A Visual Analytics Approach to Anomaly Detection in Hydrocarbon Reservoir Time Series Data

Aurea Soriano-Vargas[a,*], Rafael Werneck[a], Renato Moura[a], Pedro Mendes Júnior[a], Raphael Prates[a], Manuel Castro[a], Maiara Gonçalves[b], Manzur Hossain[b], Marcelo Zampieri[b], Alexandre Ferreira[a], Alessandra Davólio[b], Bernd Hamann[c], Denis José Schiozer[b] and Anderson Rocha[a]

[a]*Institute of Computing, University of Campinas – UNICAMP, 13083-852 Campinas, SP, Brazil*
[b]*Center for Petroleum Studies(CEPETRO), University of Campinas – UNICAMP, 13083-970 Campinas, SP, Brazil*
[c]*Department of Computer Science, University of California, Davis, CA 95616, U.S.A.*

ARTICLE INFO

Keywords:
hydrocarbon reservoir
anomaly detection
time series
visual analytics

ABSTRACT

Detecting anomalies in time series data of hydrocarbon reservoir production is crucially important. Anomalies can result for different reasons: gross errors, system availability, human intervention, or abrupt changes in the series. They must be identified due to their potential to alter the series correlation, influence data-driven forecast, and affect classification results. We have developed a visual analytics approach based on an interactive visualization of time series data involving machine learning approaches for anomaly identification. Our methods rely upon a z-score normalization technique along with isolation forests. The methods leverage the prior probability of anomalies from a time-window, do not require labeled training data with normal and abnormal conditions, and incorporate specialist knowledge in the exploration process. We apply, evaluate, and discuss the methods' capability using a benchmark data set (UNISIM-II-M-CO) and real field data in three visual exploration setups. The ground-truth annotations were done by human specialists and considered different interventions in the reservoir. Our methods detect approximately 95% of the human intervention anomalies, and about 82%-89% detection rate for other anomalies identified during data exploration.

1. Introduction

Nowadays, time series of relevant sequential data is generated continuously in various domains, from environmental and natural phenomena monitoring to financial markets to population statistics (Bernard et al., 2012) to hydrocarbon reservoir monitoring. In general, data analysis of time series requires the ability to explore the variables thoroughly, intending to identify patterns, analyzing their behavior (Steed et al., 2017), and making sense of long multivariate time series (Bernard et al., 2012). Specifically, hydrocarbon reservoir management collects massive sequential data measured by specific equipment. Subsequently, it is analyzed by specialists to verify the reliability of the data on representing the real reservoir conditions and make decisions based on the possible findings.

However, as with every time series, hydrocarbon reservoir data may be subject to unexpected or even uncontrollable events. These events can lead to erroneous observations that are somehow inconsistent with the other observations in the series. Typically, these observations are called outliers, anomalies, aberrant values, atypical data, and discrepant observations.

In many situations, the process of identifying unusual observations that could be generated by unexpected behavior is critical. Such undesirable behavior may be due to any problems that the reservoir or the registration process may be experiencing. Outliers may occur for different reasons: gross errors, human interventions, or abrupt changes in the series. Gross errors are defective observations such as measurement, recording, and typing errors. Human interventions refer to any operation carried out on a well during or at the end of its productive life that alters the well state. Therefore,

*Corresponding author: Aurea Soriano-Vargas. All authors contributed to important aspects of the work and to the research presented, as well as to manuscript preparation. ASV: code development and manuscript preparation; ASV, RW, RM, PMJ, RP, MC, MH, AF: research decisions, coding decisions, and code testing for further tasks; MZ, MG, MH, AD: generation of simulated case data and interpretation of results; DS: supervision of petroleum result generation; BH: contributions to development of visual analysis methods; AR: study planning and coordination of all research aspects. ASV drafted the manuscript, and the other co-authors reviewed and provided revisions for the drafts.

aurea.soriano@ic.unicamp.br (A. Soriano-Vargas)
ORCID(s): 0000-0002-8429-4119 (A. Soriano-Vargas)

inferring these outliers aim to provide well diagnostics or manage the production of the well. These two types should be adequately identified whenever possible due to their potential to alter the series correlation, influencing forecast, and classification results. Because of the large amount of data, the specialist's manual identification could delay identifying irregularities and generate financial losses.

Exploring time series data for identifying anomalies is particularly challenging. Typically, these data contain hundreds, thousands, or even millions of instances; analysis may be conducted with limited prior knowledge; the definition of what is a normal behavior in the series can be complicated; the notion of normal behavior may continue to evolve, and the magnitude of different anomalies may be different. In the reservoir context, a new challenge is added: although we are dealing with a big data problem, anomalies are seldom, lacking annotations, and rare events.

Strategies are necessary to diminish the time invested in the anomaly identification and allow better decision-making processes later on. Most of them were developed for a specific domain, such as identifying fraud in banking and credit operations (Gupta et al., 2020), identifying cardiological problems through analysis of electrocardiogram (ECG) (Pereira and Silveira, 2019), and anomalous events in the stock market analysis (Close et al., 2020). Meanwhile, some consider the entire range of data to establish anomalies without considering that normal behavior may change over time (Tian et al., 2019), often defining parameters automatically without considering specialists' domain knowledge.

Visual Analytics approaches (Habibi and Shirkhodaie, 2012; Sun, 2013) combine Machine Learning and Information Visualization strategies to support processes that require extracting information from data. They have been successfully applied to several domains from different machine learning fronts, based on creating graphical representations to favor understanding the machine learning processes and user knowledge. Thus, our initial hypothesis herein is that exploratory visualization can be successfully applied to identify reservoir data anomalies and understand how the parameters may affect the identification result.

We developed a visual analytics approach based on interactive visualizations of time series connected with machine learning approaches to anomaly identification. We explored different anomaly detection techniques to discover patterns that do not behave as expected. For quantitative analysis, we use a simulated dataset annotated by specialists with interventions related to partial and complete wells closures. The best results were obtained with an approach using a z-score (Yadav et al., 2018) formulation allied with Isolation Forests (Liu et al., 2008). Our approach considers the prior probability of anomalies from a time-window, does not require any labeled training data with normal and abnormal conditions and includes specialist knowledge in the exploration process.

We organized the paper into five sections. In Section 2, we discuss related literature addressing four approaches — Statistical, Supervised, Unsupervised, and Information Visualization — to anomaly detection. We define anomaly detection and explain different anomaly types in the reservoir context in Section 3. We present the proposed pipeline for anomaly detection in Section 4 while Section 5 discusses the adopted datasets and illustrates the application of the *Visual Analytics* framework with two case-studies. Section 6 summarizes our main results and contributions and points to possible future research.

2. Related Work

After extensive research on published works to detect anomalies in multiple time series from different domains, we can roughly subdivide the prior art into four categories.

Statistical methods rely upon past measurements to approximate a correct behavior data model (Jung et al., 2015). Whenever a new measurement is recorded, it is compared with the model and, if it is statistically incompatible with it, it is marked as an anomaly. A window-based approach typically aids in reducing the number of false positives (Yu et al., 2014). An example of a widespread statistical anomaly detection method is the so-called *low-high pass filter*, which classifies values as anomalies based on how different they are from the moving average of past measurements. Other strategies are based on probabilistic models with the same idea but encoding relations between measurements through time, using Bayesian Networks (Hill et al., 2007) or Hidden Markov Models (Görnitz et al., 2015). However, these methods do not scale well and are computationally intensive.

Supervised Anomaly Detection methods are related to classification tasks and typically expect a training data set in which cases have been marked as normal or abnormal classes (Elghanuni et al., 2019; Sommer and Paxson, 2010). Theoretically, supervised methods provide a better detection rate than unsupervised ones as they have access to more information (specialist annotations). For instance, Fisher et al. (2017) described experiments with two-class and one-class support vector machines (SVMs). However, some technical issues make these methods not appear as accurate as they are supposed to. One of the main problems presented in a typical reservoir is the lack of consistent training

data with enough annotations. Also, obtaining accurate labels is challenging, and training sets generally contain noise resulting in higher false alarm rates.

Another way to apply supervised methods is through regression context modeling. Past measurements are used to train a model that can predict the value of the next measurement in the time series data. If the predicted data is too different from the actual data, it is labeled an anomaly. Different regression-based algorithms were proposed for anomaly detection in this context ranging from simple linear methods to complex ones such as Deep Neural Networks (DNN) with Long Short-Term Memory (LSTM) cells (Malhotra et al., 2015; Zhong et al., 2019). The main concern with these methods is establishing the threshold of difference between the predicted and the actual values, i.e., when a different value has to be considered an anomaly.

Generative models are also used to detect anomalies, where data distribution is learnt to generate additional data points. Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) are used for this purpose. Learning input distributions can be utilized for Anomaly Detection based on GANs, see proposed (Zenati et al., 2018). They effectively identify anomalies in high-dimensional and complex datasets. However, traditional methods such as K-nearest neighbors (KNN) perform better in scenarios with fewer anomalies (Škvára et al., 2018). Moreover, models must be continuously retrained since the concept of normality may change over time.

Unsupervised Anomaly Detection methods do not need to bother with labeled data for the training. We can find two sub-approaches in this case: proximity-based methods and clustering-based methods. The proximity-based methods rely upon distances between data points to distinguish between normal and abnormal data. Local Outlier Factor (Breunig et al., 2000; Barbariol et al., 2019) assigns an outlier score to each new measurement based on the density of measurements around its k-nearest neighbors and the density of measurements around the new measurement. Measurements with high outlier scores are labeled as anomalies. Clustering-based methods, in turn, comprise a subset of proximity-based algorithms, in which past values are used to create clusters. Then, new measurements assigned to isolated and small clusters, or measurements very far from their centroids, are labeled anomalies (He et al., 2003). Similarly, Isolation-based methods use space partitioning. The Isolation Forest is often used (Liu et al., 2008; Barbariol et al., 2019), partitioning space based on random choices of variables and splitting points. The process is performed until the observation being examined is isolated. The idea underlying such methods is that normal cases are more common than anomalies in the data distribution (Chimphlee et al., 2007). If this affirmation is not valid, then the methods suffer from a high false alarm rate.

Notwithstanding, both supervised and unsupervised strategies have the problem of insufficient user engagement. Users are not involved in the learning process and, consequently, the model cannot be improved or adjusted according to user experience and needs. Besides, the interpretability may be insufficient because of the lack of graphical representations.

