Increase of Oil Production Yield in Shallow-Water Offshore Oil Wells in the Dagang Oilfield via Machine Learning

Tan, Chaodong[1) 2)]; Bangert, Patrick[3) 4) 5)]; Liu, Bailiang[6)]; Zhang, Jie[2)]

[1)] China University of Petroleum, Beijing, China

[2)] Yadan Petroleum Technology Co Ltd, Beijing, China

[3)] algorithmica technologies GmbH, Bremen, Germany

[4)] Advanced International Research Institute on Industrial Optimization gGmbH, Bremen, Germany

[5)] Department of Mathematics, University College London, London, United Kingdom

[6)] PetroChina Dagang Oilfield Co, Tianjin, China

Corresponding author: Patrick Bangert

Abstract

Several shallow-water offshore oil wells are operated in the Dagang oilfield. We demonstrate that it is possible to create a mathematical model of the pumping operation using automated machine learning methods. The basis for the model is the historical data from the data historian. No engineering or human input is required. The resulting differential equations represent the process well enough to be able to make two computations: (1) We may predict the status of the pumps up to four weeks in advance allowing preventative maintenance to be performed and thus availabilities to be increased and (2) we may compute in real-time what set-points should be changed so as to obtain the maximum yield output of the oilfield as a whole considering the numerous interdependencies and boundary conditions that exist. This method requires *no engineering changes* to be made to the rig and requires nominal human effort to implement. We conclude that a *yield increase of approximately 5% is possible* using these methods.

1 Statement of the Problem

The Dagang oilfield lies in the Huanghua depression and is located in Dagang district of Tianjin. Its exploration covers twenty-five districts, cities and counties in Tianjin, Hebei and Shandong, including Dagang exploration area and Yoerdus basin in Xinjiang. The total exploration area in Dagang oilfield is 34,629 km^2, including 18,629 km^2 in the Dagang exploration area. For this paper, we will consider data for 5 oil-wells of a shallow water oil-rig in Dagang operated by PetroChina.

An offshore platform drills several wells into an oilfield and places a pump into each one. If the pressure of the oilfield is too low – as in this case – the platform must inject water into well in order to push out the oil.

Thus, the pump extracts a mixture of oil, water and gas. This is then separated on the platform. External elements like sand and rock pieces in this mixture cause abrasion and damage the equipment. When a pump fails, it must be repaired. Such a maintenance activity requires significantly less time if it can be planned as then the required spare parts and expert personnel can be procured and made available *before* the actual failure. If we wait until the failure happens, the amount of time that the well is out of operation is significantly longer. Thus, we would like to know several weeks in advance when a pump is going to fail.

Each pump can be influenced via two major control variables: the choke-diameter and the frequency of the pump. These parameters are currently controlled manually by the operators. Thus, the maximum possible yield of the rig depends largely on the decisions of the operators, defined by the knowledge and experience of the operator as well as the level of difficulty of any particular pump state. However, the employment of continuous and uniform knowledge and experience for the pump operation is not realistically possible as no one operator controls the plant over the long-term but usually only over a shift. Observation results show oscillations of parameters in a rough eight-hour pattern which supports the argument that a fluctuation in the knowledge and experience of human operators may lead to a fluctuation in the decision making and thus a varying influence on the operation of the plant. While some operators may be better than others, it is often not fully practical and/or possible to extract and structure the experience and knowledge of the best operators in such a fashion as to teach it to the others.

Pumps in an oilfield are not independent. Demanding a great load from one will cause the local pressure field to change and will make less oil available for neighboring pumps. Obtaining a maximum yield output, therefore, is not a simple matter but requires careful balancing of the entire field. In addition, certain external factors also influence the pressure of the oilfield, e.g. the tide. This high degree of complexity of the pump control problem presents an overwhelming challenge to the human mind to handle and the consequence is that suboptimal decisions are made.

In this paper, a novel method is suggested to achieve the best possible, i.e. optimal, yield at any moment in time, taking into account all pumps as well as their complex interconnections. This method yields a computed yield increase in the range of five percent. Moreover, this yield increase is available uniformly over time effectively increasing the base output capability of the rig.

