Metin Türkay Metin Trkay, Erdal Aydin (Eds.)
Proceedings of the 31th European Symposium on Computer Aided Process Engineering
June 6th to 9th, 2021, Istanbul, Turkey.

Data-driven modelling of choke valve erosion using data simulated from a first principles model

Jan Henrik Jahren[a], Jose Matias[a] and Johannes Jäschke[a*]

[a]*Department of Chemical Engineering, Norwegian University of Science and technology, Sem Sælands vei 4, 7034 Trondheim, Norway*
johannes.jaschke@ntnu.no

Abstract

Maintenance of subsea operating equipment is a significant part of the operational costs of running an oil production system. For instance, on the Norwegian continental shelf, it amounts to 60 billion NOK in operational expenses (DNV-GL (2015)). One of the principal mechanisms of degradation in subsea process equipment is erosion by sand, which is a very complex process and, thus, difficult to model using physical domain knowledge. Because of this difficulty, we propose in this paper the use of data-driven approaches for modelling erosion in critical equipment of a subsea oil production rig. In such systems, a multitude of available process measurements such as flowrates, gas lift injection rates, pressures etc. can be combined to a soft-sensor for component degradation. This approach could save significant amounts of resources by allowing fewer cost intensive inspections and monitoring schemes. A soft-sensor approach was tested with simulated data of the sand degradation of a choke valve in a gas lifted oil production system with three wells. In this paper, we present results from soft-sensor methods like multiple linear regression, regression trees, ensembles methods and kernel methods. The approaches were tested and compared in two case studies, the first with constant sand outflow from the reservoir for initial exploration of the data driven approaches, and the second with a more realistic profile in which the sand rate is increasing. In both cases, we compare the soft sensor modelling techniques in terms of their basic requirements of accuracy as well as transparency and interpretability.

Keywords: equipment degradation, data-driven modelling, machine learning

1. Introduction

Failure to detect faults in large scale, expensive or critical equipment can have immense consequences. Both financial and in the most extreme cases, loss of life. One of the main mechanisms in equipment degradation in subsea oil extraction, is sand erosion. Accurately modelling this process is vital for monitoring of equipment health (Hansen (2016),Si et al. (2012)). Erosion by sand is a very complex process and, thus, difficult to model using physical domain knowledge. Because of this difficulty, we propose in this paper the use of data-driven approaches for modelling erosion in critical equipment of a subsea oil production rig. In such systems, a multitude of available process measurements, such as flowrates and pressures, can be combined to a soft-sensor for component degradation. This approach could save significant amounts of resources by allowing fewer cost intensive inspections and monitoring schemes as well as improving safety. To investigate

the usability of such a data-driven modelling approach models will be tested on simulated data using the subsea gas lifted oil well network model proposed by Krishnamoorthy et al. (2016) and adapted by Verheyleweghen and Jäschke (2018).

2. Gas lifted oil well network

In some cases oil wells do not have sufficient reservoir pressure to lift fluids to the topside facility. Gas lift, the injection of compressed gas through the annulus, can be used to overcome this. The annulus is the void between the well piping and the casing. This leads to a reduction in the fluid mixture density. That, in turn reduces the hydrostatic pressure loss in the well. Consequently, the pressure at the well bottom decreases, compensating for the low reservoir pressure.

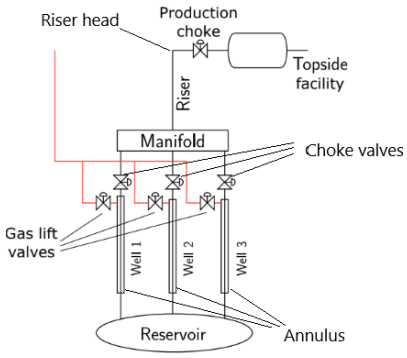

Figure 1: Illustration of a gas lifted oil production system with three wells showing how production goes from reservoir to topside facility (Verheyleweghen and Jäschke (2018)).

