Rio Oil & Gas Expo and Conference 2022

ISSN 2525-7579

Conference Proceedings homepage: https://biblioteca.ibp.org.br/riooilegas

Technical Paper

Detecting anomalies in production data using machine learning techniques

Aurea Rossy Soriano Vargas [1]
Rafael de Oliveira Werneck [2]
Maiara Moreira Gonçalves [3]
Eduardo dos Santos Pereira Eduardo Pereira [4]
Leopoldo André Dutra Lusquino Filho [5]
Soroor Salavati [6]
M. Manzur Hossain [7]
Alexandre Mello Ferreira [8]
Alessandra Davólio Gomes [9]
Denis José Schiozer [10]
Anderson de Rezende Rocha [11].

1. UNICAMP, INSTITUTO DE COMPUTAçãO, . CAMPINAS - SP - BRASIL, aurea.soriano@ic.unicamp.br
2. UNICAMP, INSTITUTO DE COMPUTAçãO, . CAMPINAS - SP - BRASIL, rafael.werneck@ic.unicamp.br
3. UNICAMP, CENTRO DE ESTUDOS DE PETRóLEO, . CAMPINAS - SP - BRASIL, maiara.moreira.gocalves@gmail.com
4. UNICAMP, INSTITUTO DE COMPUTAçãO, . CAMPINAS - SP - BRASIL, eduardo.santos@inpe.br
5. UNICAMP, INSTITUTO DE COMPUTAçãO, . CAMPINAS - SP - BRASIL, leopoldolusquino@gmail.com
6. UNICAMP, INSTITUTO DE COMPUTAçãO, . CAMPINAS - SP - BRASIL, s264967@dac.unicamp.br
7. UNICAMP, CENTRO DE ESTUDOS DE PETRóLEO, . CAMPINAS - SP - BRASIL, manzur_hossain@yahoo.com
8. UNICAMP, INSTITUTO DE COMPUTAçãO, . CAMPINAS - SP - BRASIL, melloferreira@ic.unicamp.br
9. UNICAMP, CENTRO DE ESTUDOS DE PETRóLEO, . CAMPINAS - SP - BRASIL, davolio@unicamp.br
10. UNICAMP, CENTRO DE ESTUDOS DE PETRóLEO, . CAMPINAS - SP - BRASIL, denis@unicamp.br
11. UNICAMP, INSTITUTO DE COMPUTAçãO, . CAMPINAS - SP - BRASIL, anderson.rocha@ic.unicamp.br

Abstract

Oil production data may present anomalous behavior that does not reflect the actual reservoir dynamics. Some causes are human interventions, abrupt increase of water, severe slugging, flow instability, amongst others. The data relating to these anomalous events needs to be identified and removed from the dataset due to their potential to change the correlation of the series and influence forecasting and classification results. This work describes the use of two unsupervised anomaly detection techniques (DBSCAN and GMM) and one supervised strategy based on recurrent neural networks underpinned by machine learning to discover observations that do not behave as expected. Our experiments were performed using two datasets: UNISIM-II-M-CO (synthetic benchmark model) and 3W (a real Brazilian field dataset), and we evaluated them considering two metrics: recall and balanced accuracy. These strategies show promising results, with over 93% of recall and 86% of accuracy for UNISIM-II-M-CO and over 99% of recall and 80% of accuracy for the 3W dataset with both strategies. Such results help us conclude that the methods are accurate, precise, and robust to identify different types of data anomalies before performing machine-learning techniques using the data.

Keywords: reservoir production. anomaly detection. time series. machine learning

Received: October 04, 2021 | **Accepted:** August 25, 2022 | **Available online:** September 26, 2022
Article nº: 298
Cite as: Proceedings of the Rio Oil & Gas Expo and Conference, Rio de Janeiro, RJ, Brazil, 2022.
DOI: https://doi.org/10.48072/2525-7579.rog.2022.298

1. Introduction

During the life of a producing oil field, multiple variables are measured over time, such as fluid production and pressure data, that create very large data-sets of valuable data that if analyzed appropriately using advanced ML/AI techniques, can provide deep insights and knowledge on the reservoir behavior. This type of data is called time-varying multivariate data, a sub-category of data streams in which data instances are described by multiple time-stamped variables recorded sequentially.

Production time-varying multivariate data contains the necessary information to analyze the reservoir state subjected to a given production system. This data can be subject to unexpected or uncontrollable events like any time series. Analyzing time-varying multivariate data requires the ability to explore the variables thoroughly to identify patterns, analyze their behavior (Steed et al., 2017), and detect anomalous values related to those events. Detecting patterns and anomalies is particularly challenging (Martin and Quach, 2016) because this data contains hundreds, thousands, or even millions of instances, and sometimes the analysis is conducted with limited prior knowledge, if any.