Visual analytics approaches to identifying anomalies in hydrocarbon reservoir data have not received significant attention. Most reservoir data visualizations are static representations using line charts, scatterplots, and others. For instance, Stoffel et al. (2013) present visual analytics for anomaly detection in computer networks, based on the perception of similarities between vertically oriented line charts compared with a reference model of the data. Shi et al. (2011) presented a sensor anomaly visualization approach that uses graph visualizations to perceive network failures and faults for the user diagnosis of anomalies.

Suschnigg et al. (2020) present a visual analytics approach to anomaly detection of industrial time-series data. This approach is based on a glyph representation to visualize anomaly scores of cycles. This work is only applicable for cyclic (also periodic or seasonal) data, a common characteristic in many industrial applications. An application for anomaly detection in buildings' power consumption has been proposed by Janetzko et al. (2014). It proposed a similarity-based anomaly score illustrated in several visualization techniques such as recursive patterns, spiral graphs, and line charts. In the work of Wu et al. (2018), anomalies are detected for equipment condition monitoring in smart factories of the process industry by a model-based approach. The deviation of estimated and real values is visualized in a river plot view. Kalamaras et al. (2017) introduced an interactive visual system to explore historical data and predict future traffic; this system supports the detection of anomalies. The Local Outlier Factor (LOF) is used for different roads and different periods. The degree of being an outlier is determined by sparsity, relative to the same roads and historical behavior.

Our hypothesis in this work is that such visual analytics methods hold potential for anomaly detection in reservoir data; we propose exploring this potential in the next sections.

3. Data Representation and Anomalies Definition

In this work, we adopt a similar definition of Multivariate Time Series to those used by Soriano-Vargas et al. (2019); Vargas et al. (2019). Multivariate time series (MTS) data consists of n time-stamped observations ($\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, ..., \mathbf{x_n}\} \forall\ n \in \mathbb{Z}$ and $n \geq 1$) of variables, recorded at a particular temporal scale (per minute, hour, day, month or year).

Each MTS i ($\{\mathbf{x_i}\}$), that represents a variable, is an ordered temporal sequence of p observations taken at different times t. It can be described as:

$$\mathbf{x_i} = \{x_i^{t_1}, x_i^{t_2}, x_i^{t_3}, ..., x_i^{t_p}\} \tag{1}$$

A data instance at time t_j can be represented as a vector $\mathbf{i^{(t_j)}}$ with k values, which are related to the k selected variables ($k \leq n$):

$$\mathbf{instance^{(t_j)}} = [x_1^{t_j}, x_2^{t_j}, x_3^{t_j}, ..., x_k^{t_j}]. \tag{2}$$

Therefore a multiple time series data set is defined by a time series describing multiple variables, see Eq. 3. It can be conceived as a matrix, where each row corresponds to the time series relative to a particular variable, see Eq.1, where each column corresponds to a multivariate observation at a particular timestamp, i.e., a multidimensional data instance, see Eq. 2.

$$\mathbf{D} = \begin{bmatrix} x_1^{t_{initial}} & x_1^{t_{initial+1}} & x_1^{t_{initial+2}} & \dots & x_1^{t_p} \\ x_2^{t_{initial}} & x_2^{t_{initial+1}} & x_2^{t_{initial+2}} & \dots & x_2^{t_p} \\ x_3^{t_{initial}} & x_3^{t_{initial+1}} & x_3^{t_{initial+2}} & \dots & x_3^{t_p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_k^{t_{initial}} & x_k^{t_{initial+1}} & x_k^{t_{initial+2}} & \dots & x_k^{t_p} \end{bmatrix} \tag{3}$$

Anomalies or outliers are particular values ($x_i^{t_j}$) that deviate from observations on data (see Figure 1), which may indicate measurement variability, an entry/experimental error, or a human intervention. An outlier can come in a group ($x_i^{t_j}, x_i^{t_{j+1}}, x_i^{t_{j+2}}$) as well and not only as an individual ($x_i^{t_j}$) as Figure 1 depicts.

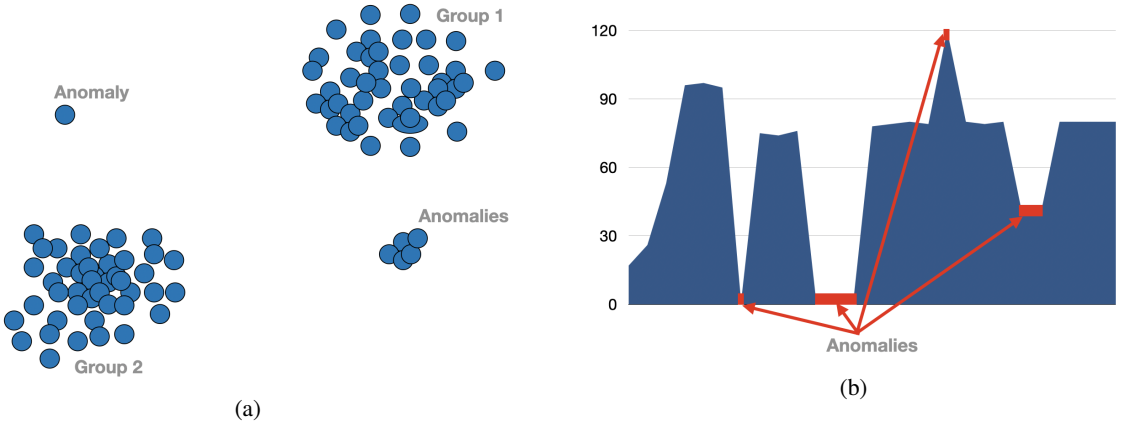
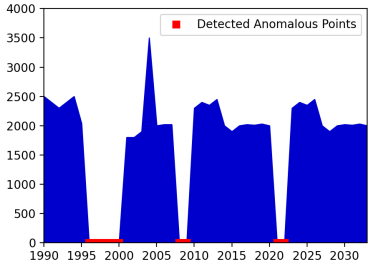


Figure 1: Anomalies are values out of the "normal" that may arise individually or in groups in a dataset (a). This definition can also be applied in time series data, where specific anomalies may appear in a time instant or interval (b).

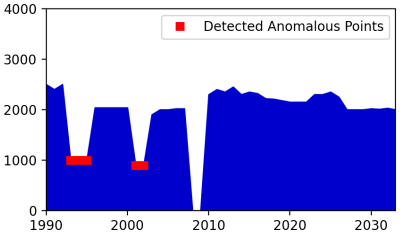
After many discussions with reservoir experts, we found the need to analyze time-series data to identify situations in which there was some unexpected or abnormal behavior. Such behavior can be caused by equipment malfunction,

poorly executed maintenance, human interventions, and others. Regardless of the cause, its identification is essential to ensure that the monitoring process remains adequate to represent the reservoir's real condition and improve data predictions.

The most common problems found in the reservoir data are characterized by the absence of data (presence of zeros, nulls, or not-available entries) and rapid and temporary changes in the values level (valleys and peaks). The absence of data can be associated with a failure to acquire or record data or human interventions. Typically, in real data, reliable annotations are not usually found. In these cases, we must detect and disregard this data as it impacts different algorithms, i.e., correlation and forecast analyses, resulting in skewed and misleading results. For instance, the quality of the historical data directly affects the quality of the forecasting algorithms. In this sense, we are interested in identifying the absence of data in short and long intervals related to possible failures, partial or total closure of the well. Figure 2a shows an example of complete closure.



(a) Complete closure highlighted in red.



(b) Partial closure highlighted in red.

Figure 2: Examples of complete and partial closure observed in the time series related to the liquid rate of a certain well.

Valleys in fluid rate time series are observations that differ from the values of nearby measurements and are related to a partial closure of the well. They begin with a negative slope and finish with a positive slope without assuming a zero value in the interval, as illustrated in Figure 2b. Besides, we are interested in detecting peaks – data with a high positive slope – as illustrated by the top blue squares in Figure 3. These peaks are related to possible failures of capacity control or the wells reopening after some closure.

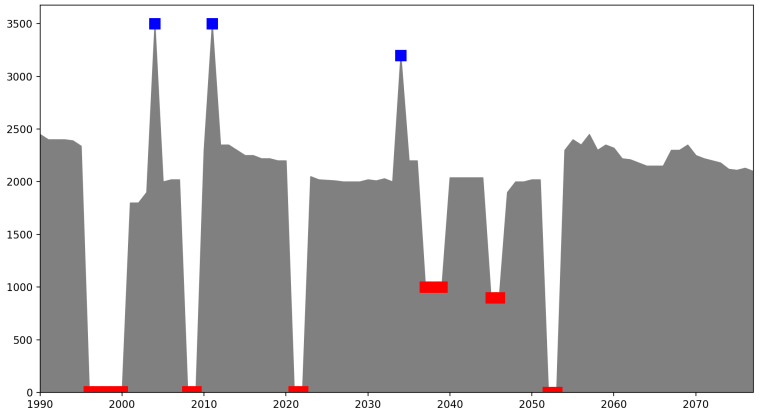


Figure 3: Different anomalies in daily production oil. Peaks (data with a high positive slope) are represented by blue squares, and valleys (data with a high negative slope) by red squares.

4. Visual Analytics Approach

In this work, we adopt hydrocarbon reservoir time series, which contains mainly production data (oil or liquid rates). In such data observations, we identified critical requirements that guided the rationale of our studies.

