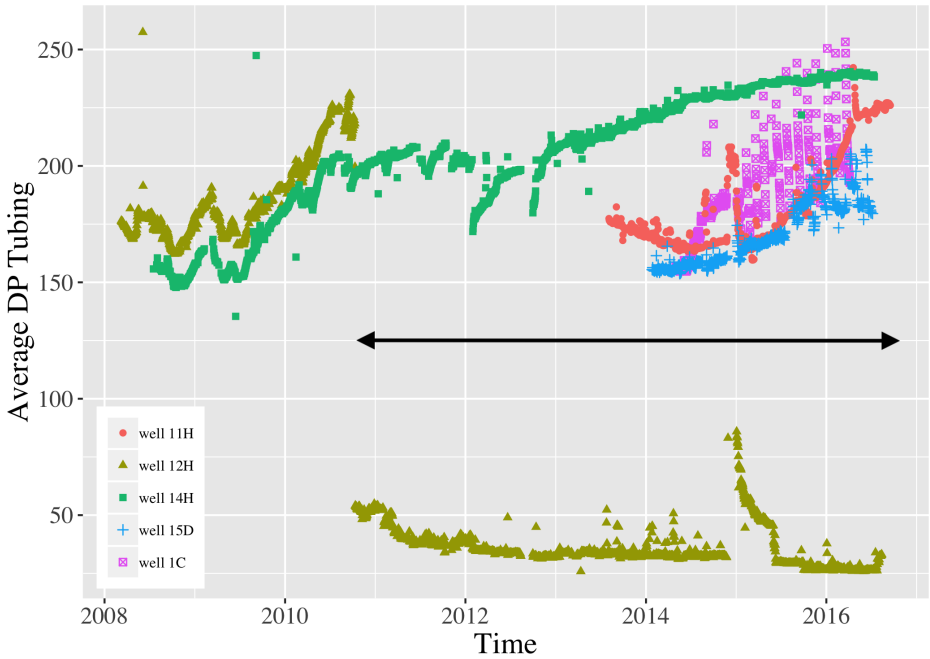
Results of imputation using SVR and MLP:

- Order of imputation
- Using all available non-missing features/data from all wells
- Splitting the dataset into two training (80%) and test (20%) sets
- Cross-validation with 10 fold in training set
- Grid search for *cost* and *sigma* values (from 4 up to 7 values for each)

Imputed feature	Model	R-squared	RMSE	Sigma	Cost	Layer & Neurons	Epochs
Avg. ΔP tubing	SVR	0.90	8.00	3	8	—	—
Avg. ΔP tubing	MLP	0.91	0.308	—	—	[32, 16]	100
Avg. downhole pressure	SVR	0.89	7.44	5	6	—	—
Avg. downhole temperature	SVR	0.94	0.63	5	3	—	—
Avg. downhole pressure	MLP	0.86	0.37	—	—	[32, 16]	100
Avg. downhole temperature	MLP	0.95	0.23	—	—	[32, 16]	100
Avg. annulus pressure	SVR	0.77	2.30	5	20	—	—
Avg. annulus temperature	MLP	0.74	0.50	—	—	[64, 32]	200

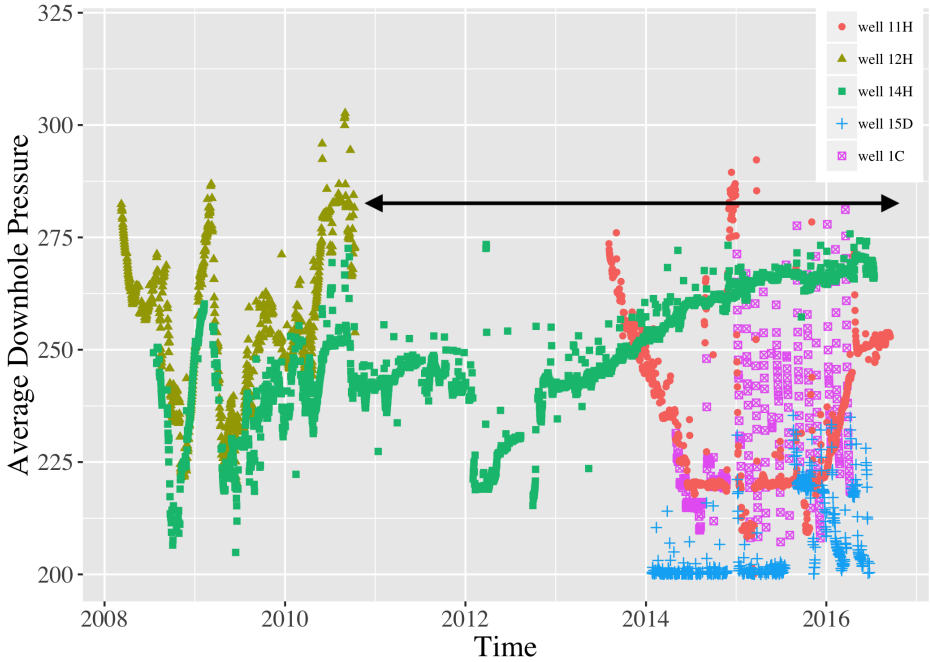
Data Cleansing - Missing Data Imputation