In the production stage, anomalies may occur for different reasons, e.g., gross errors or human interventions. We must detect and disregard these data as they affect different correlation and human intervention-free (HIF) forecasting analysis algorithms, resulting in distorted and misleading results, i.e., the quality of historical data directly affects the quality of prediction algorithms for reservoir behavior. Our goal is to identify patterns that do not behave as expected and remove them later so we can use the remaining data confidently to perform other tasks, such as production forecasting as accurately as possible.

Many anomaly detection techniques have been proposed based on distribution, distance, density, clustering, and classifications in the literature, whose applications vary depending on the problem domains and even the dataset. Distribution-based approaches use statistical distributions to model the data points; for instance, one of the approaches proposed by Soriano-Vargas et al. 2021, suggests that deviations from the model are labeled as anomalies. Distance-based approaches label how distant a data point is from a subset of points; however, they cannot cope with time-series datasets given the time characteristics and the presence of dense and sparse regions (Breunig et al., 2000). Density-based anomaly detection approaches have been proposed to overcome the problem of dense and sparse regions employing the local outlier factor (LOF), depending on the local density of its neighborhood (Breunig et al., 2000). However, it fails to properly deal with data in different granularities. Clustering algorithms can detect anomalies as data points that do not belong to, or that are near, any cluster (Juang et al. 2001). We must, however, guarantee the temporal aspect with time-series data. With classification approaches, we need to identify the categories to which a data point belongs, considering two phases: first, learning a model based on subset data points, and second, inferring a class for new data points based on the learned model. Anomaly detection approaches based on classification can be grouped into two categories: multi-class (Barbara et al. 2001) and one-class techniques (Roth 2004).

In this context, we leverage three machine learning-based anomaly detection techniques: the Gaussian Mixture Models (GMM) (Reynolds, 2009), the Density-Based Spatial Clustering of Application with Noise (DBSCAN) (Khan et al., 2014), and the application of deep learning approaches in the anomaly detection task, such as Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). These strategies are explored in an anomaly detection pipeline to deal with these problems: temporal aspect, dense and sparse regions, multiple granularities, and lack of annotated data. Our approaches consider the prior probability of anomalies from a time window; the unsupervised strategies do not require any labeled training data with normal and abnormal conditions and include specialist knowledge in the exploration process.

We apply the three anomaly detection techniques to identify anomalies in two independent datasets of production data. One of those datasets contains High-Frequency Data (HFD), i.e., time-series sampling at every second. Our work novelties include (1) a more efficient pre-processing phase, in which we highlight the samples' differences to project them in another feature space; (2) the adaptation of anomaly detection algorithms for event identification over time. The first point also allows us to work with HFD from a real field as input to our algorithms, dealing with the different temporal aspects through moving time windows.

2. Production Anomalies

Anomalies or outliers are values that deviate from observations on data, which may indicate a measurement variability, an entry/experimental error, or a human intervention. An anomaly can also occur individually or in a group. Lack of data (presence of zeros or nulls) and rapid and temporary changes in level (valleys and peaks) are the most common problems found in production data (Wising et al., 2009). Missing data may be associated with a failure to acquire or record data or due to human interventions. Valleys are observations that differ from the values of nearby measurements and are usually related to a (1) partial closure of the well. Specifically, they start with a negative slope and end with a positive slope without reaching a zero value in the range. When it reaches zero values, we are in a complete closure situation. In addition, we are also interested in detecting peaks, i.e., data with a high positive slope. These peaks are related to (2) possible failures in capacity control or the reopening of wells after closure.

Other types of anomalies can be associated with (3) abrupt increase of Basic Sediment and Water (BSW), (4) spurious closure of Downhole Safety Valve (DHSV), (5) severe slugging, (6) flow instability, (7) rapid productivity loss, (8) temporary restriction in Production Choke (PCK), (9) scaling in PCK, (10) hydrate in the production line and others (Vargas et al. 2019).