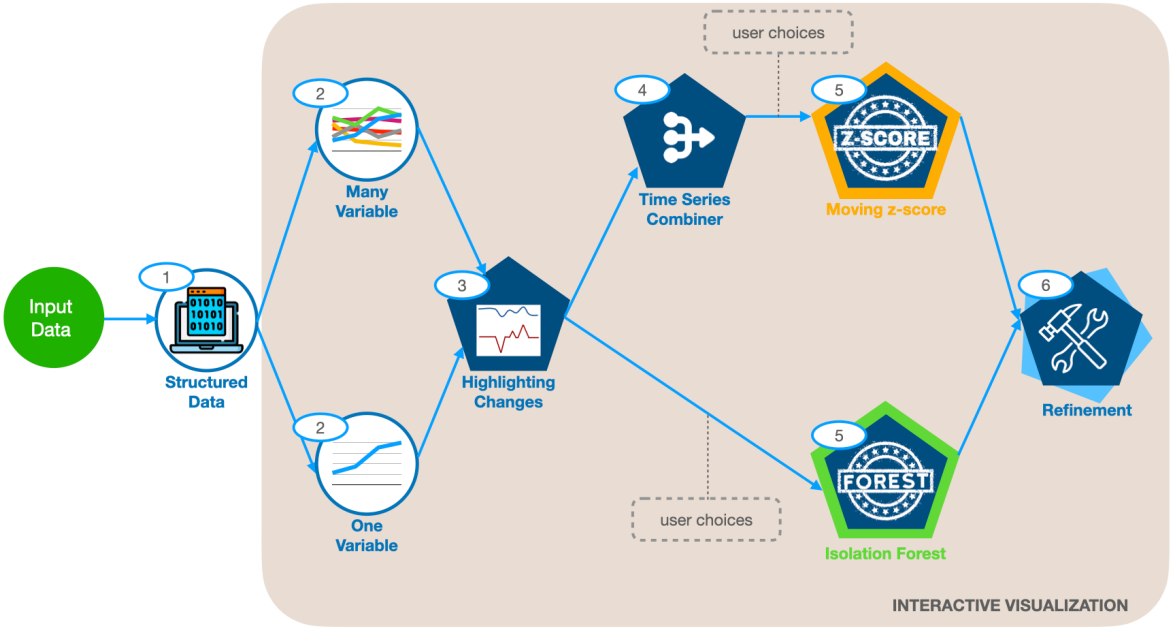


Figure 4: Overview of the proposed interactive anomaly detection process. From the data selected from one well or a set of wells (producers or injectors) **(Stage 1)**, the user must initially specify the variables (or single variable) **(Stage 2)** that will be considered in the anomaly detection process. Once we have filtered the data points, the approach includes highlighting the changes **(Stage 3)**. Then, we can work with all variables without a projection (the entry for anomaly strategies that work with more than one variable) and with a projection (for those strategies that work with just one variable) **(Stage 4)**. In **Stage 5**, the anomaly detection strategies are applied to the modified (or not) time series. Finally, in a post-processing step **(Stage 6)**, values that fell outside the decision cut-off but whose score is very close to the limit and whose two contiguous neighbors were detected as anomalies are analyzed. These steps are encapsulated in an interactive visualization process.

- **Human in the loop**: We do not have access to sufficient annotated data; consequently, supervised models no longer apply. We need to include visualizations of data distribution and behavior, bringing human experts back to the decision process. For this reason, any algorithms and visualizations should be tightly integrated.
- **Moving time-window**: Since normality within a reservoir can evolve, anomaly calculations need to consider a time window amenable to dynamic updates. By this means, the costs of retraining the entire model and user labeling can be avoided, and the history data can also be utilized.
- **Anomaly situation awareness**: The results of identifying anomalies must be shown along with variables used in a decision so that specialists can estimate the reason for the presence of those anomalies.

As we have daily rates, if a well is closed for one or more days, it will be easy to detect it, since rates tend to zero. However, when we have partial closures for a few hours, for example, there will be a drop in production rates. It is difficult to differentiate whether such a drop is related to human intervention, such as operational management or well/platform problems, or a consequence of reservoir behavior, such as kicks, reservoir pressure drop or changes in GLR. We consider data of anomalies caused by human intervention, annotated in the simulated data set by specialists. We focus on the identification of abrupt changes. The sooner these anomalies are detected faster can be the decision making of the operator.

Considering the outlined motivations, Figure 4 presents an overview of our proposed approach, comprising six stages. From the data selected from one well, or a set of wells (producers or injectors) **(Stage 1)**, the user must initially specify the variables (or single variable) ($\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, ..., \mathbf{x_p}\} \forall\ p \leq n$) **(Stage 2)** to be considered in the anomaly detection. The formulation we propose is robust to different configurations (single or multiple).

Once we have filtered the well's variables, the approach includes a step to highlight the changes (**Stage 3** in Figure 4). The basic idea is to increase the value of the time instants, where changes occur, as Figure 5a illustrates. For this purpose, we include first derivatives using finite difference approximations applied to the sequential values:

$$\hat{\mathbf{x}}_\mathbf{i} = \frac{\mathbf{x_i(j+h) - x_i(j)}}{\mathbf{h}}, \tag{4}$$

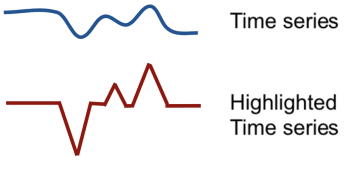
where $h = 1$, representing the change for one day, and j is the index of the time series data of variable x_i.

Our strategies depend on the exploration needs of specialists relative to the number of time series used, since some strategies only work for one time series. In this case, multivariate time series data is combined into one data set. Then, we can work with all variables without a projection (the entry for anomaly strategies that work with more than one variable) and with a projection (for those strategies that work with just one variable) (**Stage 4** in Figure 4). For this purpose, we performed experimental tests with different projection techniques and combined strategies, from which we highlighted the results with the method proposed by Keogh and Pazzani (1998), given the same influence factor for all variables.

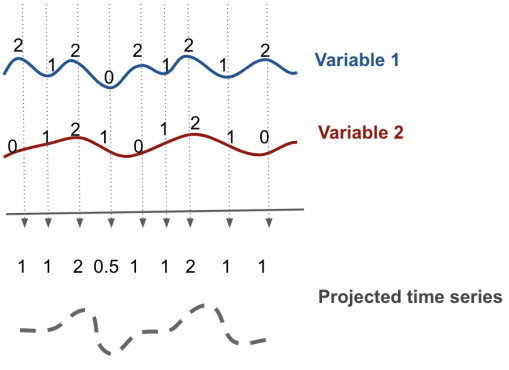
Keogh and Pazzani (1998) proposed a merge operator to combine information from two time series and repeated application of the merge operator that allows a combination of information from time series. We find a combined value of a set of variables at each instant of time, as Figure 5b depicts. Certainly, detected anomalies in different time series is beneficial for a more effective process for causal explanation. At this time, we do not have access to more types of anomalies, and the combination of time series turned out to be sufficient and more efficient.

In **Stage 5** of Figure 4, two anomaly detection strategies are applied to the different time series. We use the anomaly detection strategies to analyze each data point considering prior time-window data. Our approach provides default parameters and time-window size. However, these can be modified through the interactive visualization process with the participation of a human expert.

Finally, in a post-processing step **(Stage 6)**, values that fell outside the cut-off but whose score is very close to the limit and whose two contiguous neighbors were detected anomalies are analyzed. All of these steps are involved in an interactive visualization process, whose visualizations are described below.



(a) High-lighting significant changes in the time series.



(b) Combining variables into one time series.

Figure 5: Strategies applied to times series in pre-processing step. (a) High-lighted: result of emphasizing differences using first derivative. The original time series is transformed as follows: Areas without many transitions are smoothed out; areas with many variations are exaggerated. (b) Merging: Application of merging strategy proposed by Keogh and Pazzani (1998). Some strategies only work with one time series. To satisfy this restriction multivariate time series are combined into one.

4.1. Anomaly Detection using Z-score

A typical assumption in large datasets is that the samples' populations tend to be normally distributed. In this context, the z-score is a common metric to calculate the standard deviations that the value of a variable is from the mean. Based on that idea, there is a straightforward statistical method to detect anomalies data using this metric and cut-offs from variable data.

However, the assumption of normality is critical when we analyze production data since values may range from large to small orders of magnitude over time, i.e., normality can fluctuate. In this sense, we established a window size to analyze the data z-score. Moving z-score (Yadav et al., 2018) is a model for measuring each data point's anomalousness in a time series. Given a time interval, the Moving Z-Score is the number of standard deviations. Each observation is away from the mean, where the mean and standard deviation are computed only over the previous interval observations.

In this strategy, we consider two cut-off decisions, one positive and one negative, which can also be defined by the user, according to the data distribution. Re-calculations are performed in real time. The two cut-off decisions are related to two possible behavior, positive when the outlier presents a high value and negative when the outlier presents a low value, compared to the average value. Through these two cut-off decisions, we can configure our exploration according to our needs.

There are situations in which we want to analyze the anomalies presented, considering more than one variable. Since these concepts can be applied only to one-variable data, we explore projection concepts applied to time series. The objective is to project many time series in a representative time series, whose shape is a compromise between the original time series and can convey the reservoir data behavior. First attempts were focused on punctual projections where each instance is considered a point in a multidimensional space, and then we obtained the 1-dimensional projection of each point to build the time series. However, better results were obtained with the strategy proposed by Keogh and Pazzani (1998).

The Keogh merge operator combines information from two time series at each step using a weight vector, which reflects how much corresponding time series segments agree. They also associate a term called influence to each time series to be mixed. In our case, we give the same influence to all time series, i.e., the same contribution to all variables. For more information see (Keogh and Pazzani, 1998).

4.2. Anomaly Detection using Isolation Forest

When we work with more than one variable separately, the Isolation Forest (Liu et al., 2008) option is available and adjusted based on the user-defined time interval. Re-calculations require between 30 and 40 seconds. Since this method originally works with high-dimensional datasets, we do not have to employ any projection or combination strategy.

The Isolation Forest method is an unsupervised learning-based anomaly detection algorithm leveraging the idea that anomalous values are easily separable from the rest of the samples in a dataset (Liu et al., 2008; Barbariol et al., 2019). It splits the dataset by randomly selecting a feature and a threshold value by several decision trees. During this process of isolating points, anomalies are identified as the ones getting isolated in fewer steps. This process can be described in more detail as follows:

Random Partition. The algorithm uses a binary tree structure and recursively creates partitions by randomly selecting a variable and a split value between the variables' ranges, generating isolation tree proxies where anomalies have shorter paths in the tree. The process consists of these steps: a) for each variable, identify the minimum and the maximum value; b) choose a variable randomly; c) choose a random value in the variable range; d) repeat steps b) and c) until the maximum depth is reached. Due to the same normality principle of production data in which the normal concept may vary, we defined a time window from which the algorithm gets the samples to build the Isolation Forest.

Binary Tree. The random partitioning produces noticeably shorter paths for anomalies since fewer instances (anomalies) result in smaller partitions and are more likely to be separated during early-stage partitioning. The result is a binary tree, where each node is either an internal or external node. Internal nodes are non-leaf and contain the proxies to evaluate a variable, given a split value and a split variable, and have two child sub-trees. External nodes are leaf nodes that hold the size of the un-built subtree (number of resting values). The process is repeated until one reaches the maximum tree depth, set to $\lceil log_2(n) \rceil$, where n is the number of samples used to build the tree Liu et al. (2008), i.e., the number of instances. As we work with time intervals, we are not performing sub-sampling. We apply the same random process to generate 100 isolation trees. Once the iterations have terminated, we generate an anomaly score for each point.