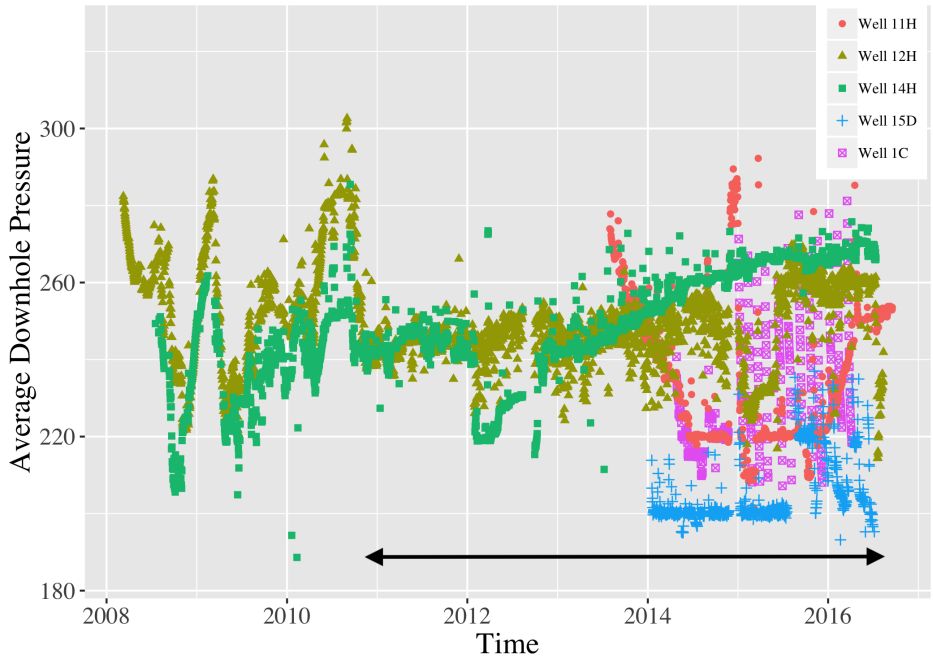
Some visual inspections on features before and after missing data imputation



Data Cleansing - Missing Data Imputation

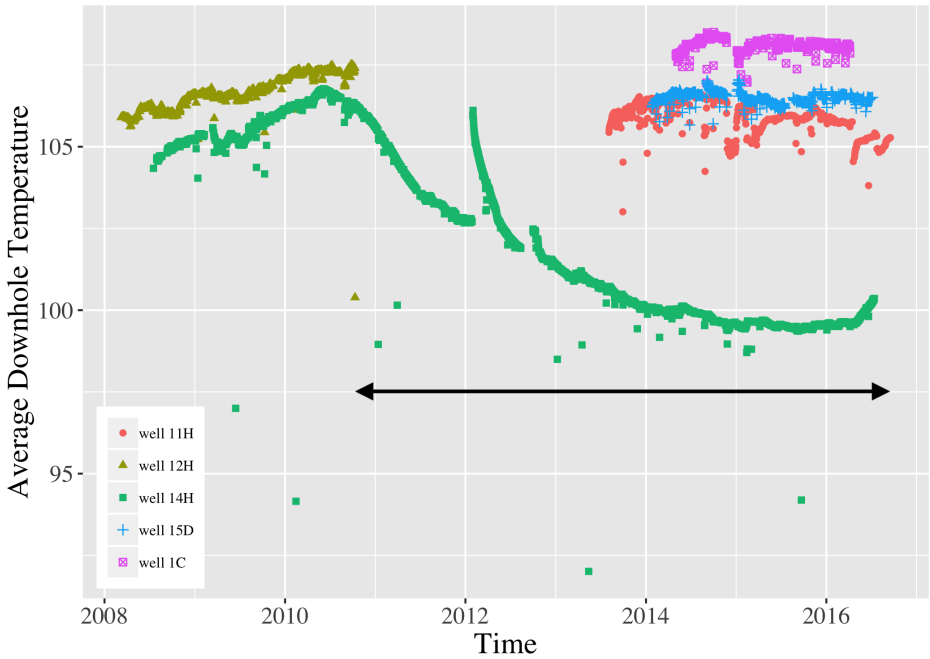
Some visual inspections on features before and after missing data imputation

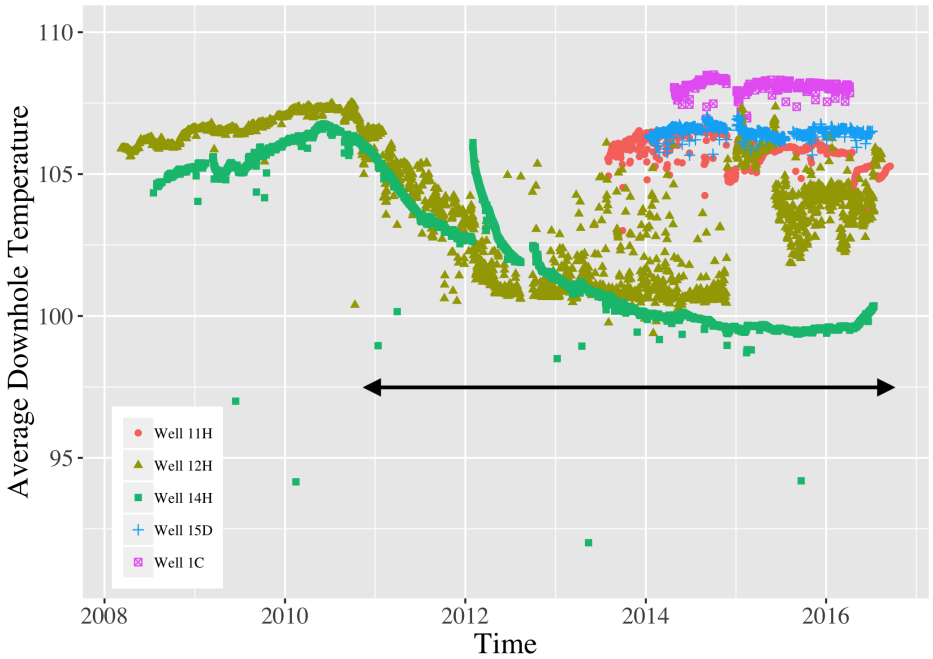




Data Cleansing - Missing Data Imputation

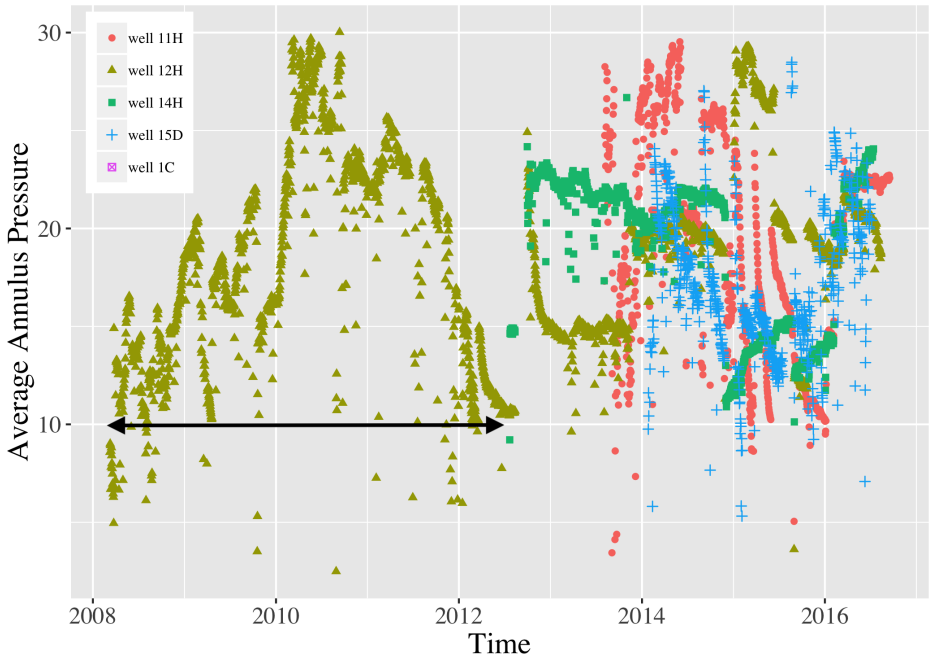
Some visual inspections on features before and after missing data imputation

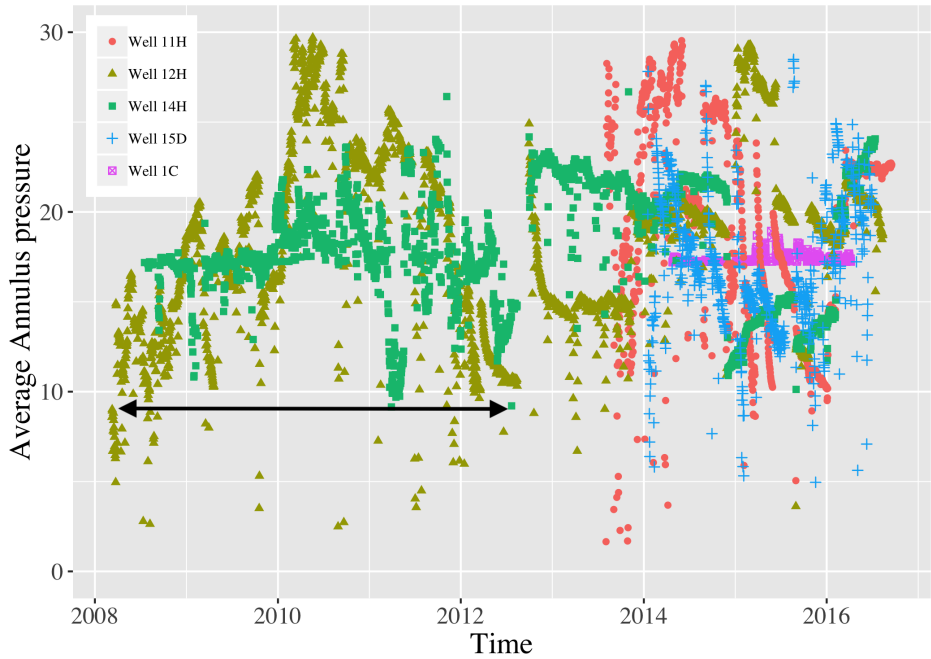




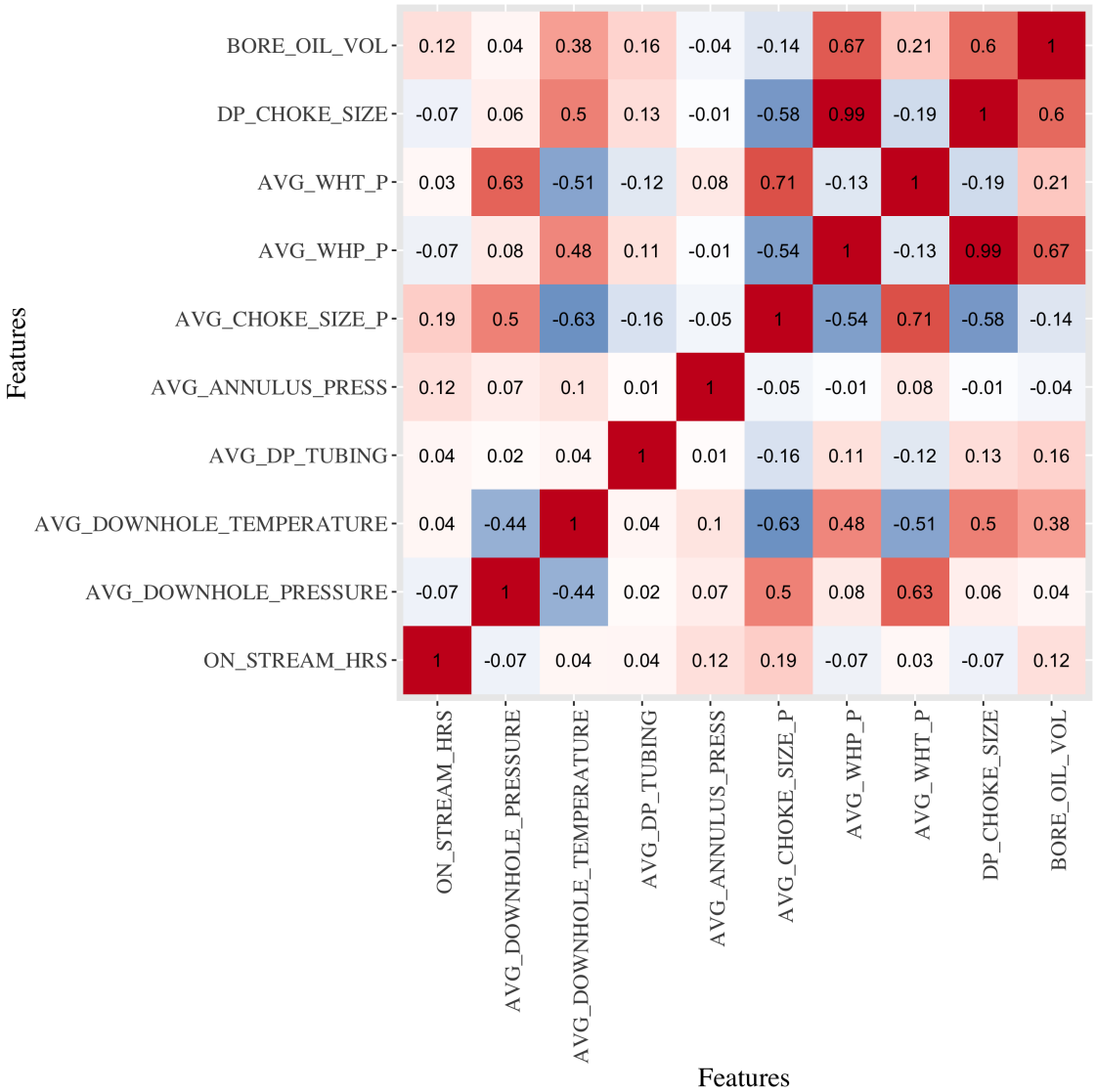
Data Cleansing - Missing Data Imputation

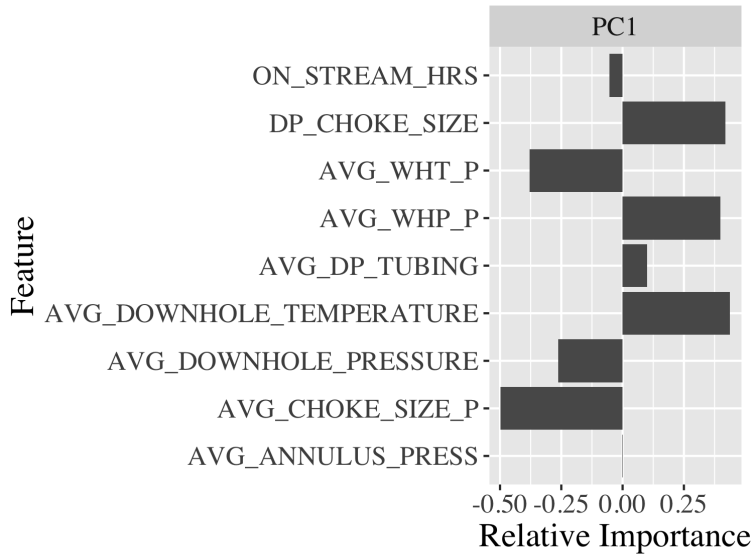
Some visual inspections on features before and after missing data imputation





Data Cleansing - PCA and Correlation





Preliminary Results

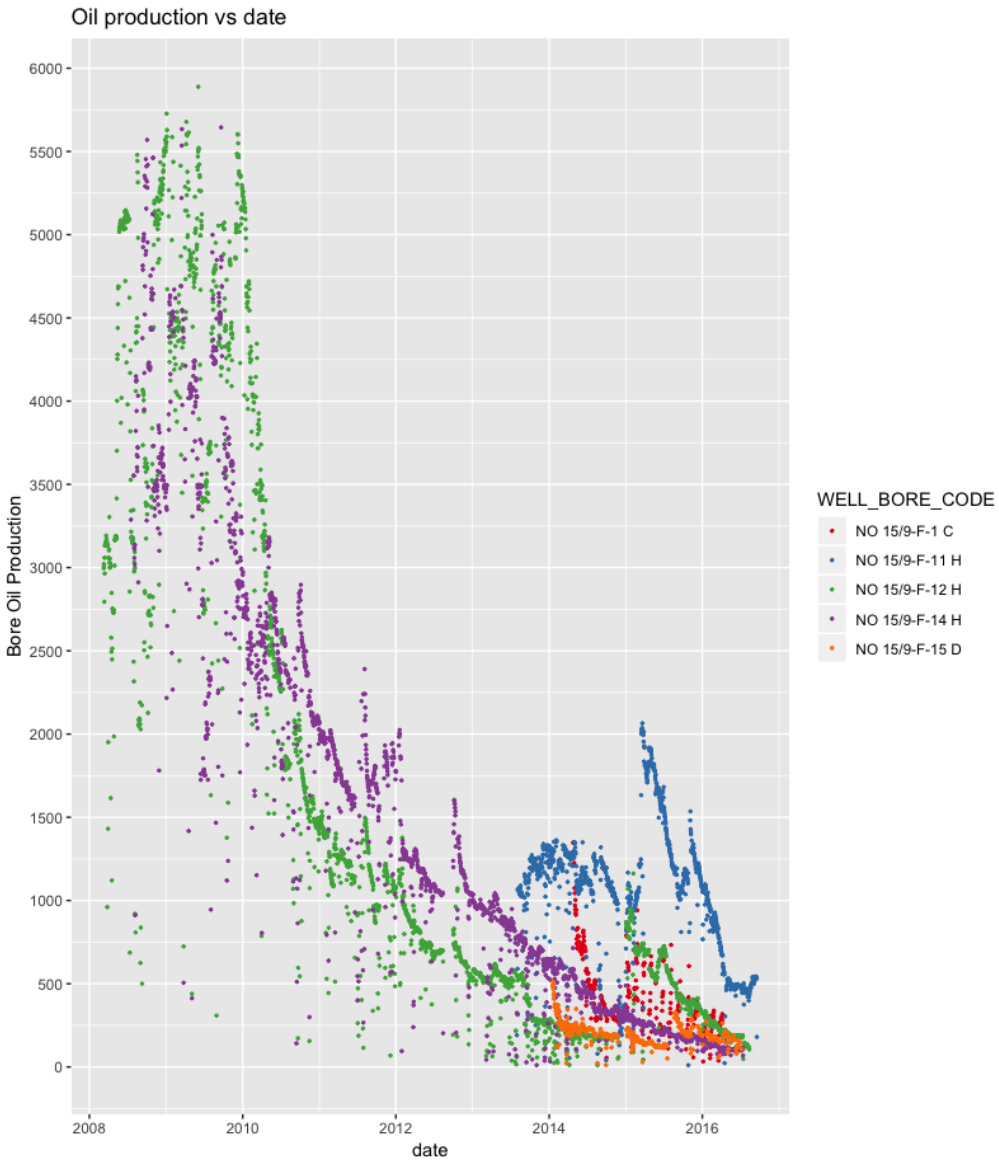
- Using cleaned data
- Using all the features except water injection values
- Without considering history as a feature

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

Time-Series Prediction

- Using cleaned data
- Using all the features as well as water injection values
- Considering history as a feature and constructing time-series data
- Two approaches for normalizing the datasets
- Testing the models on different combinations of datasets

Time-Series Prediction - normalization approaches



Oil production range

- 1 C: 480 $\sim$ 2100
- 11 H: 250 $\sim$ 2100
- 12 H: 0 $\sim$ 6000
- 14 H: 0 $\sim$ 6000
- 15 D: 250 $\sim$ 1200

Time-Series Prediction - normalization approaches

Approach 1 for normalization process

Normalization based on the data from all wells

- Setting the random seed equal to 7
- Constructing the time series data
- Calculating mean and standard deviation of the data
- Normalizing the data based on mean and standard deviation
- Shuffle the time series data
- Splitting first 80% data as the training set and the rest 20% data as the test set

Time-Series Prediction - normalization approaches

Approach 2 for normalization process

Normalization based on the data of each well

- Setting the random seed equal to 7
- **procedure** Recursion(*Well Code*)
- Constructing the time series data
- Calculating mean and standard deviation of the data
- Normalizing the data based on mean and standard deviation
- Shuffle the time series data
- Splitting first 80% data as the training and the rest 20% as the test set
- **end procedure**
- Concatenating the training set and the test set

Time-Series Prediction -SVR

Constructing time series data for SVR model

$$X_{SVR} = [Var_1(t-n), Var_2(t-n), \cdots, Var_m(t-n), label(t-n), \cdots, Var_1(t-1), Var_2(t-1), \cdots, Var_m(t-1), label(t-1), \cdots, Var_1(t)Var_2(t)\cdots Var_m(t)], \quad (1)$$

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:

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

Results- SVR

Average Results of time series oil production prediction using SVR
Grid search on different sigma and cost values
Training and testing the model using different time-steps; 1,3,5,10 and 20

dataset	normalization	R-squared	RMSE	time-step	cost	sigma
{1C}	Each	0.93	0.25	5	100	0.0001
{11H}	Each	0.86	0.34	5	100	0.001
{15D}	Each	0.95	0.21	5	100	0.0005
{12H, 14H}	All	0.99	0.08	5	16	0.001
{1C, 11H}	All	0.97	0.18	5	64	0.004
{1C, 11H}	Each	0.93	0.25	5	64	0.004
{1C, 11H, 15D}	All	0.98	0.12	5	100	0.001
{1C, 11H, 15D}	Each	0.92	0.29	5	100	0.001
{12H, 14H, 1C, 11H, 15D}	All	0.99	0.08	5	100	0.001
{12H, 14H, 1C, 11H, 15D}	Each	0.97	0.18	5	100	0.001

Time-Series Prediction - LSTM

Constructing time series data for LSTM model

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

and Output Y_{LSTM} is:

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

Results - LSTM

Average Results of time series oil production prediction using LSTM
Training and testing the model using different time-steps; 1,3,5,10 and 20
Best time-step: **10 days**
Number of neurons: 50
Training epochs: 100

dataset	normalization	R-squared	RMSE
{1C}	Each	0.91	0.26
{11H}	Each	0.96	0.2
{15D}	Each	0.84	0.36
{12H, 14H}	All	0.99	0.08
{1C, 11H}	All	0.98	0.122
{1C, 11H}	Each	0.96	0.2
{1C, 11H, 15D}	All	0.98	0.11
{1C, 11H, 15D}	Each	0.86	0.38
{12H, 14H, 1C, 11H, 15D}	All	0.87	0.33
{12H, 14H, 1C, 11H, 15D}	Each	0.96	0.25

Conclusion

- The best result of LSTM and SVR when using history as a feature is for the cases below:
  - {12H , 14H}: **R-squared=0.99** and **RMSE=0.08**
  - {1C, 11H, 15D}: **R-squared=0.98** and **RMSE** $\cong$ **0.12**
- LSTM and SVR show different results when we train the model on well 1C, 11H and 15D datasets separately
- Unlike LSTM, the SVR model shows a very good result when training the model on all datasets. But, can not be reliable
- ARIMA and the Gated Recurrent Unit (GRU) models will be deployed as the next step

question:
Separating the datasets into two sets {12H, 14H} and {11H, 1C, 15D}?
Or combining the data from all the wells together?

The End