In Figure 1 the different parts of a gas lifted oil well network are shown. The oil flows, from a single reservoir into three wells, then, into a manifold before the riser takes the fluid mixture to the surface. Maturing fields experience a significant increase over time in sand production, which can be approximated by an exponential function (Hettema et al. (2006)).

3. Methods

In this section we give an overview of the simulations of the oil well network which were used to generate the data. The pre-processing step that was applied to facilitate modelling of the data. Next, we introduce the statistical learning methods used for predicting the erosion rate.

3.1. Simulations

A small additional adaption to the model was made to incorporate a varying sand production rate (SPR). For each case study datasets of 200 time series of 500 days were simulated containing the total erosion length of the wells choke valves. The gas lift rate was varied randomly every 50 days within a given range:

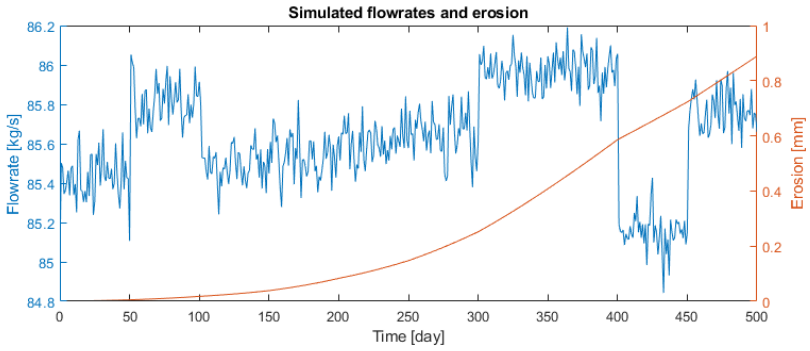

Figure 2: Illustrative plot of cumulative erosion [mm] (right axis) and flowrate [kg/s] (left axis) from one 500 day time series.

Constant sand production rate yielding a greatly simplified model, but a good starting point for initial exploration of the data-driven methods. This is equivalent to: $m_{sand}(t) = m_{sand}(0)$.

Exponential sand production rate yielding a more realistic model of a field increasing sand production as it matures. An exponential function, $m_{sand}(t) = m_{sand}(0)e^{(0.005t)}$ was chosen to emulate a qualitative description of the sand production rate in a field over time (Hettema et al. (2006)). Figure 2 shows the cumulative erosion of the choke valve for one simulated time series in addition to the total well production flowrate. Note that the choke valve erosion is a function of the well production flowrate, as expected, but the dominating force is the sand production flowrate. The more sand produced by the well, the faster the choke valves erode.

Process variables were measured every time step (day), with added noise to process measurements. Erosion, gas lift injection rate and sand production rate were recorded directly without noise. The pressure is measured in the annulus, well head (top of each well), riser head (top of the riser) and in the manifold connecting the wells. The production rates of gas and oil in the top of the riser as well as the top of each individual well are also measured. In addition to the aforementioned process measurements, the sand production rate of the reservoir and the gas lift injection rate (control input of the optimisation) are used to simulate the erosion in the choke valves atop each well. The simulated data is then split in two parts of equal size, a training set and an independent test set. Due to the large availability of data for validation, since simulated data is used, we choose to use holdout validation, holding out 30% of the training set for model selection and validation.

3.2. Pre-processing

The gradient of the simulated cumulative erosion is used as response variable (i.e. the erosion rate). Predictions at each time step, are made using the corresponding process variables for that time step. Sand production rate is usually not a constantly monitored variable, as such it was supplied to the models with a realistic sampling interval of 50 days, but the impact of lowering this to 30 days is also studied. Models were tested on normalized data with unit variance and mean zero. Tests were also done using PC scores

as predictors, in this case PC's explaining $> 95\%$ variance.

3.3. Data-driven models

For obtaining the soft-sensor for erosion degradation, multiple traditional statistical learning methods are implemented and compared for the use cases, a brief introduction is given below but the interested reader is referred to Friedman et al. (2017). All models were implemented in Matlab using the Statistics and Machine Learning toolbox (The MathWorks (2020)).

Multiple linear regressions (MLR) are the simplest of all statistical models, where we assume a linear relationship between predictors (or combinations of predictors when including interactions). When such a relationship exists these models yield good interpretability and room for extrapolation. For the linear regression models, least squares loss is used and a stepwise approach (*Stepwise MLR*) is applied for model selection, with iterative addition of predictors (linear terms and interaction terms) to a null model until additional predictors no longer lead to significant reduction in loss.

Ensemble methods essentially aggregate several simple models, in this case regression trees to yield a better prediction than individual models can. In this work boostrap aggregation (*Bagged trees*) and gradient boosting (*Boosted trees*) were used. Ensemble methods can provide very good models of linear and non-linear phenomena but this comes at a significant cost in transparency as interpreting relationships from these methods is very difficult.

Support vector regression (SVR), which is a kernel method adapted from the famous classification algorithm support vector machines, was used. This method uses the SVM algorithm to create a separating hyperplane in the data space which will allow prediction on new data samples. SVM's are particularly suited for high dimensional data and are robust against outliers (Drucker et al. (2003)). The reported results are from *Cubic kernel SVR*, which had the strongest overall performance.

In this work the loss function for parameter optimisation, model selection and model validation is always the mean square error of prediction (MSE). Hyperparameters of the ensemble and kernel methods were tuned using bayesian optimization, using the expected improvement as acquisition function.

4. Results and discussion

Learning methods were tested on both the data simulated from constant sand production rate and exponential sand production rate. The performance of the methods is measured by independent test set mean squared error of prediction (MSE). The performance results are given in Table 1, where the first two columns give results for the first case study and the last for the second case study. There are some considerations outside of model performance to be taken into account when selecting a model for actual applications. There are significant differences in model training time, with the outlier being cubic kernel SVR, requiring significantly more training than bagged trees, boosted trees and stepwise MLR. The other consideration that needs to be made is that to be acceptable in real world scenarios "black box" models of safety critical components are undesirable. Model transparency and interpretability is preferred. None of the models considered here are completely black

Table 1: Table showing the test set performance of the models on simulated data. With columns showing test set MSE's when trained on different data sets: constant sand production rate normalised data, constant sand production rate PC scores, exponential sand production rate normalised data and exponential sand production rate PC scores. The performance on data with a 30 day sampling rate is also shown for comparison to the 50 day sampling rate of the sand production rate.

Method	Const. SPR	Const. w/ PCA	Exp. SPR	Exp. w/ PCA
Stepwise MLR	0.0096	0.0166	0.0182	0.0191
Bagged trees	0.0091	0.0162	0.0184	0.0192
Boosted trees	0.0125	0.0320	0.0184	0.0239
Cubic kernel SVR	0.0120	0.0129	0.0164	0.0169
30 day sampling rate	Const. SPR	Const. w/ PCA	Exp. SPR	Exp. w/ PCA
Stepwise MLR	N/A	N/A	0.00740	0.00946
Bagged trees	N/A	N/A	0.00537	0.0134
SVR	N/A	N/A	0.00687	0.00879

box, in general models with transparent coefficients showing how a prediction is made are easier to interpret. Linear regressions is a class of such methods, allowing easy analysis of coefficients and, consequently its predictions. As such, when performance is close the preferred methods will be MLR models due to fast training times and superior interpretability.

4.1. Constant sand production rate data

All the methods show a small MSE value when tested on unseen data. Since the underlying phenomenon of erosion behaves linearly when sand production rate is held constant as in this test case. Very good fit is, thus, expected. The usefulness of this test case was primarily for initial exploration on a very simplified system. Additionally we note that the performance of all models is degraded when PCA pre-processing is applied. This could be because the overall degrees of freedom afforded to the model is lowered. If multiple variables are mainly represented in one principle component, the interactions between them cannot be properly modelled by the interaction terms in a linear regression model, similar arguments hold for the other methods.

4.2. Exponential sand production rate

For the data simulated with exponential sand production rate, a significant decrease in the performance was observed for all methods, as expected with a more complex phenomenon being emulated. Similarily to the initial test case, there is a drop in performance when PCA is applied, but this effect is relatively weaker for the exponential data. In this case cubic kernel SVR proved to have the strongest performance, but in general Bagged trees, Boosted trees and Stepwise MLR all provided fairly accurate predictions of the erosion rate. The sampling rate of the sand production rate measurement as expected has a very significant impact on the model accuracy with exponential distribution, increasing sampling to once every 30 days instead of 50 for example reduces the MSE of a optimised bagged ensemble from 0.0066 to 0.0184. Similar trends are seen in the other methods as

well, with model performance improving significantly when sampling rate is increased. Such a effect is expected, as the models are working with more accurate data.

5. Conclusion

It is observed that for constant sand production rate, very accurate predictions of the erosion rates are made. Additionally a significant degradation of accuracy is seen for all the constant sand production rate models except for the kernel methods when PCA is used. This could be due to a lower reliance on variable interactions which to some extent is hidden when PCA is applied. This provided a useful initial exploration and foundation for the second case study. On the data simulated from an exponential sand production rate the performance is overall worse, which is expected since the phenomenon that the models are attempting to reproduce is more complex. The methods are still relatively accurate with under 0.02 MSE on normalised unseen test data. However, when the slope of the sand production rate profile gets very steep (i.e time increases) the models suffer quite significantly from the sampling rate. With current model performance there is a strong case to be made for selecting linear regression based methods as they provide superior model transparency and interpretability. Having observed that simulated data can be predicted well using statistical models, further investigation on real world data is merited to ascertain applicability to real industrial facilities.

6. Acknowledgments

This work was carried out as a part of SUBPRO, a Research-based Innovation Centre within Subsea Production and Processing. The authors gratefully acknowledge the financial support from SUBPRO, which is financed by the Research Council of Norway, major industry partners, and NTNU.

References

DNV-GL (2015), 'Recommended practice rp-o501: Managing sand production and erosion.'. **URL:** *https://rules.dnvgl.com/docs/pdf/dnvgl/RP/2015-08/DNVGL-RP-O501.pdf*

Drucker, H., C, C., Kaufman, L., Smola, A. and Vapnik, V. (2003), 'Support vector regression machines', *Advances in Neural Information Processing Systems* **9**.

Friedman, J., Hastie, T. and Tibshirani, R. (2017), *The Elements of statistical learning, Data Mining, Inference, and Prediction*, Springer.

Hansen, S. K. (2016), Modelling failure mechanisms in subsea equipment, Master's thesis, NTNU.

Hettema, M. H., Andrews, J. S., Papamichos, E. and Blaasmo, M. (2006), The relative importance of drawdown and depletion in sanding wells: Predictive models compared with data from the statfjord field, *in* 'Proceedings of the SPE International Symposium and Exhibition on Formation Damage Control, Lafayette, 15 – 17 February 2006', Society of Petroleum Engineers.

Krishnamoorthy, D., Foss, B. and Skogestad, S. (2016), 'Real-time optimization under uncertainty applied to a gas lifted well network', *Processes* **4**, 52.

Si, X., Wang, W., Hu, C., Zhou, D. and Pecht, M. G. (2012), 'Remaining useful life estimation based on a nonlinear diffusion degradation process', *IEEE Transactions on Reliability* **61**(1), 50–67.

The MathWorks, I. (2020), *Statistics and Machine Learning Toolbox*, Natick, Massachusetts, United State. **URL:** *https://www.mathworks.com/help/stats/*

Verheyleweghen, A. and Jäschke, J. (2018), 'Oil production optimization of several wells subject to choke degradation', *IFAC-PapersOnLine* **51**(8), 1–6.