Scale Treatment Optimisation
Preliminary Report

Kevin Fung
Imperial College London
MSc Applied Computational Science & Engineering
kkf18@imperial.ac.uk

5th July 2019

Supervisors

Prof. Olivier Dubrule
Imperial College London
Dep. of Earth Science & Engineering
o.dubrule@imperial.ac.uk

MSc Lukas Mosser
Imperial College London
Dep. of Earth Science & Engineering
lukas.mosser15@imperial.ac.uk

Dr Meindert Dillen
Wintershall Dea GmbH
EOR Projects
meindert.dillen@wintershalldea.com

MSc Peter Kronberger
Wintershall Dea GmbH
Resevour Engineering
peter.kronberger@wintershalldea.com

1 Introduction

1.1 Background

Scale formation is a severe yet common occurrence during oil well production that can significantly impact an oil wells performance in short time. Two types of scaling can typically occur in oil wells, Calcium Chloride (CaCO3) and Barium Sulphate (BaSO4). CaCO3 is formed from calcium rich fluids and is largely influenced by its pressure, temperature, and other ionic concentrations[1]. BaSO4 is formed when sulphate rich water and incompatible reservoir fluids mix[1]. These types of scale buildup cause increasingly restrictive flows which can drastically reduce oil production if left untreated. Furthermore, they can cause corrosive damage and equipment wear within well systems [2]. Because of this, scaling if left untreated becomes a dangerously expensive problem for operators and thus they are required to act quickly.

By tradition, production engineers analyse live data collected by a multitude of sensors across the well system to judge the occurrence of scaling. When a scale buildup is identified, the engineer must determine the appropriate scale treatment and decide when to apply it with respect to its operational costs. This is a reactive approach. As an initiative to improve production efficiency, operators now look to more proactive approaches by scheduling future scale treatments from predictions in scale forming.

1.2 Motivation

The motivation behind this project is driven by this initiative. Given the recent rise in applicable machine learning, the project will focus on developing a model to predict scaling in the future whilst identifying when to apply scale treatment. The project is in collaboration with Wintershall Dea, who will provide the backlog of well sensor and modelling data spanning back to 2010/16 from the Brage oil field (North Sea). Due to large amounts of data for different wells, the project shall focus on predicting the buildup of CaCO3 scaling in the A07 well (refer to Section 4). In addition, a usable and sustainable piece of software for visualisation shall be developed for engineers to use.

2 Project Proposal and Objectives

From discussions with production engineers, it was identified that the key feature which engineers look for in detecting scale are the differences in well head and down hole pressures. When scale builds up over time the well diameter decreases thus pressure in the fluid increases and so pressure provides a somewhat direct indicator for engineers. The uncertainty in using this data is that scale build up may not be the only influence on pressure readings. Manual interventions, reservoir pressure fluctuations, well fluid injections, and sand clogging can affect the pressure also. From Wintershall Dea engineers experience, scale build up is known to be one of the largest influences in pressure change.

The project proposal is to predict this pressure change then to investigate the use of unsupervised pattern recognition techniques to detect the underlying influence of scale build-up. The desired outcome would be a prediction of the implicit scale buildup in terms of pressure, which would better inform engineers whether scale treatment is required and help optimise the decision process.

A top-level view of project objectives is outlined and can be split into four parts:

1. Data Engineering
   (a) Clean and preprocess data
   (b) Set up pipeline with Azure Databricks
2. Pressure Prediction Modelling
   (a) Feature analysis
   (b) Time series modelling of pressure
3. Scale Build-Up Investigation
   (a) Apply time series signal decomposition methods
4. GUI for Visualisation
   (a) Simple dashboard for graph plots

Discussions with Wintershall Dea engineers highlight that being able to predict pressure will already be a project success. This is because they would be able to use it as an assistive model in monitoring pressure and thus make a better-informed decision on scaling. Nevertheless, the project will attempt

to achieve the stated objectives in the given time frame, although Wintershall Dea expectations will be kept in mind.

3 Relevant Research

3.1 Previous works

One consultancy has recently contributed to this problem by predicting the oil and liquid rate of each well using multivariate linear regression on well test separation data. The intuition is to predict the decrease in flow rate then to pair these trends with the timings of treatments previously applied. This data includes pressure, temperature, and ionic data and is collected every few weeks to determine the theoretical liquid, gas and oil rates for oil wells which engineers use as a benchmark. The predictive results produced for each well was mixed, with relative RMS errors ranging from 8% to 12%. The issue with the approach taken was that well test data was limited (number of test samples per well ranged from 30 - 266) and thus 10-fold validation was applied to try alleviate under fitting. The approach taken is questionable as it trains on data which has been influenced by multiple sources and thus cannot identify scaling with definite certainty. In addition, although the timing of scale treatment was predicted, the models were essentially biased to when the engineer decided to apply scale treatment.

3.2 Supervised Models for Predicting Pressure

In literature, there has been a growing number of machine learning papers in predicting oil well fluid properties. Tian and Horne 2015, was successfully able to predict basic pressure readings using only flow rate and temperature data [3]. Interestingly, they used the Kernel Ridge Regression (KRR) method for modelling. KRR is essentially linear regression but involves the use of a filter to train the regression on a higher dimensional representation of the selected features. The benefits of using linear regression shine when large datasets are provided, a simpler method means better computational efficiency, and they argue that the good accuracy in predictions makes this a viable choice [3]. On the other hand, it is noted that their findings are based on simple experimental cases, which may not upscale to more complex pressure signals encountered in real life.

Onwuchekwa 2018, applied a number of popular machine learning techniques to predict a variety of reservoir fluid parameters with a majority of good accuracies [4]. However, an attempt at accurately predicting the PVT gas data was not successful, and the lack of gas data samples were highlighted. It is not known whether k-fold validation or any feature augmentation methods were applied. One highlight of the paper was the use of collaborative filtering to estimate missing data. The method provides an estimation to an oil wells missing data derived by looking at other oil wells of similar features [4].

In addition, the application of neural networks to predict oil well properties have been a popular research topic. Cao et al 2016, were able to predict the production rate of oil wells, beating standard industry forecasting models [5]. They used a simple neural network architecture with four input nodes, one hidden layer with five nodes, and one output. The objective loss function was the squared error loss with L2 regularisation. To prevent overfit early stopping was applied as well, increasing generality of the model. No quantifiable performance metric was given in the paper, but graphs show predictions

to map the trend better than other models [5].

3.2.1 Time Series Modelling

There is little literature in time series forecasting for oil well parameter predictions and so it becomes an attractive area to explore given our time series data. The multivariate autoregressive model (VAR) is one of the most popular methods for time series forecasting. The method is an extension of autoregression (AR). AR is simply a linear regression whose features are the previous results in the time series with a bias and white noise constant [6]:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \epsilon_t$$

Where β_i are our model parameters, ϵ_t is white noise, and Y_t is our time series data. VAR models regress a vector of time series data instead, where all lagged features are considered in the summation. The main hyperparameter is the lag window p, too much lag would overfit the model whilst too little would result in an estimation bias (underfit), Two approaches can be used to determine the lag window, the F-test approach or an information criterion [6].

Long Short Term Memory (LSTM) networks are another identified method suitable for predicting pressure based on Recurrent Neural Networks (RNN). RNNs uses an internal memory component updated from it's previous states to process incoming sequence data. A disadvantage of this is that it can only capture the short term dependencies in the data. LSTM networks builds upon RNN models to capture the short and long term dependencies. LSTMs are constructed from multiple units of RNN, each corresponding to a respective lagged data point, and each passing information onto the next unit. The RNN unit consists of a Forget gate layer, Store gate layer, input and output layers. The Forget gate layer chooses what information to drop through adaptive weighting and the Store gate layer decides what information to store [7].

3.3 Unsupervised Methods for Detecting Scale Build Up

In this section, two methods of interest was researched. Singular Spectrum Analysis (SSA) is the first method of interest. The time series data is interpreted as a signal composed of a trend, noise, and periodicity. The periodicity may contain a set of separate individual periodic components and each periodic component could then correspond to an underlying influence which built up to the original signal [8]. The SSA method can be generalised as:

1. Produce a sequence of multidimensional lagged vectors, called a Trajectory Matrix.
2. Apply singular value decomposition (SVD) to the Trajectory Matrix.
3. Determine elementary matrices from the SVD components.
4. Diagonally average elementary matrices to form the time series components.
5. Group together similar elementary matrices to identify the individual components.

The second method of interest falls in applying independent component analysis to separate a multivariate signal into maximally independent subcomponents [9]. This may be applicable to the pressure time series data which has been affected by operational influences, and could be interpreted as independent signals building on top of the natural oil well pressure.

4 Data Overview

A general overview of the data is explained in this section. Wintershall Dea provided six different well data sets in the Brage oilfield, however some wells had missing features which rendered them unusable for the project. A table organising the available data is illustrated in figure 1 (Appendix A).

The A07 and A01 well data were the most approachable for applying machine learning. Discussions with engineers led to A07 being chosen as it was known to currently and historically have CaCO3 scaling, as well as having no known evidence of BaSO4 scaling. We choose to not focus predicting BaSO4 scale as it is already difficult for engineers and sensors to detect, and the data available was deemed insufficient to successfully apply machine learning on.

In addition, simulated engineering data is provided, and can be used as a second performance metric during predictive modelling.

5 Proposed Approach

5.1 Project Management

To keep to time and ensure project objectives are met, a Gantt Chart has been produced (see Appendix C). The largest work impacts will be spent on modelling time series data and investigating detection of the underlying scale build up. Thus, two weeks are allocated to each task. A contingency allowance of one week is given if original deadlines are not achieved.

A schedule of meetings twice a week with the project supervisor has been arranged to allow regular feedback across the project period. This is to also ensure sub tasks have been met daily and facilitates any appropriate adjustments to the original plan.

5.2 Achieving Milestones

The milestones in this project align with the main tasks in the Gantt chart, these are:

1. Planning
2. Data Engineering
3. Time Series Modelling
4. Time Series Pattern Recognition

5. GUI
6. Report Writing

5.2.1 Planning

This is currently being achieved through actively speaking to production and reservoir engineers, and extensive literature review in the field of machine learning. In addition, time has been taken out to learn how to use external libraries and video lectures on time series analysis. Practice examples include using Facebook prophet, and Kaggle notebooks. A Gantt chart has been produced and regular meetings scheduled in advance.

5.2.2 Data Engineering

The data is accessible via the company's Azure Data Lake. Azure Databricks will be used for coding and visualisation of data, as the advantage lies in the use of multiple processors for machine learning. Spark library will be used with other standard python libraries for its ability to parallelize the management of data during calculations, hence improving computational speed. Spark and Azure Databricks workshops have been taken to improve knowledge in this area, and practice with it is ongoing.

In terms of data, engineers have highlighted the uncertainty in data which stem from noise from equipment or natural phenomena, missing value, and zero value results. During the week period beginning 24th June, cleaning of the data begun. To visualize certain data better, thresholding and rolling averages are applied to see trends.

The challenge in the milestone lies in cleaning data without risking changing too much of the potential information in the data. To overcome this, regular discussions with production engineers in preprocessing data are scheduled.

5.2.3 Time Series Modelling

A standard data science approach in modelling is taken (see Appendix B). The main challenge will be to find a high performing model within the given time frame regardless of contingency given. Therefore, to maximise time efficiency, a smaller test set shall be used for training, validation, and test sets. For each new modelling method investigated, the contribution of performance improvement by data preprocessing, modelling tuning, and hyperparameter tuning will be compared. This will help decide where time and effort should be spent most on out of these pipeline processes.

Based from research, increasing levels of complexity of predictive models based on the research will be explored, beginning with multivariate Linear Regression with kernelling. Time series methods will be explored as well, however, the data will need to be tested for stationarity which will be a challenge (ADF test, KPSS test). If the data is not stationary, then techniques like differencing and log transformations will need to be applied.

On top of standard python libraries (Numpy, Pandas, Matplotlib). For low level modelling of NN and LSTMS, Pytorch is suitable and will be used. Linear regression methods are available in the scikit

learn libraries, and autoregressive models can be taken from the Stats model library. More blackbox method libraries like FB Prophet and Keras are considered if time is an issue.

Finally, In terms of performance assessment, the speed and accuracy of methods will be compared.

5.2.4 Time Series Pattern Recognition

The success criteria defined for this will be to decompose the pressure signal out and identify major influences on it. As it is an exploration task, a more fluid structure of work will be followed. Two time series sets, one of actual sensor data and the other predicted data, will be prepared for this then signal separation techniques are applied and analysed. A challenge in this milestone lies in the potential of data being used for pattern recognition. To overcome this, further data preprocessing will be applied, as well as using the simulated engineering data for separation.

In terms of coding, SSA can be produced with Numpy methods. Scikit learn has a method called Fast ICA which could be one feasable choice.

5.2.5 GUI

A basic GUI showing animations of the data and predictions will be produced. Simple unit tests will be set up to ensure its adaptability and sustainability when certain functionalities are integrated into the application. If progress moves ahead of schedule, further functionalities may be added to the GUI. Code documentation will be provided.

5.2.6 Report Writing

A first draft report shall begin on the 8th July, where the pipeline for modelling is set up. As the project progresses, the first draft shall continually be updated and adapted up to the Time Series Pattern Recognition deadline. After this, a week will be taken to solely produce a second draft if no contingency has been used. Meetings with project supervisors will give feedback on the report at the end of the week. The final report will then be produced a week after this feedback is received and finally submitted along with the code.

References

[1] M. C. . D. E. . P. F. . M. M. . A. J. . G. King, "Fighting scale - removal and prevention," *Oilfield Review*, pp. 30–45, 1999.

[2] E. Otumudia and A. Ujile, "Determining the rates for scale formation in oil wells," *Int. Journal of Engineering Research and Application*, vol. 6, pp. 2248–9622, 09 2016.

[3] C. Tian and R. Horne, "Applying machine learning techniques to interpret flow rate, pressure and temperature data from permanent downhole gauges," in *Application of Machine Learning Ideas to Reservoir Fluid Properties Estimation*, 04 2015.

[4] C. Onwuchekwa, "Application of machine learning ideas to reservoir fluid properties estimation," in *Application of Machine Learning Ideas to Reservoir Fluid Properties Estimation*, 01 2018.

[5] Q. Cao, R. Banerjee, S. Gupta, J. Li, W. Zhou, and B. Jeyachandra, "Data driven production forecasting using machine learning," in *Data Driven Production Forecasting Using Machine Learning*, 01 2016.

[6] A. G. Christoph Hanck, Martin Arnold and M. Schmelzer, *Introduction to Econometrics with R*. Bookdown, 2018.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[8] H. Hassani, "Singular spectrum analysis: Methodology and comparison," *University Library of Munich, Germany, MPRA Paper*, vol. 5, 01 2007.

[9] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks : the official journal of the International Neural Network Society*, vol. 13 4-5, pp. 411–30, 2000.

Appendices

A Well Table

		Well (*Samples)					
Data	Format	A07	A01	A04	A17	A31	A37
Well Sensors	CSV	X	X	X	X	X	X
Test Separator	CSV	X (117)*	X (209)*	X (266)*	X (30)*	X (32)*	X (36)*
Caliper Test	LAS	X	X	-	X	X	X
Scale Treatment Records	Excel	X (1)*	X (4)*	-	-	-	-
Ion Composition	Excel	X	X	X	-	-	-
Daily production, WHP, Gas lift rate, DHP	Excel	X	X	X	-	-	-

Figure 1: General overview of provided data

B Flow Chart

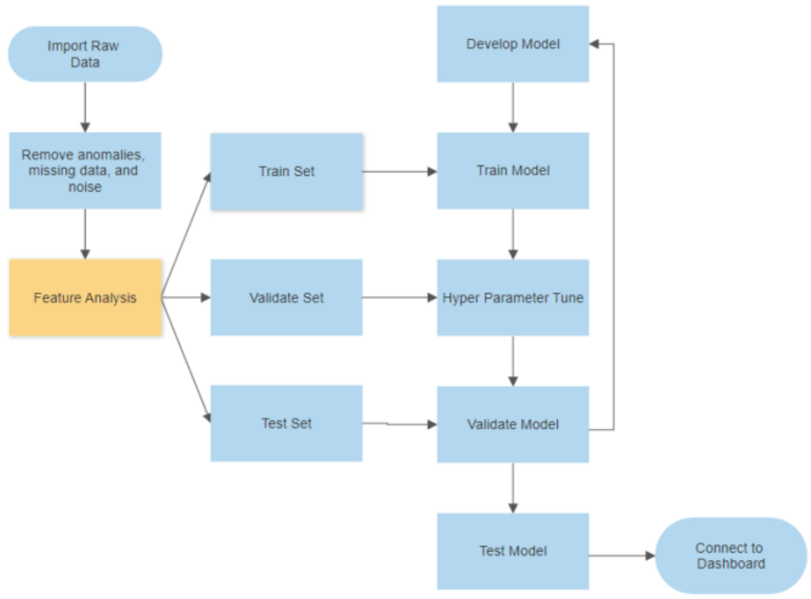

Figure 2: Predictive modelling pipeline flowchart

C Gantt Chart

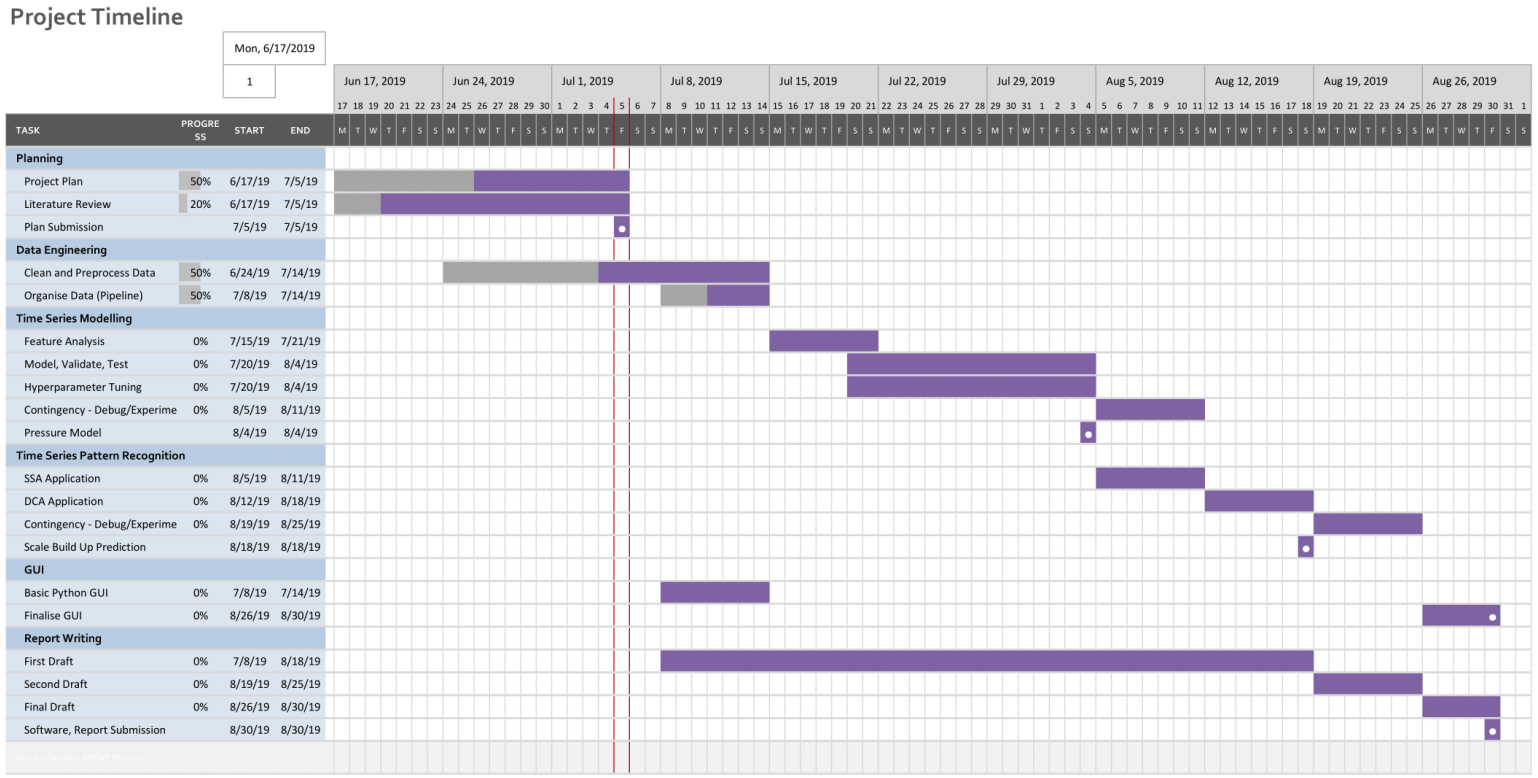

Figure 3: Project Gantt chart