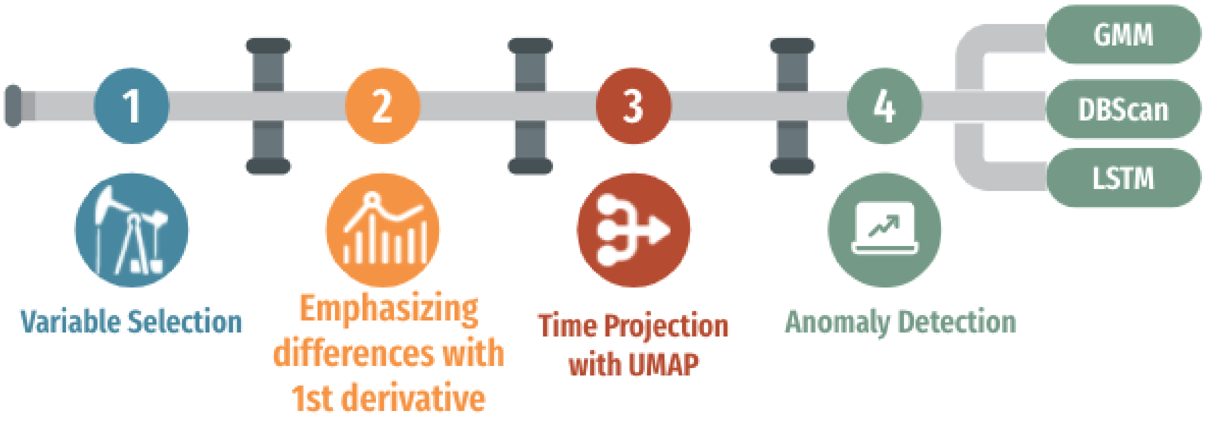
3. Methodology

Our focus is to identify data related to anomalous events. An overview of the data processing and exploration stages is shown in Figure 1, which is divided into three stages. First, from the data selected from one well, or a set of wells (producers or injectors), the user must select the variables considered in the anomaly detection.

Once we have selected the well variables, we emphasize differences using the first derivative. With this approach, we deal with the multi-granularity problem. Then, as the anomaly detection strategies work with just one variable, we applied a time projection by using UMAP (McInnes et al., 2018) in each time instant. UMAP is a manifold-learning technique for dimension reduction, constructed from a theoretical framework based on Riemannian geometry and algebraic topology. Finally, we apply UMAP respecting the temporal characteristics of the variables, i.e., we find a projection of a set of variables at each instant of time.

The anomaly detection strategies are applied to the modified time series. We use these strategies to analyze each data point considering a temporal interval. We leverage anomaly detection strategies to analyze each data point considering a time interval of 15-time units once the best results are found with intervals between 15- and 18-time units.

Figure 1 – Overview of the data processing and exploration stages.


1
2
3
4
GMM
DBScan
LSTM
Variable Selection
Emphasizing differences with 1st derivative
Time Projection with UMAP
Anomaly Detection

Source: Authors

Our approach includes three tailored anomaly detection strategies, GMM (Reynolds, 2009), DBSCAN (Khan et al., 2014) applied to time series, and Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). GMM and DBSCAN are unsupervised approaches, while LSTM is supervised.

In GMM, we fit *k* Gaussians and find the Gaussian distribution parameters, such as mean and variance, for each cluster and cluster weight. Then, for each point, we calculate its probability of belonging to each cluster and compare each probability with a decision cutoff.

DBSCAN finds high-density core samples and expands clusters from them. To include the temporal characteristic, we also apply it in the sliding time window. Forming a dense region requires two parameters: eps and the minimum number of points (*minPts*). After empirical experiments, we obtained the best results with *eps* = 3 and *minPts* = 2.

Recurrent Neural Networks (RNNs) are a kind of artificial neural network, where each unit (or neuron) has a specific value, known as weight, that is attributed through a process known as training. During the training, many examples are supplied to the network to learn the best weights in an optimization process, trying as much as possible to generalize and make the least mistakes when exposed to unknown examples. Unlike the unidirectional neural networks (known as feed-forward), RNNs have cycles between their units, i.e., units can have connections to units from previous layers or from the same layer, which creates much more complex models to solve a broader range of problems.

We choose the model called LSTM (Hochreiter and Schmidhuber, 1997), a specific recurrent architecture that allows longer sequences as input. The main idea is to capture the regular patterns in a previous time window from which the algorithm will predict the next day's value. We can then calculate the prediction errors from the actual and predicted data, see Figure 2. We hypothesize that minor errors will be obtained from typical values and, therefore, they may be more frequent. Conversely, anomalies will generate high-value errors, which will be less frequent.