Anomaly Score. The split number determines the isolation level. For a point x, we determine the number of edges that x traverses in the Isolation Tree, called $h(x)$. As discussed in the original proposal Liu et al. (2008), the average path length is calculated based on an unsuccessful search in Binary Search Trees, defined as:

$$c(n) = 2H(n-1) - (2(n-1)/n), \tag{5}$$

where n is the number of instances, $H(i)$ the harmonic number (estimated using the Euler's constant).

The anomaly score of an instance x is given as:

$$s(x, n) = -2^{-\frac{E(h(x))}{c(n)}}, \tag{6}$$

where $E(h(x))$ is the average of $h(x)$ from a collection of isolation trees, and $c(n)$ is a normalization factor. Differently from the original paper, we use a negative sign for the anomaly score.

When $E(h(x))$ is close to $c(n)$, the score tends to be -0.5, when $E(h(x))$ is close to 0, the score tends to be -1, and when $E(h(x))$ is close to $n-1$, the score tends to be 0. Consequently, the anomaly score is in the interval $[-1, 0]$, where a value closer to -1 is related to outliers, and a value closer to 0 to inliers.

We must define a cut-off value that classifies points as inliers or outliers based on the scores. This cut-off decision can be made by a user, as part of the definitions made in the user interface. The resulting identified anomalies can be seen as a vector of Boolean values. In Figure 6, we illustrate the strategy by showing each step of the anomaly verification with Isolation Forest of the green star value. In addition, the detailed algorithm is provided in Appendix A in Algorithm 1.

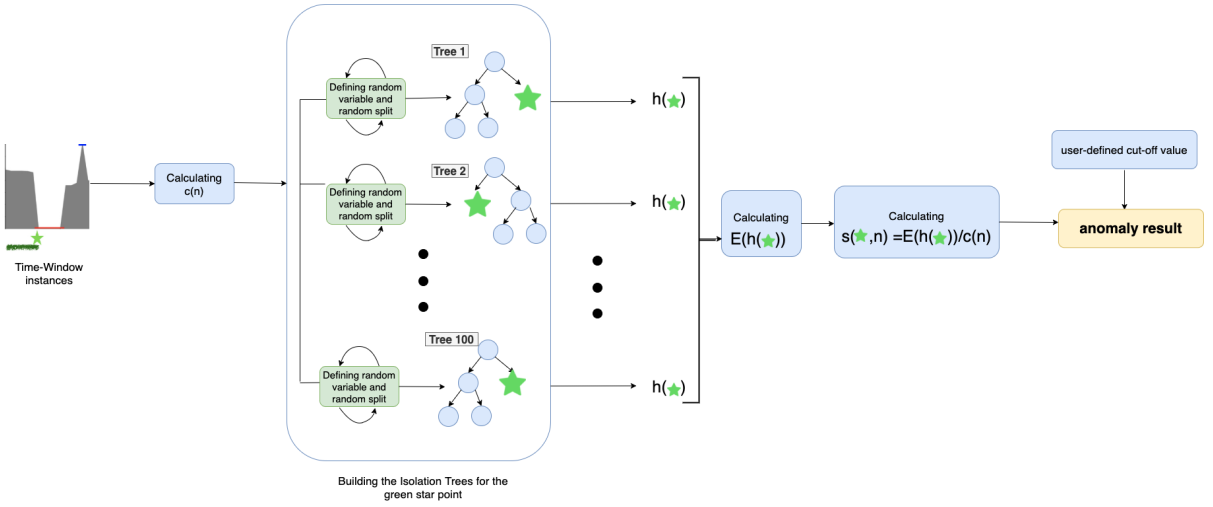


Figure 6: Isolation Forest strategy applied to time series data. Each instance is analyzed using data from the previous time window, using a user-defined step size. The strategy produces a score between -1 and 0, where a value closer to -1 is related to outliers and a value closer to 0 to inliers. A user-defined cut-off value determines inliers or outliers based on the scores.

4.3. Variable Visualization

We worked with petroleum experts in the team to identify the most critical issues for anomaly detection. The specifications suggest that our analyses have to consider (i) the temporal behavior of variables, (ii) the visual identification of the anomaly result for one well, (iii) the visual interplay of variable behavior – to provide intuitive visualizations supporting hypothesis formulation for potential causes or correlations–, and (iv) the comparison among wells. Hence, it is essential to capture the multivariate nature and temporal behavior while preserving detailed information. Besides, it is crucial to have enough display space to visualize details over time. Consequently, our goal is to propose and validate novel anomaly detection approaches that consider these requirements and incorporate interaction mechanisms to support diverse user-driven anomaly analysis tasks.

Our Visual Analytics Tool combines visualizations that convey the temporal behavior of individual target variables with anomaly detection strategies to (i) provide data analysts with a visual exploration loop that supports the inspection of the multivariate values recorded; (ii) show how the behaviors of multiple variables are related employing a small multiples visualization (van den Elzen and van Wijk, 2013); and (iii) allow interactive exploration of the anomaly detection results.

The initial view is illustrated in Figure 7. The **Choices** window includes all settings and options described in the previous section. The **Distribution** window shows the data distribution from which it is possible to select better parameters. The **Statistics** window includes statistics about partial or total closures of the selected well considering the user's selected variables. The **Main** window conveys the original data in a green area, the modified time series (projected or highlighted) in a dotted-gray line, and the detected anomalies with circles.

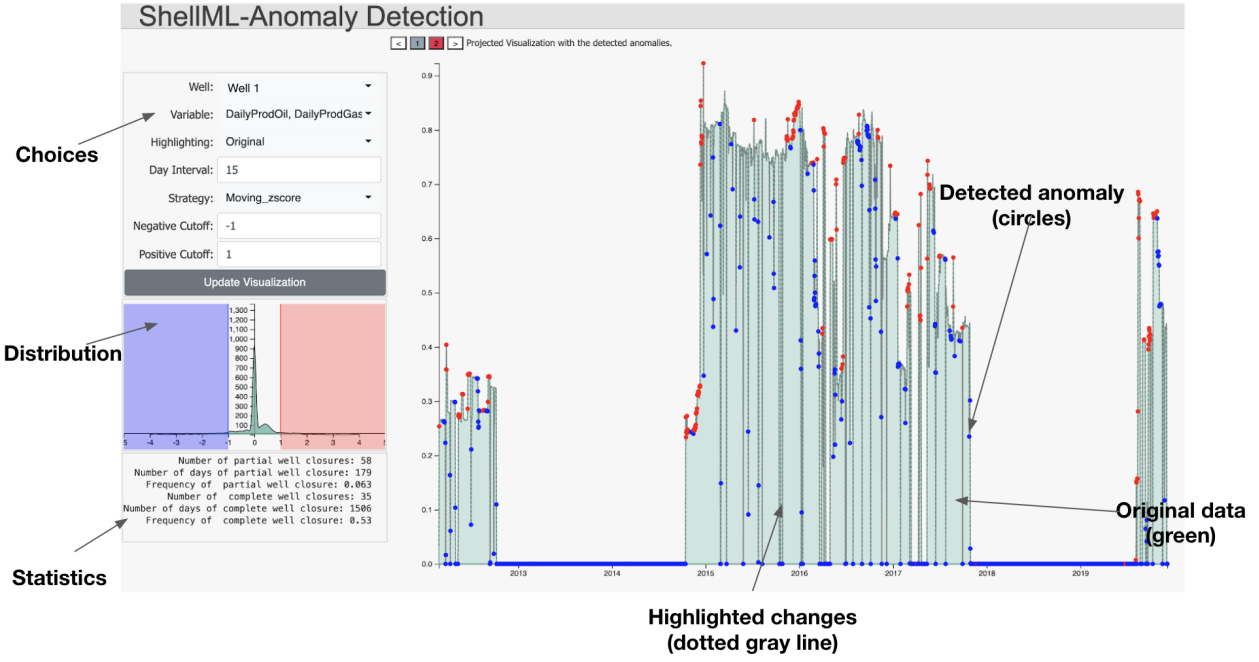


Figure 7: Initial view – anomaly detection for one variable. The **Choices** window includes all settings and options which are defined by the user. The **Distribution** window shows the data distribution from which it is possible to select better parameters. The **Statistics** window includes statistics about partial or total closures of the selected well considering the variables selected by the user. The **Main** window conveys the original data in a green area, the modified time series (projected or highlighted) in a dotted-gray line, and the detected anomalies with circles.

Also, by selecting the red rectangle (2) in the **Main** window at the top center, we have access to an alternative initial view based on small multiples visualization (Figure 8), in which the projected data (in a gray area), the highlighted data (in a dotted gray line) and the different variables used (green for oil, red for gas, blue for water and yellow for BHP) are shown. The name of each variable is also shown on the left side of each visualization.

When we select a set of wells (either producers or injectors), we have access to new visualizations related to partial or total closure statistics considering the selected variables. A large visualization is also based on small multiples of the considered variables for all the wells. The views contain:

1. Number of partial closures.
2. Number of days of partial well closures.
3. Frequency of partial well closures.
4. Number of complete well closures.
5. Number of days of complete well closures.
6. Frequency of complete well closures.
7. The considered variables for the set of wells.

Figure 9a shows an example of the statistics regarding the number of days of partial closure for all producers. Each producer's name is on the left side of the visualization, accompanied by a horizontal bar in proportion with other wells and a specific statistic value. Figure 9b illustrates the two variables for each well, which were considered in the computation of the statistics. Green represents the oil rate and red for the gas rate. The name of each well and variable is placed on the left side of the visualization. Through this small multiples visualization, it is possible to recognize the variables' behavior at different time instants.

Figure 8: Alternative Initial view - multiple variable visualization. The projected data is shown in a gray area. The highlighted data in a dotted gray line and the different variables used (green for oil, red for gas, blue for water and yellow for BHP).

5. Experiments and Results

We present two case-studies for validation of the proposed strategies: (a) a controlled experiment with simulated reservoir data from a benchmark model and (b) real reservoir data without annotations of human-interventions. We evaluated our approach based on two criteria: recall and accuracy for the identified anomalies related to human interventions.

5.1. Evaluation Metrics

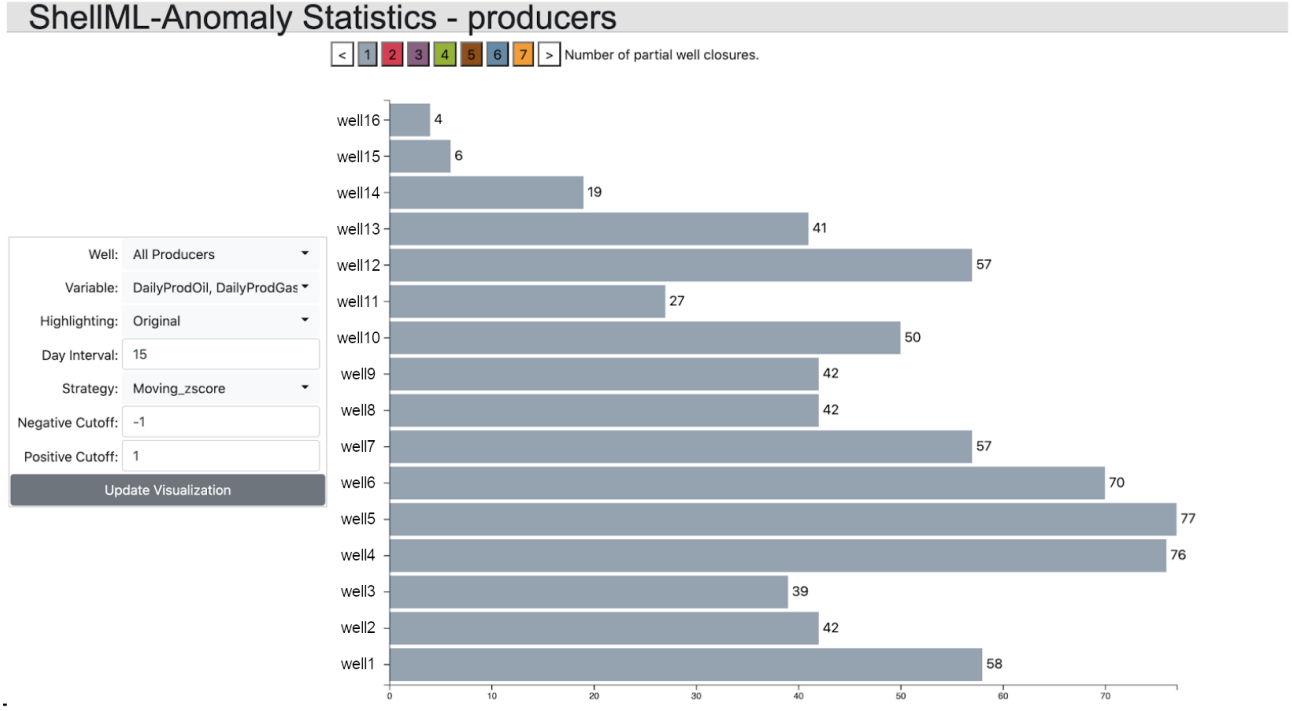
We analyze the results using two metrics: Recall related to the fraction of anomalies from the ground-truth that were successfully identified (Equation 7), and Classification Accuracy related to the fraction of corrected anomalies and no anomalies identified by our approach (Equation 8).

$$\textbf{Recall} = \frac{T_{Anom}}{T_{Anom} + F_{Norm}}, \tag{7}$$

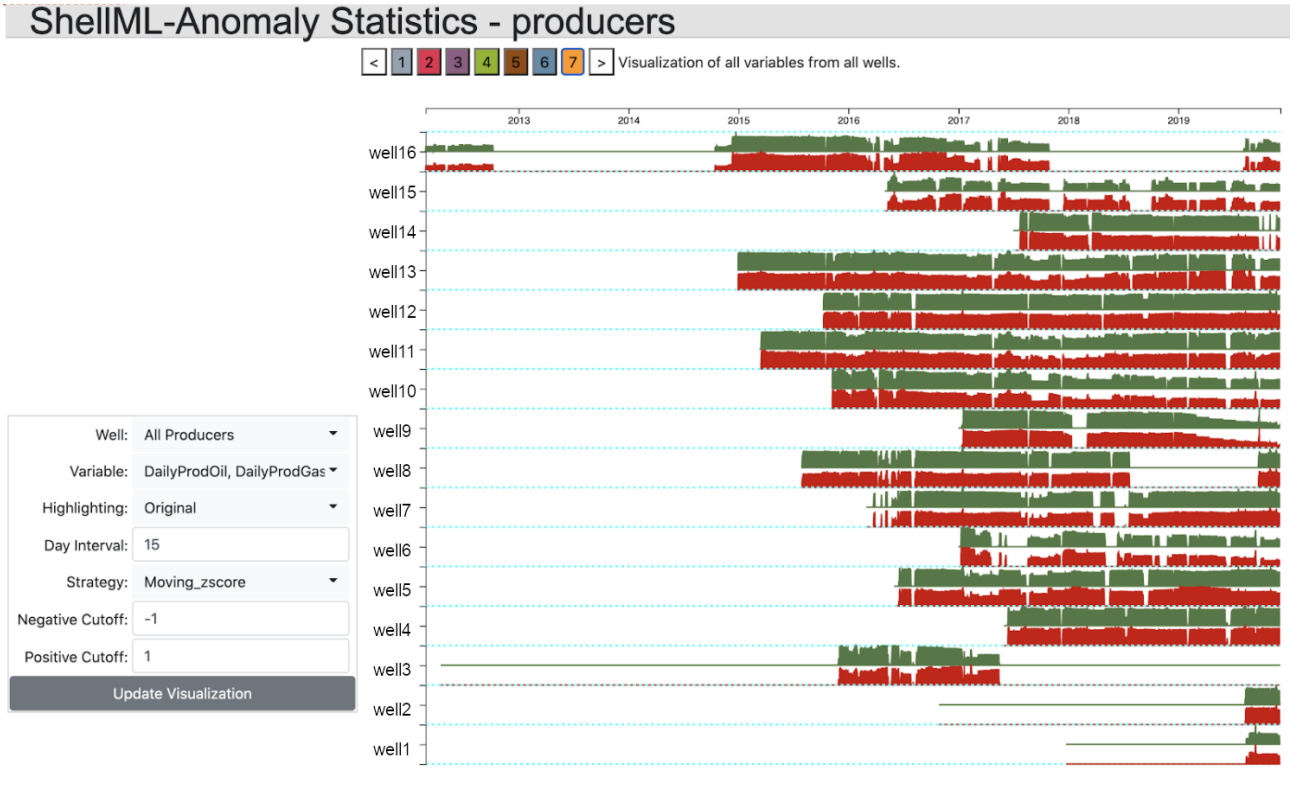
$$\textbf{Accuracy} = \frac{T_{Anom} + T_{Norm}}{T_{Anom} + F_{Anom} + T_{Norm} + F_{Norm}}, \tag{8}$$

where T_{Anom} is the number of actual anomalies identified as anomalies. T_{Norm} is the number of actual normal values identified as normal values. F_{Anom} is the number of normal values identified as anomalies F_{Norm} is the number of anomalies identified as normal values.

We do not use precision to analyze the results. For our experiments, we considered data from one of the possible anomalies related to human interventions, which were annotated by specialists in the data set. However, there are many other anomalies that a reservoir may present, which had not been annotated but could still be detected with our strategies. The precision considers all detected interventions, including the ones without annotations. Many of them



(a) Number of days of partial closure for all producers.



(b) Variables of each well can be displayed.

Figure 9: Visual Analytics for all producer wells considered in the statistics.

Table 1
Results of Recall and Accuracy for the sequential steps applied for all producer wells.

	Moving z-score (seq. process 1)		Moving z-score (seq. process 2)		Isolation Forest (seq. process 3)	
	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy
PRK014	0.95	0.83	0.95	0.82	0.96	0.77
PRK028	0.91	0.95	0.89	0.95	0.9	0.86
PRK045	0.96	0.96	0.96	0.92	0.96	0.90
PRK052	1.0	0.7	1.0	0.71	1.0	0.67
PRK060	0.92	0.94	0.91	0.95	0.95	0.9
PRK061	0.96	0.96	0.96	0.96	0.96	0.88
PRK083	0.98	0.87	0.98	0.88	0.97	0.77
PRK084	0.92	0.96	0.92	0.95	0.93	0.87
PRK085	0.87	0.76	0.87	0.79	0.87	0.69
wildcat	0.96	0.84	0.96	0.99	0.96	0.85
	0.94	**0.88**	**0.94**	**0.89**	**0.95**	**0.82**

might not necessarily be false positives, but anomalies of a type different from human interventions that the system has detected. However, our method already is capable of detecting different types of anomalies, even those not yet annotated and those for which the types of interventions are not known.

5.2. Anomaly Detection in Simulated Reservoir Data

We adopted a benchmark case, UNISIM-II-M-CO, as our base dataset to evaluate our methodology, created by the UNISIM group at the University of Campinas, Brazil. This synthetic reservoir model is based on a real field (Correia et al., 2015). It includes injection and production trends similar to a private field, accounting for the wells' partial and total frequency closure based on the private real field statistics data. The model is a synthetic carbonate light-oil based on a combination of the Pre-Salt characteristics, such as fractures, Super-K layers, and high heterogeneity. The fluid model is compositional, with seven components in the oil phase. The simulation model has six and a half years of history production, containing eight well injectors and ten well producers. We use Daily Production of Oil, Gas, and Water for producer wells from the simulated data variables. Other experiments could use other or more variables.

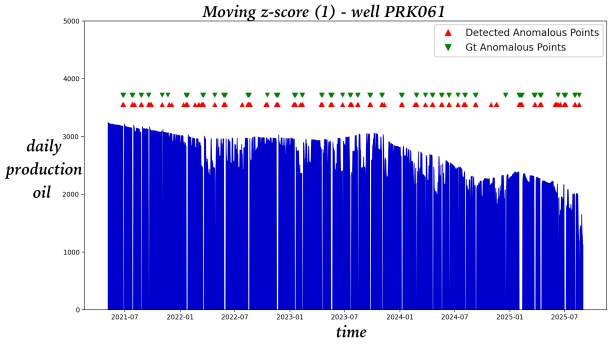
UNISIM-II-M-CO contains annotations related to human interventions, which facilitated the evaluation of our experiments. We applied three sequential steps of our exploration using an interval of 15 days and considering the following information, whose parameters were defined by the visual interactions.

1. **Moving z-score (1):** We considered one variable (Daily Production of Oil), without projection, and highlighting with `diff` (first derivative), by applying the moving z-score strategy. As we are interested only in the valleys, we consider a negative cut-off of -4 and a positive cut-off of 5, which allows us to analyze just the distribution's left part.
2. **Moving z-score (2):** We considered the same previous setup (**Moving z-score (1)**) but considering three variables (Daily Production of Oil, Daily Production of Gas, Daily Production of Water).
3. **Isolation Forest (3):** We considered three variables (Daily Production of Oil, Daily Production of Gas, Daily Production of Water), without projection, and highlighting with `diff` (first derivative). We applied the isolation forest strategy and set the decision cut-off at -0.75.

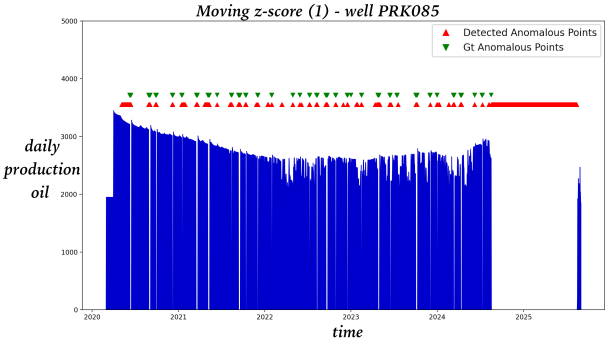
All sequential steps were applied in the ten producers' wells: PRK014, PRK028, PRK045, PRK052, PRK060, PRK061, PRK083, PRK084, PRK085, and Wildcat. The results can be visualized in Table 1. The recall is very high, which tells us that most of the anomalies noted were identified with our strategies. However, the accuracy results indicate several values considered as anomalies that were not found in the ground-truth (as it is only related to human interventions). Therefore, we explored the results for each of the wells and were able to identify anomalies related to well-closure that were not annotated because this ground-truth is related only to human interventions.

In Figure 10, we selected the wells with very good and regular results: PRK061, and PRK085, with the three sequential processes. Variables employed to identify anomalies are shown in blue, resulting from the sequential process in red and the ground-truth in green. The more similar the red line graph is to the green line graph, the more accurate to the ground-truth our strategies are.

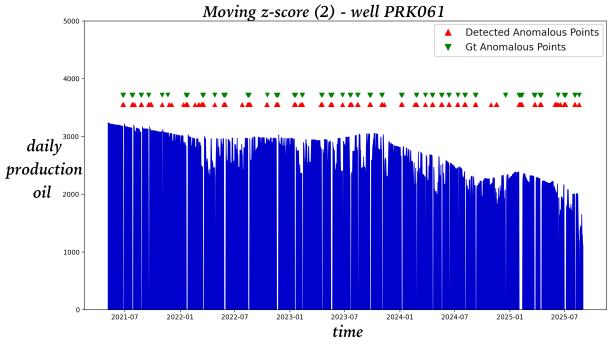
The ground-truth of well PRK061 is very accurate, and our strategies identified all the human interventions plus other anomalies, which may be related to other problems. However, visually we can see a well closure at the final part of PRK085, from which there is no annotation, and our strategies detected it.



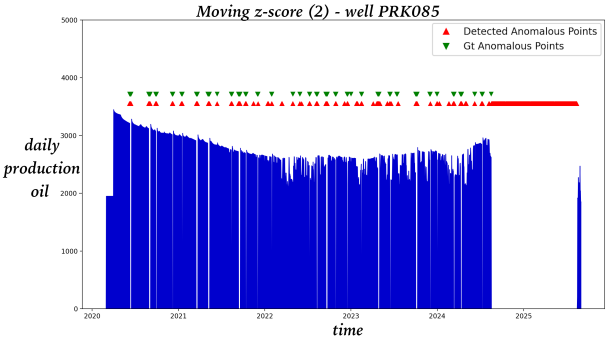
(a) Sequential Process 1 for PRK061.



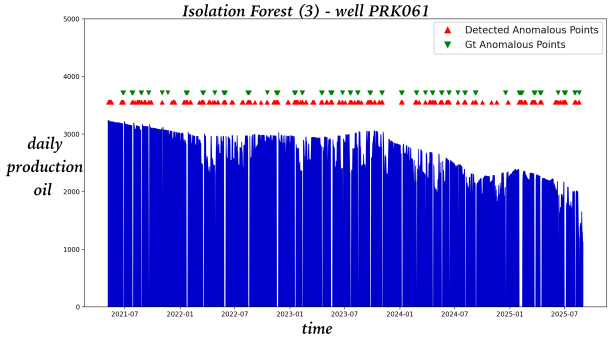
(b) Sequential Process 1 for PRK085.



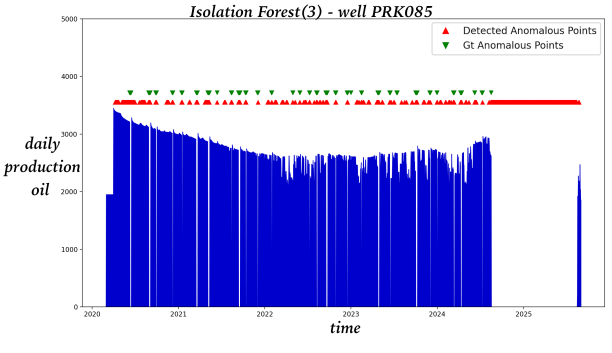
(c) Sequential Process 2 for PRK061.



(d) Sequential Process 2 for PRK085.



(e) Sequential Process 3 for PRK061.



(f) Sequential Process 3 for PRK085.

Figure 10: Variables employed (in blue) to identify anomalies versus the result of the sequential process (in red) and the ground-truth (in green). The ground-truth of well PRK061 is quite accurate, and our strategies identified all the human interventions plus other anomalies, which may be related to other problems. However, visually we can see a well closure at the final part of PRK085, from which there is no annotation, and our strategies detected it. The well wildcat presents some small production data initially. After that, it was affected by a well closure, which was not annotated but was identified for our strategies. For those reasons, the accuracy results are not good because of the imprecision of the ground-truth data annotations.

For space reasons, we include the images of the visual analysis of producer PRK061 with the sequential process

Table 2
Results of Recall for the sequential steps applied for all producer wells in a real data.

	Moving z-score (seq. process 1) Recall	Moving z-score (seq. process 2) Recall	Isolation Forest (seq. process 3) Recall
Entire reservoir	0.67	1.0	1.0

2 (see Figure 11). We can see that our strategy detected all the existing declines in the data. These declines were specified as less than -3. Looking at the data distribution panel, we can see that the selected region consists of a small region away from the distribution hood center. More details could be considered by increasing that cut-off, such as -2 or -1.

5.3. Anomaly Detection in Private Reservoir Data

Complementing our previous analysis, we also evaluate the proposed methods with a private dataset comprising production and reservoir data from a field in Brazil.

This dataset provides information on production rates (oil, gas, and water), pressure (bottom-hole), and the ratio between them (water cut, gas-oil rate, and gas-liquid rate). The reservoir contains 16 producers and 16 injector wells, divided into ten water injectors and 6 WAG (water alternating gas) injectors. For the oldest producer well, we have five years of historical data.

We explored human interventions related to valleys in that data. For this purpose, we seek well intervention annotations in the real dataset, where only four annotations were found related to human interventions. The results are found in Table 2. Due to the shortage of annotations, the accuracy calculation would not be fair since our system would identify many anomalies that were not reported.

If we consider only one variable, the Moving z_score strategy obtained 67% correct answers. With three variables, Moving z_score and Isolation Forest achieved a 100% success rate. Figure 12 illustrates the visual analytics for one of the real wells. In the distribution panel, we identified that the different values are placed below -1. Then, we increased the negative cut-off to -1. The anomalies were detected, which should be analyzed in detail by our specialists.

We also performed experiments on a different scenario. We selected an injector well from the real dataset, which contains two annotations related to interventions (Figure 13). We used the Moving z-score with one variable: daily gas injected volume, and a negative cut-off of -1. We observed that the two annotations were detected by our approach, highlighted in magenta. However, it is possible to see that this well presents many anomalies, including interventions that were not annotated and should be analyzed.

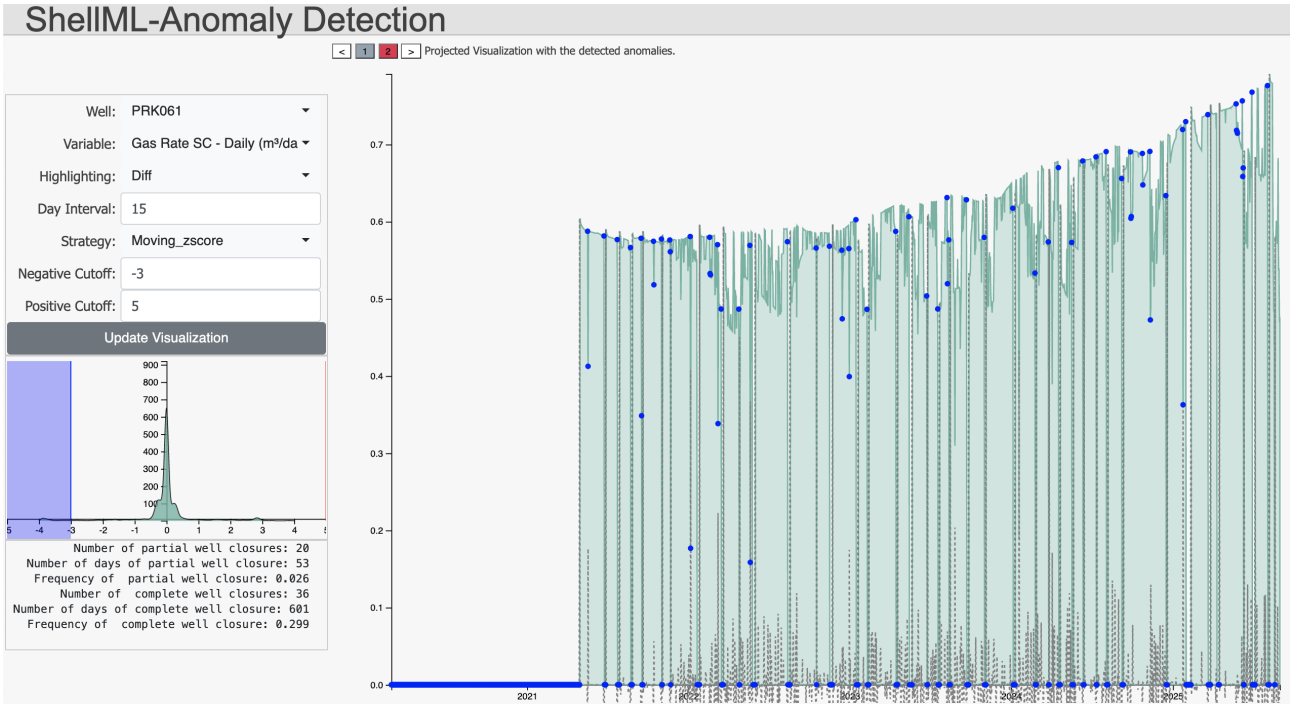
Despite the low number of annotations, our visual analytics loop provides experts with important and helpful knowledge for more effective reservoir management. According to our application domain specialists, this visual analytics tool is of great value for controlling the health of the wells and maintenance of a reservoir, and for detecting problems that one might not notice easily without the tool.

6. Conclusions

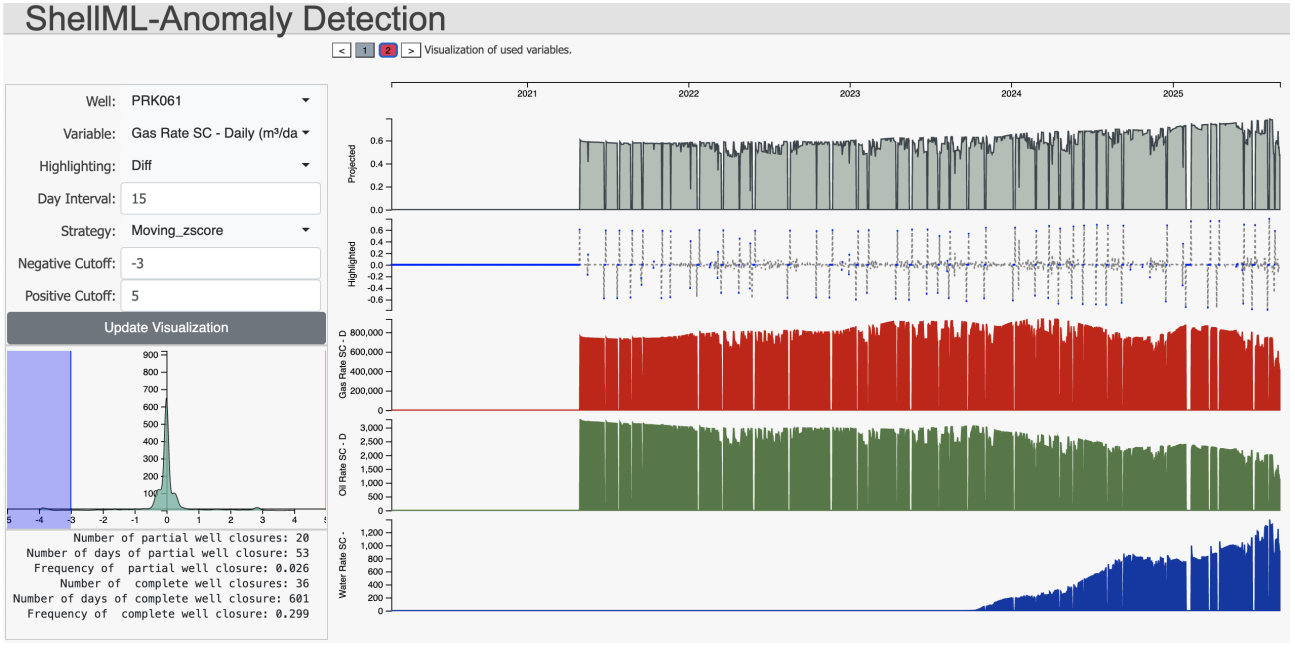
We have introduced a visual analytics approach to anomaly detection for hydrocarbon reservoir time series data. This approach combines powerful machine learning models and human perception. The approach is based on time-window exploration related to visualizations to support anomaly identification.

Human understanding of the data and machine learning-based analysis are mutually enhanced through data visualizations and user interaction. We have applied, evaluated, and discussed our approach for a benchmark data set (UNISIM-II-M-CO) and real field data. Our experiments have produced promising results (between 94% and 95% of recall and 82%-89% of accuracy rates) when identifying anomalies caused by human interventions, using the annotation information in the case of the benchmark data. The annotation is a tedious task and involves careful control, recording, and monitoring of values. For this reason, our results are significant to domain specialists because there is a lack of annotations in different real reservoir data anomaly situations. Furthermore, because our solution is included in a visualization process, the created model is transparent and self-explanatory.

Users can modify the cut-off parameters by gaining more knowledge through the distribution chart and the different visualizations. Concerning the moving z-score strategy, re-calculations can be done in real time, and the strategy based

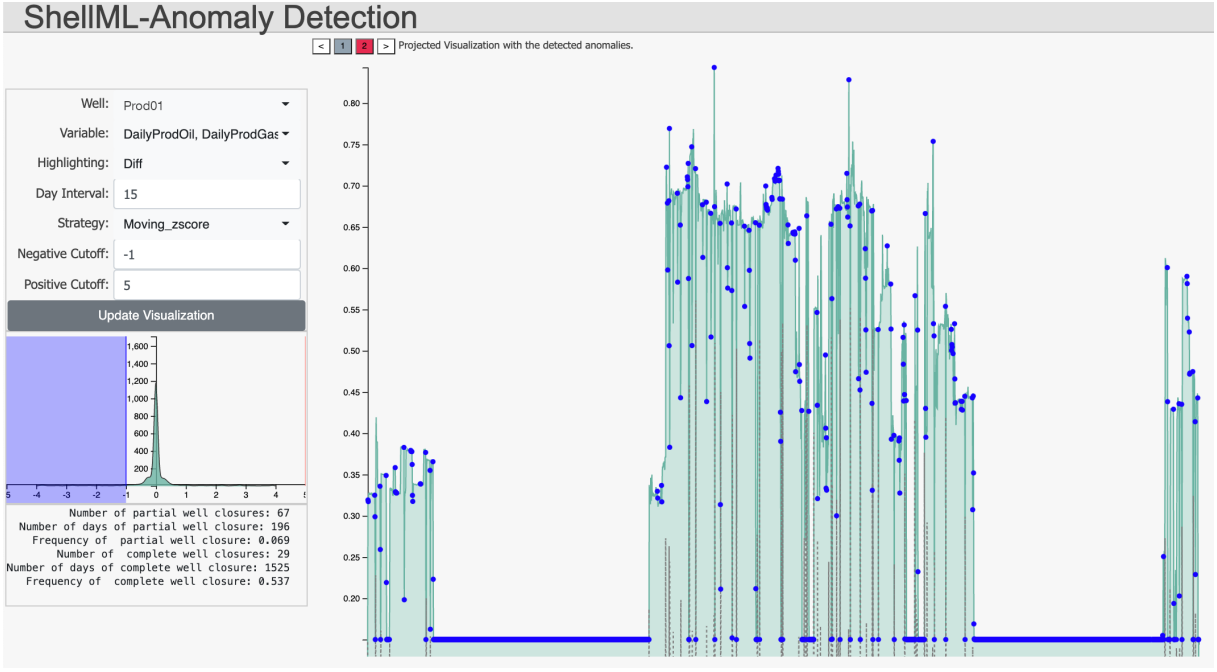


(a)

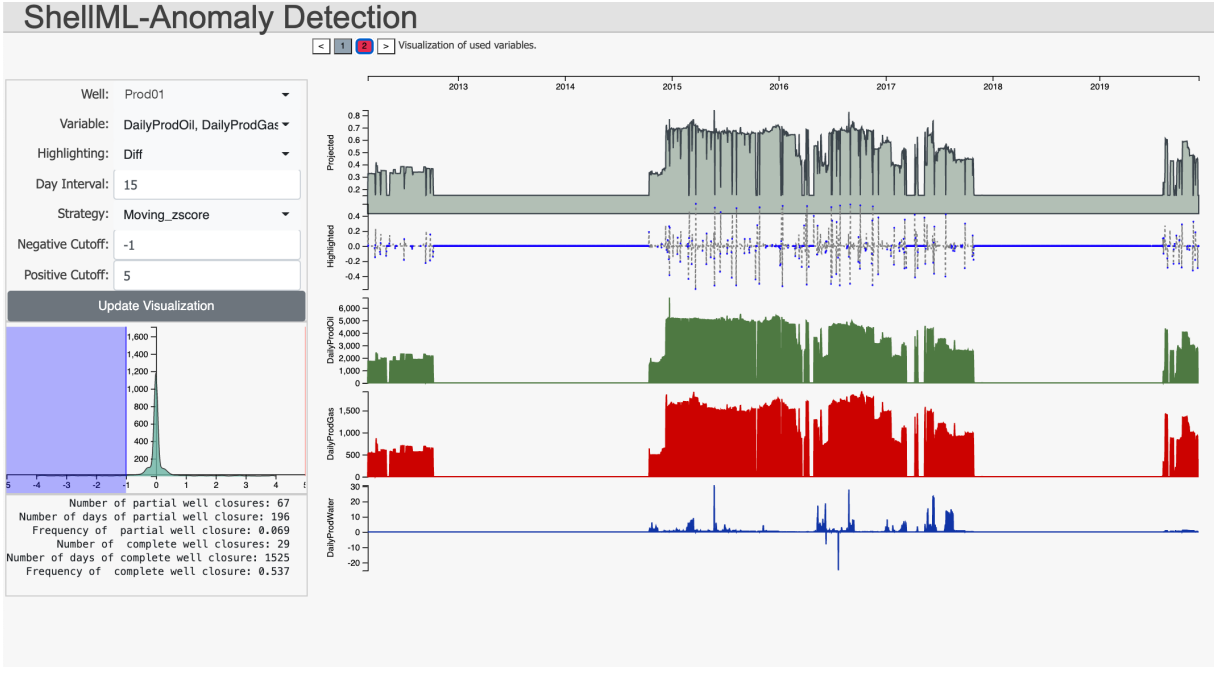


(b)

Figure 11: Visual Analytics of Producer PRK061 using the sequential process 2. We can see that our strategy detected all the existing declines in the data. These declines were specified as less than -3. Looking at the data distribution panel, we can see that the selected region consists of a small region away from the distribution hood center. More details could be considered by increasing that cut-off, such as -2 or -1.



(a)



(b)

Figure 12: Visual Analytics of a real Producer well using the sequential process 2, but with a negative cut-off of -1. In the distribution panel, we identified that the different values are placed below -1. Then, we increased the negative cut-off to -1. The anomalies were detected, which should be analyzed in detail by our specialists.

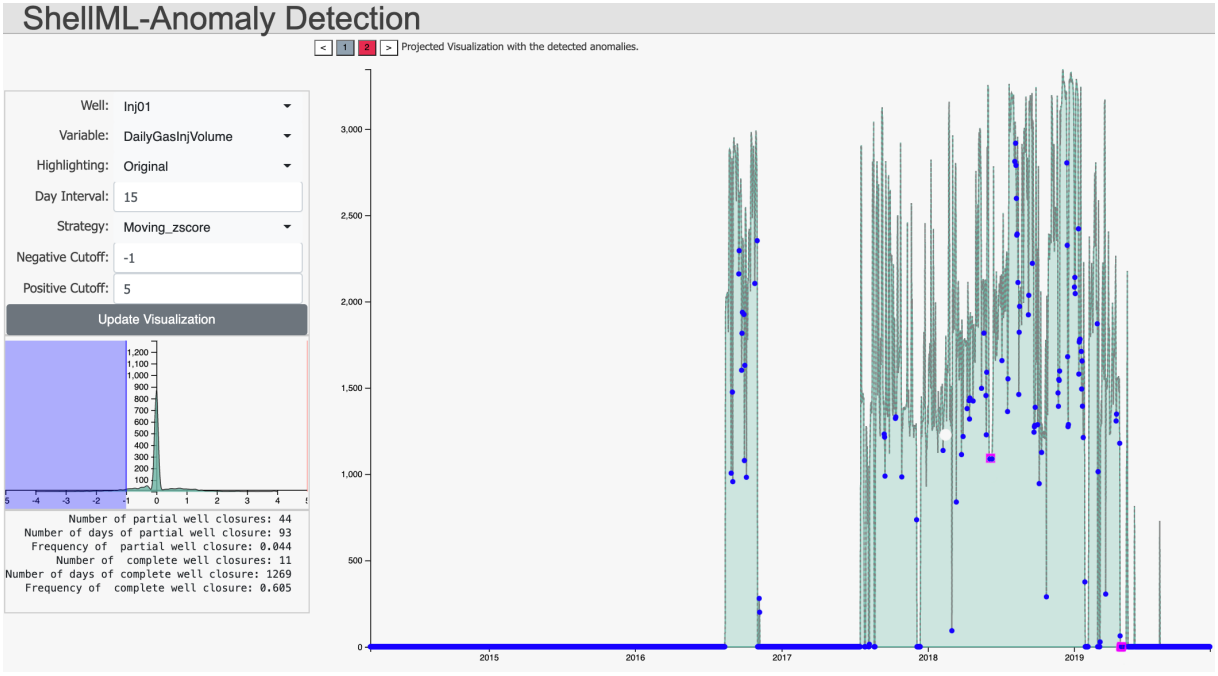


Figure 13: Visual Analytics of a real Injector well using Moving z-score with one variable: daily gas injected volume, but with a negative cut-off of -1. Two annotations were found for this well, which were identified by our approach. The two real annotations were placed in magenta.

on the Isolation Forest requires between 30 and 40 seconds. Indeed, modifying the parameters certainly reflects on the results, which were only obtained as a method for evaluation in the manuscript, given that labels are not usually available in a real scenario. As the interactive platform is self-explanatory, the specialists, who were actively involved in this effort, could adjust their choices to be more stringent when detecting anomalies or to relax them. Our contributing specialists are satisfied with the amount of time needed for re-calculations. After some experiments, we were able to define default parameter values for the hydrocarbon reservoir we work with.

Different directions are possible for future work. An idea that arose is the possibility of automatically determining data distributions to define near-optimal cut-off parameter values. In our experiments, our approach identified other anomalies that are not related to human interventions. This fact invites us to generate the following hypothesis: the detected anomalies can be clustered, and each group can be characterized to get specialists' attention. The visual analytics framework can be evolved to consider other aspects related to the anomaly identification mentioned above. The methods can also be coupled with production data forecasting in an end-to-end process, further improving predictions and eliminating data inconsistencies.

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) 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.

References

Barbariol, T., Feltresi, E., Susto, G.A., 2019. Machine learning approaches for anomaly detection in multiphase flow meters. IFAC-PapersOnLine 52, 212–217.

Bernard, J., Wilhelm, N., Scherer, M., May, T., Schreck, T., 2012. Timeseriespaths : Projection-based explorative analysis of multivariate time series data, in: Journal of WSCG, pp. 97–106.

Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. Lof: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 93–104.

Chimphlee, W., Abdullah, A.H., Sap, M.N.M., Chimphlee, S., Srinoy, S., 2007. Unsupervised clustering methods for identifying rare events in anomaly detection. International Journal of Computer and Information Engineering 1, 2568 – 2573.

Close, L., Kashef, R., et al., 2020. Combining artificial immune system and clustering analysis: A stock market anomaly detection model. Journal of Intelligent Learning Systems and Applications 12, 83.

Correia, M., Hohendorff, J., Gaspar, A.T.F.S., Schiozer, D., 2015. Unisim-ii-d: Benchmark case proposal based on a carbonate reservoir, in: Latin American and Caribbean Petroleum Engineering Conference, Society of Petroleum Engineers, Quito, Ecuador. p. 21. URL: https://doi.org/10.2118/177140-MS, doi:10.2118/177140-MS. sPE.

Elghanuni, R.H., Ali, M.A.M., Swidan, M.B., 2019. An overview of anomaly detection for online social network, in: IEEE 10th Control and System Graduate Research Colloquium (ICSGRC), pp. 172–177.

van den Elzen, S., van Wijk, J.J., 2013. Small multiples, large singles: A new approach for visual data exploration, in: Computer Graphics Forum, Wiley Online Library. pp. 191–200.

Fisher, W.D., Camp, T.K., Krzhizhanovskaya, V.V., 2017. Anomaly detection in earth dam and levee passive seismic data using support vector machines and automatic feature selection. Journal of Computational Science 20, 143–153.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. arXiv preprint arXiv:1406.2661 .

Görnitz, N., Braun, M., Kloft, M., 2015. Hidden markov anomaly detection, in: International conference on machine learning, pp. 1833–1842.

Gupta, S., Patel, S., Kumar, S., Chauhan, G., 2020. Anomaly detection in credit card transactions using machine learning. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) 8, 1–5.

Habibi, M.S., Shirkhodaie, A., 2012. A survey of visual analytics for knowledge discovery and content analysis, in: Signal Processing, Sensor Fusion, and Target Recognition XXI, International Society for Optics and Photonics. p. 83920T.

He, Z., Xu, X., Deng, S., 2003. Discovering cluster-based local outliers. Pattern Recognition Letters 24, 1641–1650.

Hill, D.J., Minsker, B.S., Amir, E., 2007. Real-time bayesian anomaly detection for environmental sensor data, in: Proceedings of the Congress-International Association for Hydraulic Research, Citeseer. p. 503.

Janetzko, H., Stoffel, F., Mittelstädt, S., Keim, D.A., 2014. Anomaly detection for visual analytics of power consumption data. Computers & Graphics 38, 27–37.

Jung, I.S., Berges, M., Garrett Jr, J.H., Poczos, B., 2015. Exploration and evaluation of ar, mpca and kl anomaly detection techniques to embankment dam piezometer data. Advanced Engineering Informatics 29, 902–917.

Kalamaras, I., Zamichos, A., Salamanis, A., Drosou, A., Kehagias, D.D., Margaritis, G., Papadopoulos, S., Tzovaras, D., 2017. An interactive visual analytics platform for smart intelligent transportation systems management. Transactions on Intelligent Transportation Systems 19, 487–496.

Keogh, E.J., Pazzani, M.J., 1998. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback., in: Kdd, pp. 239–243.

Liu, F.T., Ting, K.M., Zhou, Z.H., 2008. Isolation forest, in: International Conference on Data Mining, IEEE. pp. 413–422.

Malhotra, P., Vig, L., Shroff, G., Agarwal, P., 2015. Long short term memory networks for anomaly detection in time series, in: Proceedings, Presses universitaires de Louvain. pp. 89–94.

Pereira, J., Silveira, M., 2019. Unsupervised representation learning and anomaly detection in ecg sequences. International Journal of Data Mining and Bioinformatics 22, 389–407.

Shi, L., Liao, Q., He, Y., Li, R., Striegel, A., Su, Z., 2011. Save: Sensor anomaly visualization engine, in: Conference on Visual Analytics Science and Technology (VAST), IEEE. pp. 201–210.

Škvára, V., Pevnỳ, T., Šmídl, V., 2018. Are generative deep models for novelty detection truly better? arXiv preprint arXiv:1807.05027 .

Sommer, R., Paxson, V., 2010. Outside the closed world: On using machine learning for network intrusion detection, in: Symposium on Security and Privacy, IEEE. pp. 305–316.

Soriano-Vargas, A., Hamann, B., de Oliveira, M.C.F., 2019. Tv-mv analytics: A visual analytics framework to explore time-varying multivariate data. Information Visualization 19, 3–23.

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, 50–64.

Stoffel, F., Fischer, F., Keim, D.A., 2013. Finding anomalies in time-series using visual correlation for interactive root cause analysis, in: Proceedings of the Tenth Workshop on Visualization for Cyber Security, pp. 65–72.

Sun, S., 2013. A survey of multi-view machine learning. Neural computing and applications 23, 2031–2038.

Suschnigg, J., Mutlu, B., Fuchs, A.K., Sabol, V., Thalmann, S., Schreck, T., 2020. Exploration of anomalies in cyclic multivariate industrial time series data for condition monitoring, in: EDBT/ICDT Workshops, pp. 1–8.

Tian, H., Khoa, N.L.D., Anaissi, A., Wang, Y., Chen, F., 2019. Concept drift adaption for online anomaly detection in structural health monitoring, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2813–2821.

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., Magalhães, L.P., 2019. A realistic and public dataset with rare undesirable real events in oil wells. Journal of Petroleum Science and Engineering 181, 106223.

Wu, W., Zheng, Y., Chen, K., Wang, X., Cao, N., 2018. A visual analytics approach for equipment condition monitoring in smart factories of process industry, in: Pacific Visualization Symposium (PacificVis), IEEE. pp. 140–149.

Yadav, S.S., Vijayakumar, V., Athanesious, J., 2018. Detection of anomalies in traffic scene surveillance, in: International Conference on Advanced Computing (ICoAC), IEEE. pp. 286–291.

Yu, Y., Zhu, Y., Li, S., Wan, D., 2014. Time series outlier detection based on sliding window prediction. Mathematical problems in Engineering 2014.

Zenati, H., Foo, C.S., