Visual Analysis of Anomalous Behavior Patterns in Oil Wells Using Dimensionality Reduction with t-SNE Projection

Bruno Batista dos Santos[1], Thales Miranda de Almeida Vieira[1]

[1]*Federal University of Alagoas, Av. Lourival Melo Mota, S/N, Tabuleiro do Martins, Maceió - AL, 57072-970*
bruno.batista@ctec.ufal.br
[2]*Federal University of Alagoas, Av. Lourival Melo Mota, S/N, Tabuleiro do Martins, Maceió - AL, 57072-970*
thalesv@gmail.com

Abstract. Anomaly detection in industry processes, particularly in the oil industry, is crucial due to the complexity and criticality of these systems. Oil wells are subject to various dynamic and interdependent variables, monitored using sensors along the well. Precise analysis of these sensor records is essential for early anomaly detection, as deviations can have serious consequences. Given the multidimensionality and variability of this data, machine learning techniques are indispensable for analysis and classification. However, anomaly identification remains challenging due to the oscillatory behavior and scale disparity of the variables throughout the well's lifetime. Dimensional reduction techniques, such as Distributed Stochastic Neighbor Embedding (t-SNE), can mitigate these challenges by reducing computational cost, noise, and improving classifier accuracy. t-SNE projects high-dimensional data into a low-dimensional space, preserving clusters and bringing similar data together. This study aims to identify behavior patterns in multivariate data from producer and injector oil wells using t-SNE for dimensionality reduction. By visualizing real records of pressure and temperature sensors, the study seeks to identify patterns before and after anomalies, and explore the potential for developing artificial intelligence algorithms based on dimensional reduction data for anomaly detection.

Keywords: Anomaly Detection, Dimensionality reduction, T-SNE, Oil wells, Graphical visualization.

1 Introduction

With advancements in artificial intelligence technology, industries are increasingly turning to anomaly prediction in their processes using machine learning techniques. This approach is highly valuable in complex industries, such as the oil industry. Due to the intricate nature and the large number of variables involved in oil and gas well operations, early identification of anomalies can prevent catastrophic failures and optimize operational efficiency.

The analysis of the conditions of an oil well, whether it is a producer or injector, involves monitoring parameters such as pressure and temperature through sensors along the well (Fig. 1). Evaluating the state of the well is a delicate process, given the oscillatory behavior and the disparity of scales of the variables during operations. Consequently, developing an algorithm for anomaly detection is challenging, as distinguishing true anomalies from typical parameter fluctuations is difficult.

Graphically representing a vector with two variables is simple using a two-dimensional plot. A quantity with three variables can be represented using a three-dimensional plot. However, for a larger number of variables, as in the analysis of well parameters, there are no physical means to perfectly map this information on a plot. To address this problem, various dimensionality reduction algorithms are utilized.

According to Lopes [1], the technique of dimensionality reduction aims to decrease the complexity of data analysis by reducing the dimensionality to two or three dimensions while preserving the distances from the original dimension. These algorithms apply mathematical and statistical methods to achieve this goal. This technique offers advantages such as reduced computational cost, noise reduction, and improved classifier accuracy [2].

In this work, the t-SNE method will be used to perform dimensionality reduction of the pressure and temperature parameters of real wells to visually analyze possible patterns of anomalous behavior.

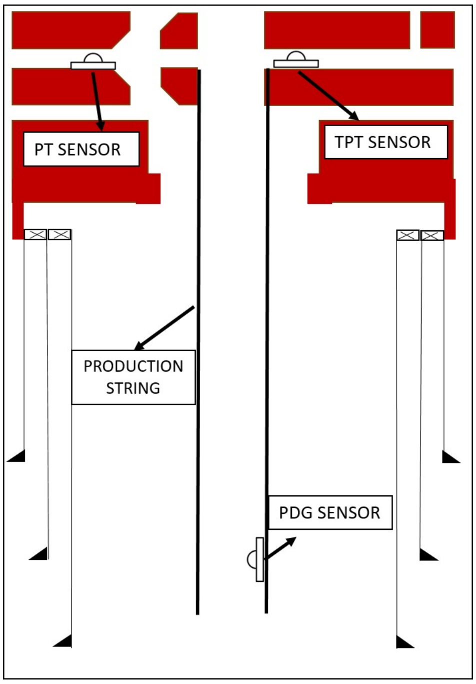

Figure 1. Some sensor to monitor the well

2 Methodology

2.1 Data Used

The data used in this study were CSV files containing real information from oil wells provided by Petrobras. In these datasets, it was known that anomalies occurred after valve maneuvers. The datasets contained time series spaced 30 seconds apart, where each time point recorded measurements from TPT, PDG, and PT sensors at 8 different locations along the well, resulting in 8 dimensions to be reduced to a two-dimensional plane using t-SNE. Anomalies were identified by the professionals responsible for monitoring the wells and are presented in the dataset in the “class” variable, which is labeled 0 for normal conditions and 1 for anomalous states. Table 1 describes the nomenclature and type of well for the datasets used.

The dataset consists of six files distributed across four wells, with the majority being injector wells, except for Well C, which is a producing well. Well A is represented by three files: A.1 and A.2 each contain 2881 observations, while A.3 contains 5761 observations. Wells B, C, and D each have a single file, with each containing 5761 observations. Notably, the three files associated with Well A (A.1, A.2, and A.3) are derived from the same well but represent data captured at different moments in time.

2.2 Dimensionality Reduction with t-SNE

The t-SNE (t-distributed Stochastic Neighbor Embedding) method is a non-linear dimensionality reduction technique widely used for visualizing high-dimensional data. Developed by van der Maaten and Hinton in 2008, t-SNE converts similarities between data points into joint probabilities and minimizes the Kullback-Leibler divergence between these distributions in high and low dimensions. This results in a visual representation where similar points are close together and dissimilar points are far apart, allowing for the identification of patterns in complex data [3] [4]. In this work, the goal is to use this technique to visualize the behavior of data when the well is anomalous, with the expectation that they will become distinct from the data recorded when the well is in a normal state.

One of the main hyperparameters of t-SNE is perplexity, which represents the number of nearby neighbors each data point considers. The perplexity value directly impacts the visualization’s density: lower values yield denser, more compact clusters by focusing on fewer neighbors, while higher values surround a broader range of neighbors, thereby capturing a more global data structure (Fig. 2). In this study, the choice of perplexity is aimed at better highlighting the behavior of anomalous points, with the expectation that when the well enters an anomalous state, a new cluster will form, distant from the others.

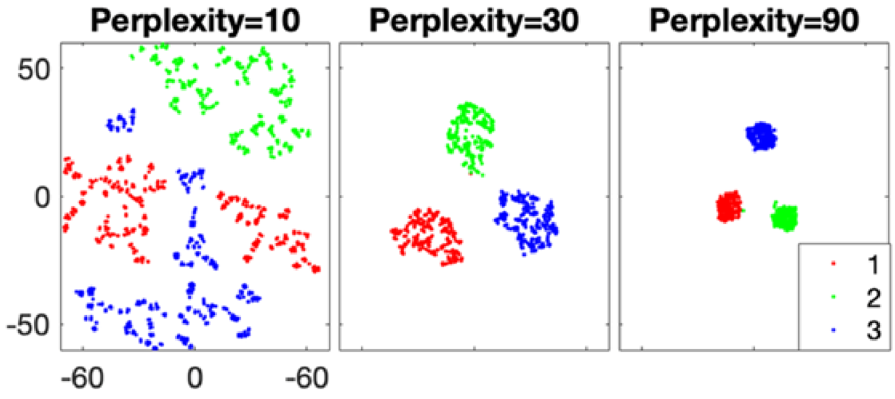

Figure 2. Visualization of clusters varying the perplexity hyperparameter [5]

2.3 Programming Language

For the computational implementation of this study, Python was chosen due to its widespread acceptance in the scientific community and its robust libraries for data analysis. The libraries employed in this work include Pandas, Matplotlib, and Scikit-learn.

Pandas is one of the most powerful and widely used tools in Python for data manipulation and analysis. It provides flexible and efficient data structures, such as the DataFrame, which organizes data into tabular form, facilitating complex operations [6]. This functionality is essential for managing the well datasets and preparing sensor information for dimensionality reduction using t-SNE.

Matplotlib is a data visualization library that enables the creation of high-quality graphs and plots. It offers a vast collection of efficient tools for tasks such as classification, regression, and clustering [6]. Scikit-learn, one of the most popular Python libraries for machine learning [6], will be used in this study to perform dimensionality reduction using the t-SNE function.

To highlight the moment when the anomaly occurs, two color scales will be applied to the graphs generated by Matplotlib: grayscale for the period before the anomaly and a palette ranging from cool to warm colors for the period after the anomaly. Additionally, the symbol ”x” will be used to mark the onset of the anomaly on the graph. This approach is intended to enhance the visualization of the transition and facilitate the identification of anomalies in the temporal context.

3 Results

Initially, all well datasets were processed using t-SNE with a perplexity value of 30. However, the results for some wells were not satisfactory. In this context, Li [7] confirmed that a higher perplexity makes the distribution of data clusters more sparse, improving visualization performance under different conditions. Therefore, the perplexity was gradually increased until more suitable results were achieved. It was observed that as the hyperparameter was incremented, there was an increase in computational cost, but the quality of the results also improved. Based on the tests conducted, it was determined that a perplexity of 100 offered the best balance between result quality and computational cost. Figure 3 presents the test results for each well.

With the perplexity parameter defined, t-SNE was applied to all wells. Figure 4 presents the results obtained from dimensional reduction of Well A.1, A.2 and Well C.

In the graphs of dimensionality reduction of the wells (Fig. 4), two color maps were assigned, which represent the moment before the anomaly (left scale) and after (right scale), as well as demonstrating the temporal progression through the change in colors. It can be observed that when an anomaly occurs, a new cluster forms in the dimensionality reduction. The color scale in the graphs, showing the onset of the anomaly (starting with blue and transitioning to red), allows identification of the moment when anomalous points begin to emerge. In most cases, the anomalous points appear within the cluster where the well is still in normal conditions. This occurs because the professional responsible for identifying the moment the anomaly occurs (variable ”class”) makes an estimate based on the observed parameters; however, the exact moment of the anomaly may be slightly earlier or

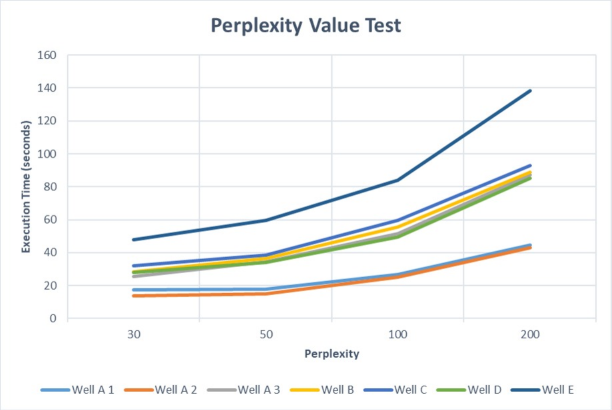

Figure 3. Varying perplexity value

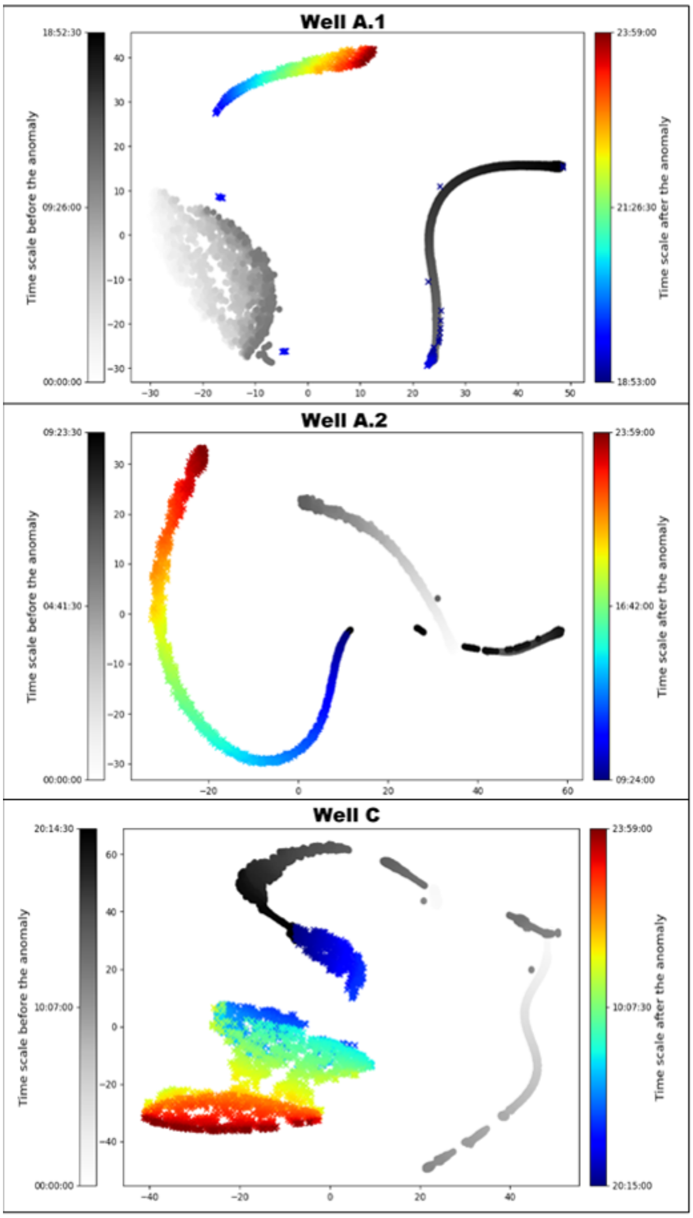

Figure 4. Dimensional reduction of Well A.1, A.2 and Well C

later than indicated.

Moreover, as in many cases a new cluster was observed to form shortly after the anomaly, it is evident that despite the variability and complexity of the data analyzed, dimensionality reduction allows the identification of patterns in anomalous data. This suggests that for the cases tested, it is feasible to identify anomalous behavior patterns using the t-SNE method for dimensionality reduction.

Given these results, the implementation of machine learning algorithms based on dimensionality reduction for the rapid identification of anomalies in oil wells can be recommended. This could not only improve operational efficiency but also enable more agile adoption of preventive measures, thereby increasing safety and reducing the risks associated with well failures.

4 Conclusions

The results of this work demonstrated the effectiveness of the t-SNE method in dimensionality reduction and pattern visualization in oil well data. It was observed that by increasing the perplexity hyperparameter value, the quality of the visualization significantly improved, allowing for a clearer and more coherent representation of the data.

The visual analysis revealed distinct patterns of data behavior before and after the occurrence of anomalies, evidenced by the clusters formed in the dimensionality reduction. These clusters facilitated the identification of anomalous transitions, highlighting the utility of t-SNE for this purpose.

Based on the visualization of data behavior patterns, it is possible to suggest the implementation of algorithms based on dimensionality reduction for anomaly detection in oil wells. These algorithms can improve operational efficiency and safety by allowing early identification of failures and the adoption of preventive measures.

Acknowledgements. The authors would like to thank PETROBRAS for the financial and technical support.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] M. A. S. L. . A parallel technique for data dimensionality reduction applied in smart cities. *Federal University of Rio Grande do Norte*, 2020.

[2] U. Maulik. Remote sensing image classification: A survey of support-vector-machine-based advanced techniques. *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, n. 1, pp. 33–52, 2017.

[3] van der L. Maaten. Visualizing data using t-sne. *Journal of Machine Learning Research*, vol. 9, n. 86, pp. 2579–2605, 2008.

[4] J. A. Le. Nonlinear dimensionality reduction. *Springer Science Business Media*, 2007.

[5] E. Ozanich. Deep embedded clustering of coral reef bioacoustics. *ResearchGate*, 2021.

[6] W. McKinney. Python for data analysis: Data wrangling with pandas, numpy, and ipython. *O'Reilly Media, Inc.*, 2017.

[7] Y. Li. Machine learning based defect detection in robotic wire arc additive manufacturing. *University of Wollongong Thesis Collections*, 2021.