3.1. Datasets

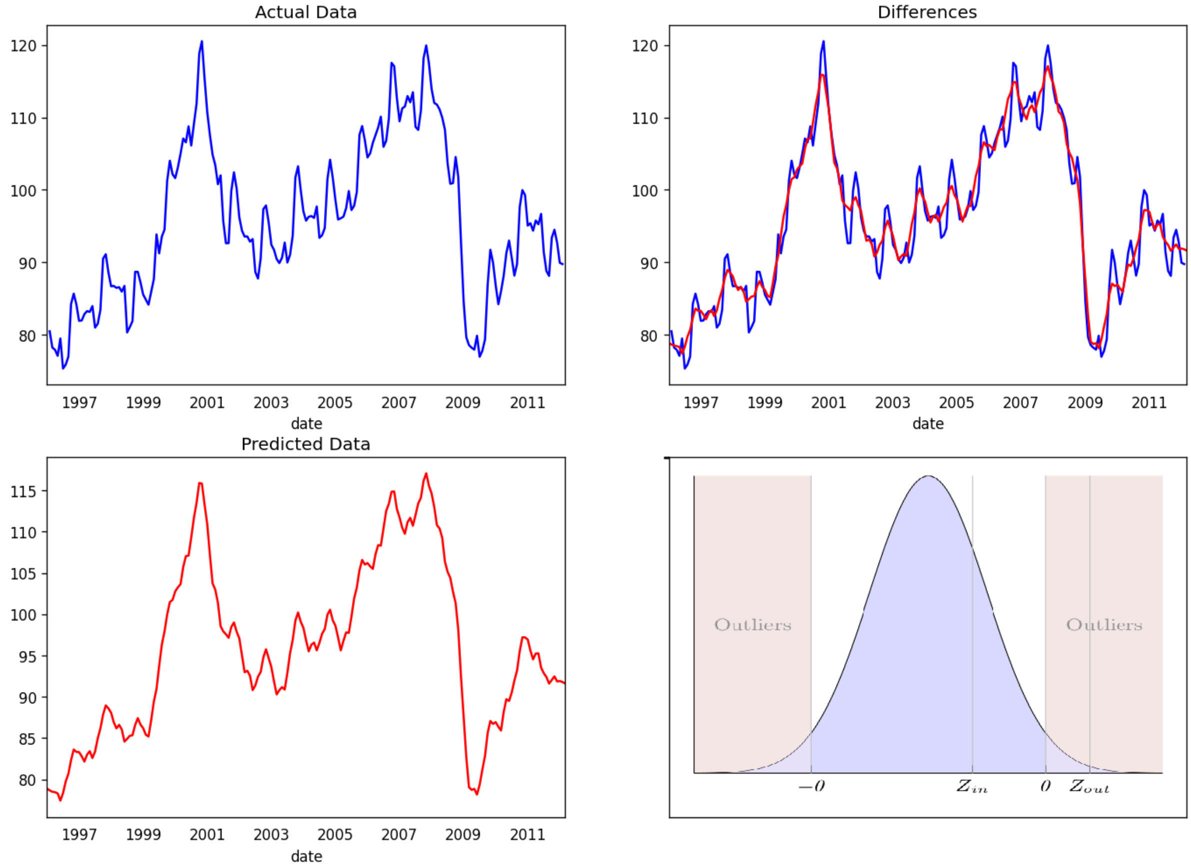
For our experiments, we used the benchmark case UNISIM-II-M-CO, which is a synthetic reservoir model based on pre-salt features created by the UNISIM group at Unicamp, Brazil (Correia et al., 2015), and a fully-annotated anomaly dataset (3W) published by Vargas et al. (2019).

The UNISIM-II-M-CO data comprise fluid rates and pressure for producer and injector wells, which contains anomalies related to the (1) partial closure of the well and (2) failures in capacity control or the reopening of wells after closure. The injection and production rates have trends mimicking actual fields that account for partial and complete closure of wells. The simulation model provides 6.5 years of production history, containing eight injection wells and ten production wells. We use the daily Oil, Gas, and Water production for production wells from the simulated data variables.

The 3W dataset contains data from offshore wells with the natural flow in a normal state, considering more common monitored variables, such as Pressure at the Permanent Downhole

Gauge (PDG), Pressure at the Temperature and Pressure Transducer (TPT), Temperature at the TPT, Pressure upstream of the Production Choke (PCK), and Temperature downstream of the PCK. Vargas et al. (2019) differentiated eight types of anomalies in this dataset: (3) abrupt increase of Basic Sediment and Water (BSW), (4) spurious closure of Downhole Safety Valve (DHSV), (5) severe slugging, (6) flow instability, (7) rapid productivity loss, (8) temporary restriction in Production Choke (PCK), (9) scaling in PCK, and (10) hydrate in the production line and others.

Figure 2 – Proposed anomaly detection using RNNs.



Source: Authors

4. Results

We present two case studies for validating these strategies using the two selected sets. For UNISIM-II-M-CO, we used Daily Production of Oil, Gas, and Water for producer wells from the simulated data variables. For the 3W dataset, we used Pressure at the Permanent Downhole Gauge (PDG), Pressure at the Temperature and Pressure Transducer (TPT), Temperature at the TPT, Pressure upstream of the Production Choke (PCK), and Temperature downstream of the PCK.

We applied the three exploration protocols using an interval of 15-time units with DBSCAN, GMMs, and LSTMs, whose DBSCAN and GMMs parameters are by default. For LSTMs, we also consider a negative cutoff of -3 and a positive cutoff of 3. We divide the time series into 70% for training and 30% for testing, respecting the time-space. From the first group, sequences of 15 continuous-time units (time window) were broken down with the prediction objective of the following day. This information was used to train our LSTM and recognize the reservoir's behavior, with a learning rate of 0.001, MSE loss function, and Adam Optimizer. The group of 30% was used for testing, breaking down into 15-day sequences in the same way and requiring the network to give us the predicted value of each of those sequences.

We evaluated our approach based on two criteria: recall and balanced accuracy for identified anomalies. The results are shown in Table 1.