2 Methodology

Sensor equipment is installed in all important parts of the plant and thus alerting the operator via the control system about the current state of the plant. The numerical values of all sensors can be arranged into a vector. Let us assume that we have a total of N measurements on and around the plant that we wish to look at.

We may represent the state of the plant at time t by an N-dimensional vector, $\mathbf{x}^{(t)}$. Via the data historian, we may obtain a set of such vectors for past times. If we order this set with respect to time, then this set is called a time-series, $\mathbf{H} = (\mathbf{x}^{(-h)}, \mathbf{x}^{(1-h)}, \mathbf{x}^{(2-h)}, \dots, \mathbf{x}^{(0)})$ where time $t = 0$ is the current moment and time $t = -h$ is the most distant moment in the past that we wish to look at. Thus the time-series $\mathbf{H}$ is effectively a matrix with h+1 columns and N rows.

Observe that this matrix contains all the decisions of the operators and all the reactions of the plant to these decisions. The knowledge and experience of the operators is thus plainly visible in the data. If the history is long and detailed enough, this information is all one needs to know about this plant in order to model it.

We recall the topic of control theory. Here we are faced with a black box that has input signals and output signals. The process that connects input to output is totally unknown and is represented by the black box. Control theory now aims to discover the relationship between input and output by performing experiments. If we send a particular signal in, then we observe another signal coming out. Given enough such data and some analysis, control theory provides tools for creating a set of (differential-) equations that govern the behavior of the black box. The resulting set of equations is called a mathematical model. A crucial element is the time evolution of the process, i.e. an action at a certain time will have some effects immediately, some effects over a short-term and other effects over the long-term – this time dependency must be contained in the model for realistic results.

Note that the model does not allow us to 'understand' the process inside the black box. But it does allow us to compute the output of the black box given a sample input. Using the results of optimization theory, we can reverse this process and compute the input needed to achieve a given desired output.

Control theory is meant to be applied manually. For a process as complex as that of a power plant, this is impractical due to the amount of work that would be required. It is suggested to use machine learning [1] to develop the set of equations automatically. There are various techniques available to achieve this such as neural networks [2]. We opt for the technique of recurrent neural networks [3]. Here we must differentiate classificatory neural networks [2] from recurrent neural networks [3]. The first can tell the difference between a finite number of types of objects while the second can represent the evolution over time. The necessary mathematical methods that allow recurrent neural networks to be trained efficiently for large datasets coming from real industrial facilities have been invented only in 2005 and thus these methods can only be applied now.

The advantages of using machine learning over a human engineered model are (1) that the model is produced within a very short time (usually days), (2) that it is adaptive (i.e. it learns continuously as it experiences more data), (3) that it can change to match new situations (the new data is learnt) and (4) that the entire prob-

lem can be modeled (and not a simplified version as in the manual approach). Thus, (5) this method is economical.

In the state vector that describes the plant, there are elements of three different types. First, there are measurements that can be directly controlled by the operator. An example is the amount of coal per hour being put into a particular mill. We call these *controllable*, $\mathbf{x}_c^{(t)}$. Second, there are measurements that cannot be controlled at all by the operators and thus represent a state of the world outside the plant. An example is the outside air temperature. We call these *uncontrollable*, $\mathbf{x}_u^{(t)}$. Third, there are measurements that are indirectly controlled via the controllable measurements. An example is a vibration in the turbine. We call these *semi-controllable*, $\mathbf{x}_s^{(t)}$.

Uncontrollable measurements provide boundary conditions for the problem and so we really have a set of models depending on the boundary conditions. This poses no problem for machine learning and is simply included in the model of the black box that is the plant. The only requirement is that it must be clearly defined which measurements belong into which of the three possible groups. Once this is known, the learning may begin.

What we obtain is a function $f(\mathbf{x}_c^{(t)} ; \mathbf{x}_u^{(t)}) = \mathbf{x}_s^{(t)}$. In words, this means that we have a function with the controllable measurements as variables, the uncontrollable measurements as given parameters and the semi-controllable measurements as functional outputs. The plant efficiency is, of course, among the semi-controllable outputs of the function f(…).

With this model and given a particular boundary condition $\mathbf{x}_u^{(t)}$, we may compute the reaction of the plant $\mathbf{x}_s^{(t)}$ to any particular operator decision $\mathbf{x}_c^{(t)}$. This is effectively a plant simulation. Such a system may be used for training and practice of the operators.

More interestingly, we ask whether the function may be inverted, i.e. whether the function $f^{-1}(\mathbf{x}_s^{(t)} ; \mathbf{x}_u^{(t)}) = \mathbf{x}_c^{(t)}$ can be obtained. Generally, it is not possible to invert functions directly. However, we do not require a closed form solution of this problem but only a numerical solution. This may be achieved using the theory of numerical methods [4].

In particular, we are not necessarily interested in general inversion but rather in a very special form of inversion, namely optimization. Given particular boundary conditions, we would wish to know what input variables lead to the optimal state of the plant. The optimum state is defined by some merit function $g(\mathbf{x}_s^{(t)} ; \mathbf{x}_u^{(t)})$. The simplest such merit function is a single measurement point but we may get complex such as the plant efficiency and even take into account market prices and other business features to define what we believe to be the optimum.

Thus we ask, what is $\mathbf{x}_c^{(t)}$ such that $g(\mathbf{x}_s^{(t)} ; \mathbf{x}_u^{(t)})$ achieves a global maximum where the relationship between the variable vector and the merit function is contained in the inverted model $f^{-1}(\mathbf{x}_s^{(t)} ; \mathbf{x}_u^{(t)}) = \mathbf{x}_c^{(t)}$. This is

a classic optimization problem. As the functions are only known numerically and they are highly non-linear and time-dependent, this is a complicated optimization problem requiring state-of-the-art treatment but such problems can be solved.

Given that the system under question (the turbine) is governed by physical laws that do not change over the history and that h is sufficiently large, then it follows that the function $\boldsymbol{f}$ exists: $\mathbf{f}(\mathbf{H}) = \mathbf{H} \prod \mathbf{x}^{(1)}$, where the symbol $\prod$ indicates concatenation of the vector $\mathbf{x}^{(1)}$ to the right side of the matrix $\mathbf{H}$. This function may be applied recursively so that $\mathbf{f}^n(H) = \mathbf{x}^{(n)}$. In this way, we may compute the n-th state of the system, i.e. the state that the system will have in n time steps from the current time.

3 Theoretical Limitations

Of course, whatever methods we choose, they cannot have arbitrary accuracy or stability. Thus, every $\mathbf{x}^{(t)}$ has an inherent measurement induced uncertainty $\Delta\mathbf{x}^{(t)}$ attached to it. This means that the true value of the state vector is somewhere in the range $[\mathbf{x}^{(t)} - \Delta\mathbf{x}^{(t)}, \mathbf{x}^{(t)} + \Delta\mathbf{x}^{(t)}]$.

Please note that no measurements made in the real world are ever completely precise. There are random and structured errors associated with the measurement process, also physical sensors drift with age and environmental effects. All of these must be taken into account to determine a reasonable measurement uncertainty $\Delta\mathbf{x}^{(t)}$.

A further limitation is the length of the history. The history must contain a record of the variations that are to be expected in the future so that these variations, correlations and other structures may be included in the model. It is thus desirable that the history be as long as possible and also the time unit (governing the frequency of measurements) be as small as possible. Together these two define a history that contains the maximum available knowledge about the system.

Our efforts are thus limited by three fundamental factors: (1) The number and identity of the measurements made, (2) the length, frequency and variability of recorded history and (3) the inherent accuracy of a measurement itself. Together these three factors will determine whether a reliable and stable model can be found.

4 Application

Initially, the machine learning algorithm was provided with no data. Then the points measured were presented to the algorithm one by one, starting with the first measured point $\mathbf{x}^{(-h)}$. Slowly, the model learned more and more about the system and the quality of its representation improved. Once even the last measured

point $\mathbf{x}^{(0)}$ was presented to the algorithm, it was found that the model correctly represents the system; see figure 1 for an example.

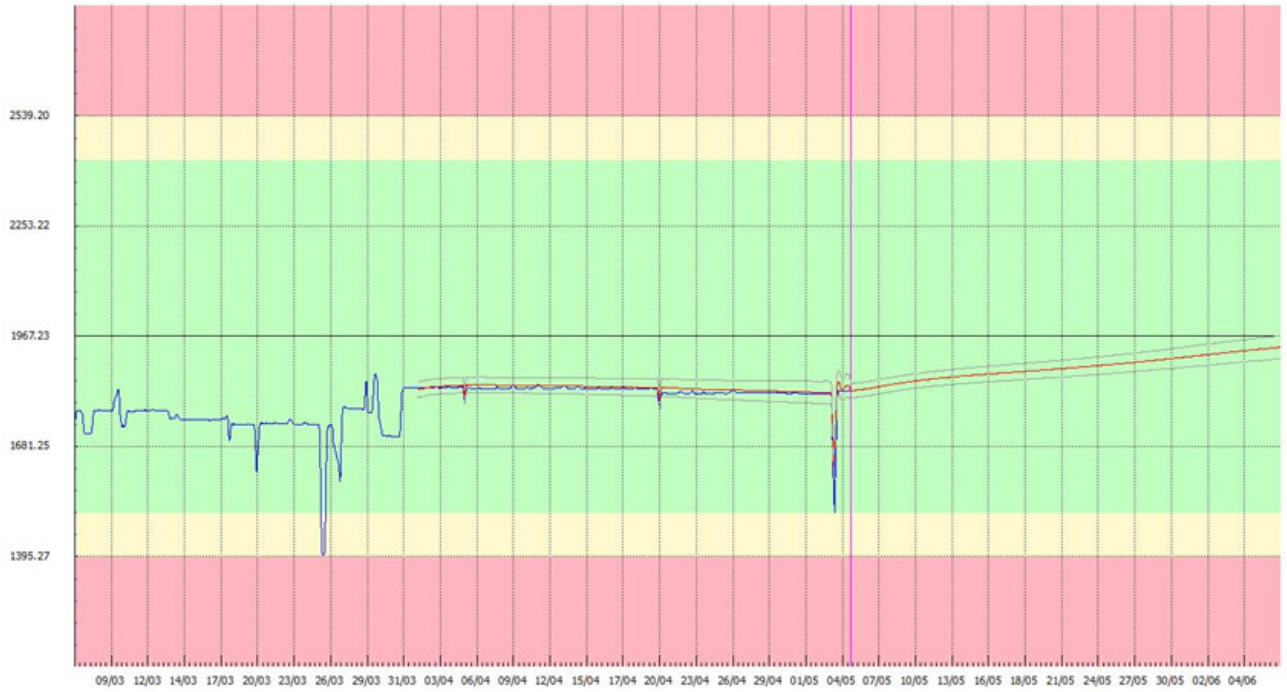

Figure 1: The discharge pressure of a pump as measured (blue curve) and calculated from the model (red curve). We observe that the model is able to correctly represent the pump as exemplified by this one variable.

The model was then inverted for optimization of yield. The computation was done for the entire history available of 2.5 years and it was found that the optimal point deviated from the actually achieved points by approximately 5% in absolute terms; see figure 2.

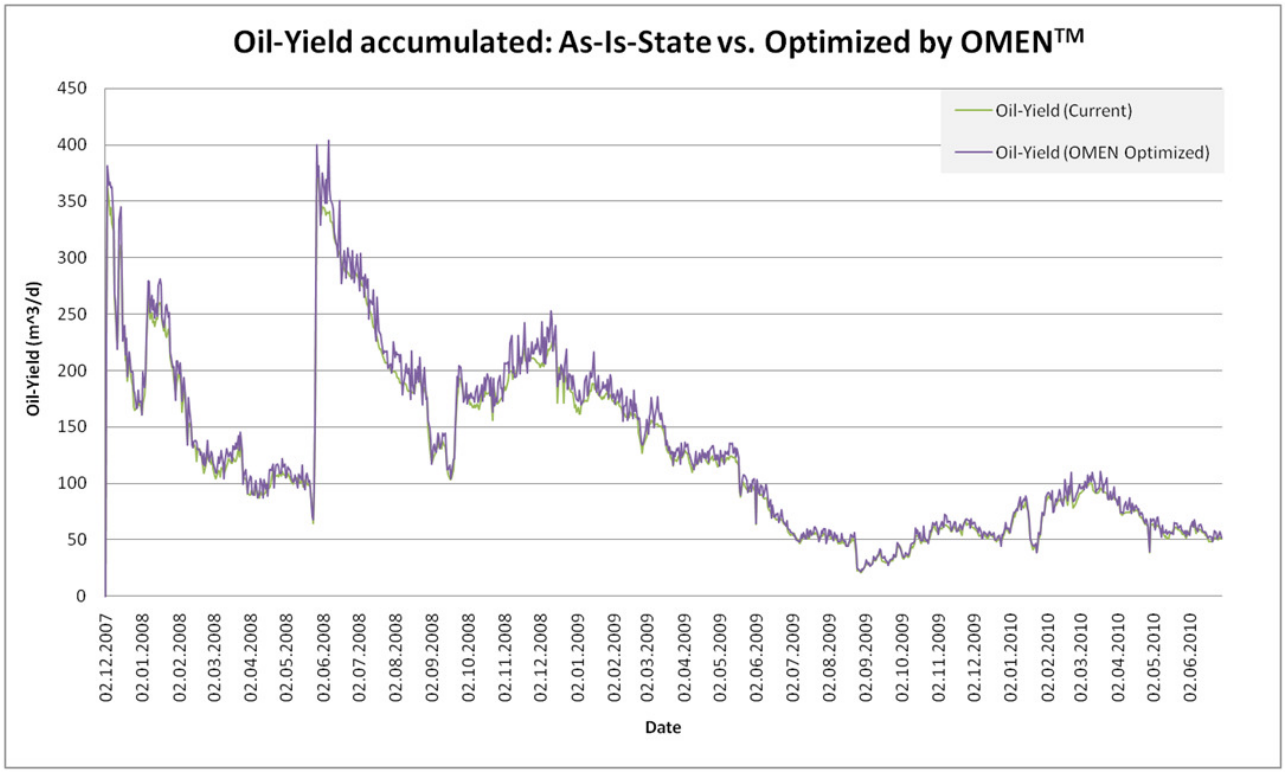

Figure 2: The yield of all considered wells together as observed (green curve) and optimized (blue curve). The difference between the two curves is 5% total yield over the history.

5 Conclusion

The main benefits of the current approach are: (1) processes *all* measured parameters from the rig in *real-time*, (2) encompasses all *interactions* between these parameters and their time evolution, (3) provides a *uniform* and *sustainable* operational strategy 24 hours per day and (4) achieves the *optimal* operational point and thus smoothes out variations in human operations.

Effectively the model represents a virtual oil rig that acts identically to the real one. The virtual one can thus act as a proxy on that we can dry run a variety of strategies and then port these to the real rig only if they are good. That is the basic principle of the approach. The novelty here is that we have demonstrated on a real rig, that it is possible to generate a representative and correct model based on machine learning of historical process data. This model is more accurate, all encompassing, more detailed, more robust and more applicable to the real rig than any human engineered model possibly could be.

The increase of approximately 5% in yield is significant as it will allow the operator to extract more oil in the same amount of time as before and thus represents an economic competitive advantage.

6 References

[1] Bishop, C.M.: Pattern Recognition and Machine Learning. Heidelberg: Springer 2006

[2] Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Reviews 65, 1958, 386-408

[3] Mandic, D., Chambers, J.: Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability. Hoboken: Wiley 2001

[4] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes. Cambridge: Cambridge University Press 2010

Increase of Oil Production Yield in Shallow-Water Offshore Oil Wells in the Dagang Oilfield via Machine Learning

Ph.D. Tan Chaodong
MOE Key Laboratory of Petroleum Engineering, China University of Petroleum, Beijing 102249, China
Tel: +86 138 0133 1255
email: tantcd@126.com

Dr. Patrick Bangert, CEO
algorithmica technologies GmbH
Außer der Schleifmühle 67, 28203 Bremen, Germany
Tel: +49 (0) 421 337-4646
email: p.bangert@algorithmica-technologies.com
www.algorithmica-technologies.com

Liu Bailiang, Vice Director
PetroChina Dagang Oilfield Company, Tianjin 300280, China
Tel: +86 22 2591 9122
email: liubailiang@petrochina.com.cn

Zhang Jie, Vice CEO
Yadan Petroleum Technology Co Ltd
No. 37 Changqian Road, Hi-Tech Park, Changping, Beijing, P.R. China 102200
Tel: +86 158 0151 1758
email: nyboc@sina.com