We also show a visual example of our three approaches in Figure 3. From the three variables DailyProdOil, DailyProdGas, DailyProdWater for the PRK045 well from the UNISIM-II-M-CO data, we were able to identify the anomalies (in green) with the three approaches (DBSCAN, GMM, and LSTMs in red). We can also notice that our strategies captured some other sudden changes, which may be related to other types of anomalies that our specialists did not list.

For the UNISIM-II-M-CO, the best results were obtained with GMMs and, for the 3W dataset, with LSTM. This could be due to the parameter setting. A subsequent experiment should include studying various parameters of these methods to improve the identification of anomalies. Nevertheless, our experiments proved promising when identifying anomalies with DBSCAN, GMMs, and LSTM.

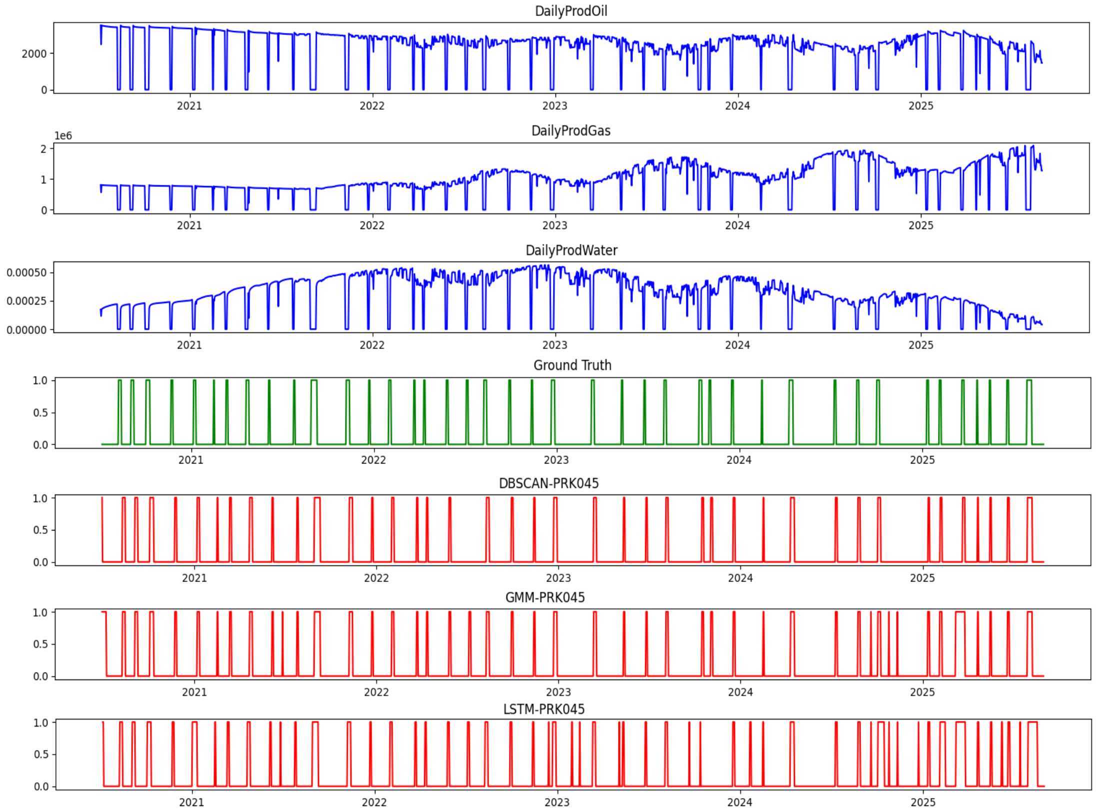
We emphasize that these anomaly detection strategies can be useful for production forecasting tasks, especially recent data driven approaches based on machine learning. The idea is to ignore the anomalous points in order to better capture the reservoir dynamics. However, this does not mean that anomalous points will be disregarded for other tasks.

Table 1: Results of anomaly detection strategies applied to UNISIM-II-M-CO and 3W.

Dataset	Recall (DBSCAN)	Accuracy (DBSCAN)	Recall (GMM)	Accuracy (GMM)	Recall (LSTM)	Accuracy (LSTM)
UNISIM-II-M-CO	93%	86%	97%	94%	91%	87.4%
3W dataset	99.9%	99.2%	99.9%	80%	99.9%	99.8%

Source: Authors

Figure 3 – Visual example of the anomaly detection strategies using three variables DailyProdOil, DailyProdGas, DailyProdWater for the PRK045 well from the UNISIM-II-M-CO data. The time-series variables are in blue. The ground-truth anomaly data is in green and the strategy result is in red.



Source: Authors

To show the ability of the anomaly detection strategies in forecasting tasks, we used a simple regression model based on LSTM, using the Mean Absolute Error (MAE) as the evaluation metric. We define a sequential model and add four blocks of a LSTM layer and a Dropout Layer combination. The LSTM layer has 50 neurons, then a dropout layer is used for regulating the network. The final Dense layer is the output layer which has 1 cell. We use Adam Optimizer, 32 batch size, and 200 epochs.

For the UNISIM dataset, we apply the regression model before and after the GMM strategy to predict the Daily Oil Rate, and for the 3W dataset, we also apply the model before and after the LSTM strategy to predict the Pressure at the Temperature and Pressure Transducer. We chose these strategies because they achieved the best results in anomaly detection (see Table 1). Below, in Table 2 and Figures 4 and 5, we show how much the results were improved in terms of the MAE metric.

Table 2: Forecasting MAE results before and after applying anomaly detection strategies to the UNISIM-II-M-CO and 3W datasets.

Dataset	Well	Before	After
UNISIM-II-M-CO	PRK028	329.457	288.66
	PRK045	264.16	234.48
	PRK052	99.80	68.74
	PRK060	278.20	298.63
	PRK061	424.38	340.40
	PRK084	260.52	252.01
	PRK085	1125.41	240.68
3W dataset	Simulated 04	77062.23	39104.15
	Simulated 05	46782.41	22606.76
	Simulated 06	61057.92	29414.03
	Simulated 07	75014.57	34694.19

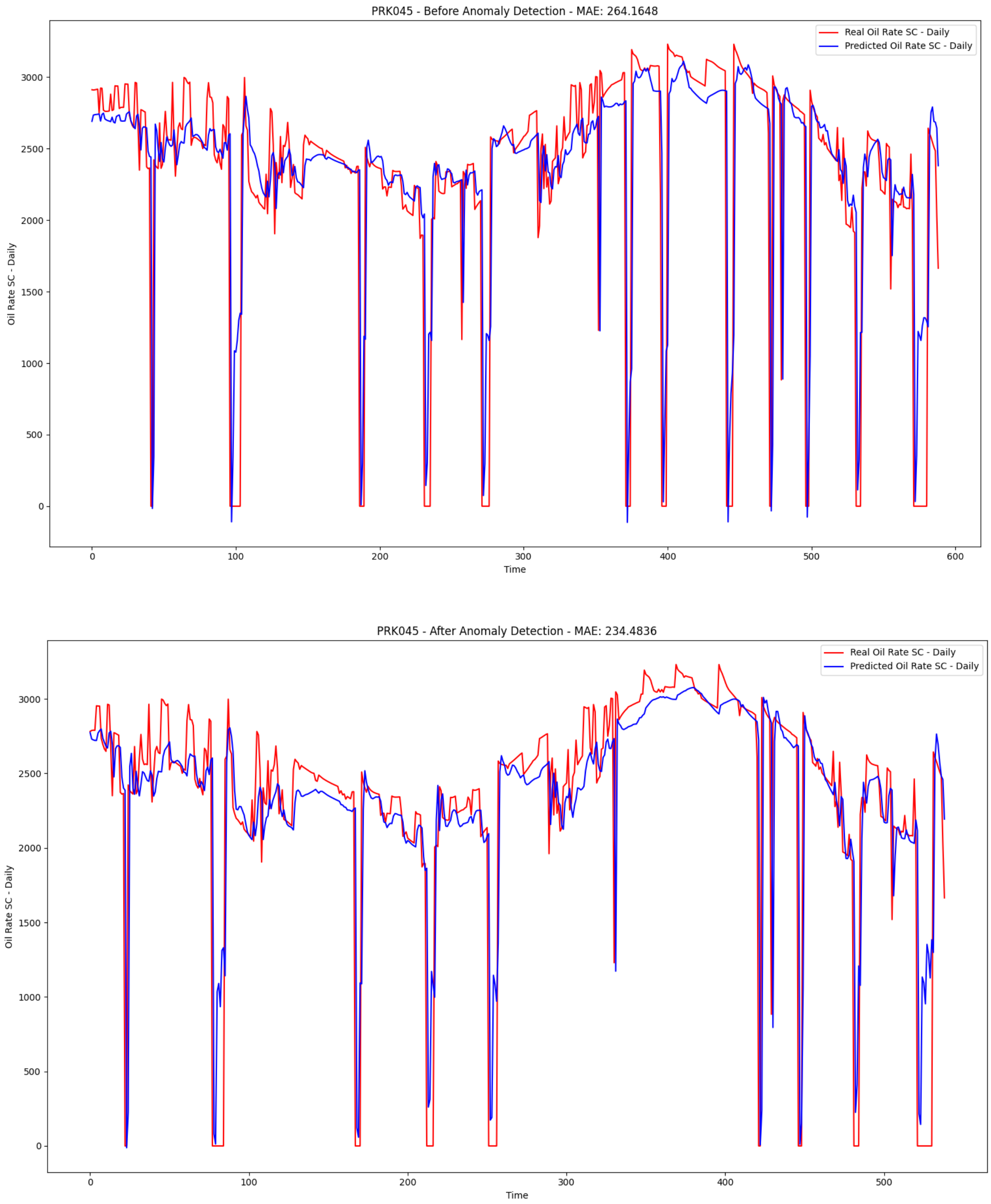
Source: Authors

5. Final Considerations

The annotation of anomalous events is complex and involves careful control, recording, and monitoring of values. For this reason, our methods are significant for the domain specialists as two of them are unsupervised and work well in different real reservoir data anomaly situations.

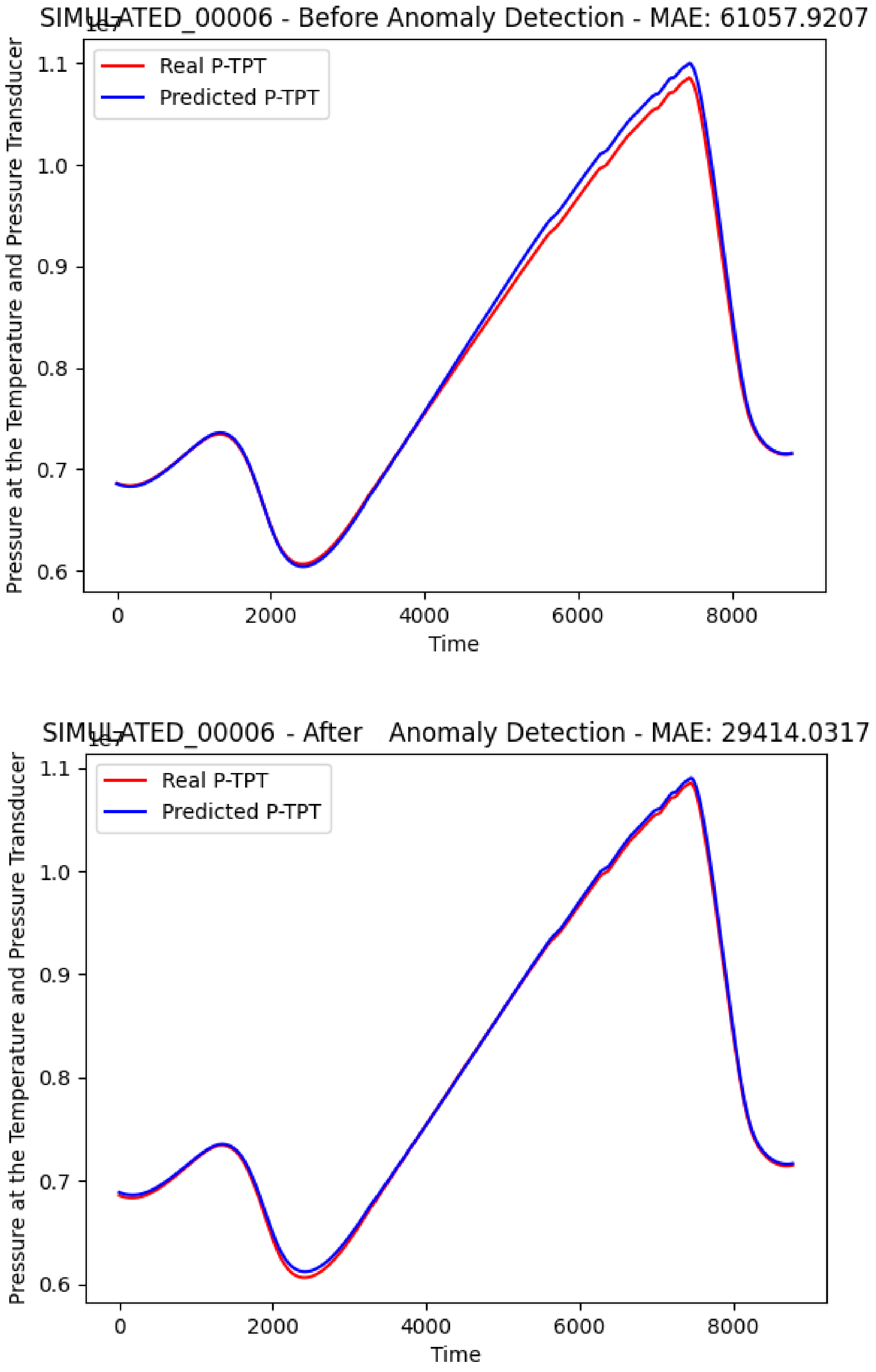
We analyzed strategies for anomaly detection in time series, considering a sliding time window (interval of time units). We applied, evaluated, and discussed our approach to a reference dataset (UNISIM-II-M-CO) and the 3W dataset. Our main contribution in this work is to tackle the problem of identifying anomalies by using a purely data-driven approach, which successfully combines a pre-processing phase, in which the samples are highlighted and projected onto another feature space with UMAP, with the adaptation of well-known anomaly detection algorithms. We compared our results to annotations made by specialists to confirm our findings and the results are promising, especially as our results were able to identify possible new events that were overlooked by the specialists. We also dealt with HFD, which is not very common practice.

Figure 4 – Forecasting results before and after applying anomaly detection strategies to the PRK045 well from UNISIM-II-M-CO (before MAE result: 264.1648 and after MAE result: 234.4836).



Source: Authors

Figure 5 – Forecasting results before and after applying anomaly detection strategies to the Simulated_00006 well from 3W dataset (before MAE result: 61057.9207 and after MAE result: 29414.0317).


SIMULATED_00006 - Before Anomaly Detection - MAE: 61057.9207

SIMULATED_00006 - After Anomaly Detection - MAE: 29414.0317

Source: Authors

Furthermore, the methods can be coupled with forecasting production data end-to-end, eliminating data inconsistencies and improving data-driven forecasts, as we showed for LSTM regressions in our datasets. We identified different characteristics from the detected anomalies, such as an interval of time units, initial slope, mean value, and other variables presenting rare values. Therefore, one of our next steps will be to explore representative characteristics to identify different clusters of anomalies.

6. Acknowledgments

This work was conducted in association with the ongoing Project registered under ANP number 21373-6 as "Desenvolvimento de Técnicas de Aprendizado de Máquina para Análise de Dados Complexos de Produção de um Campo do Pre-Sal" (UNICAMP/Shell Brazil/ANP) - "Machine-Learning Development for Analysis of Complex Production Data in a Pre-Salt Carbonate Field" - funded by Shell Brazil, under the ANP R&D levy as "Compromisso de Investimentos com Pesquisa e Desenvolvimento". The authors thank also Schlumberger and CMG for software licenses.

Referências

Barbara, D., Wu, N., & Jajodia, S. (2001). *Detecting novel network intrusions using bayes estimators*. *1*, 1–17. https://doi.org/10.1137/1.9781611972719.28

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). *LOF: identifying density-based local outliers*. *29*, 93–104. https://doi.org/10.1145/335191.335388

Correia, M., Hohendorff, J., Gaspar, A. T., & Schiozer, D. (2015). *UNISIM-II-D: Benchmark Case Proposal Based on a Carbonate Reservoir*. *1*, SPE-177140-MS. https://doi.org/10.2118/177140-MS

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jiang, M. F., Tseng, S. S., & Su, C. M. (2001). Two-phase clustering process for outliers detection. *Pattern Recognition Letters*, *22*(6–7), 691–700. https://doi.org/10.1016/S0167-8655(00)00131-8

Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014). *DBSCAN: Past, present and future*. *1*, 232–238. https://doi.org/10.1109/ICADIWT.2014.6814687

Martin, S., & Quach, T. T. (2016). *Interactive visualization of multivariate time-series data*. *9744*, 322–332. https://doi.org/10.1007/978-3-319-39952-2_31

McInnes, L, & Healy, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. ArXiv e-prints. https://arxiv.org/pdf/1802.03426.pdf

Reynolds, D. (2015). Gaussian Mixture Models. In *Encyclopedia of Biometrics*. Springer US. https://link.springer.com/referenceworkentry/10.1007/978-1-4899-7488-4_196

Roth, V. (2004). *Outlier Detection with One-class Kernel Fisher Discriminants*. *17*, 1--8. https://proceedings.neurips.cc/paper/2004/hash/1680e9fa7b4dd5d62ece800239bb53bd-Abstract.html

Soriano-Vargas, A., Werneck, R., Moura, R., Júnior, P. M., Prates, R., Castro, M., Gonçalves, M., & et al. (2021). A visual analytics approach to anomaly detection in hydrocarbon reservoir time series data. *Journal of Petroleum Science and Engineering*, *206*(1), 108988. https://doi.org/10.1016/j.petrol.2021.108988

Steed, C. A., Halsey, W., Dehoff, R., Yoder, S. L., Paquit, V., & Powers, S. (2017). Falcon: Visual analysis of large, irregularly sampled, and multivariate time series data in additive manufacturing. *Computers & Graphics*, *63*(2017), 50–64. https://doi.org/10.1016/j.cag.2017.02.005

Vargas, R. E. V., Munaro, C. J., Ciarelli, P. M., Medeiros, A. G., do Amaral, B. G., Barrionuevo, D. C., de Araújo, J. C. D., Ribeiro, J. L., & Magalhaes, L. P. (2019). A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering*, *181*(2019), 106223. https://doi.org/10.1016/j.petrol.2019.106223

Wising, U., Vrielynck, B., Kalitventzeff, P. B., & Campan, J. (2009). *Improving Operations Through Increased Accuracy of Production Data*. *1*, SPE-124766-MS. https://doi.org/10.2118/124766-MS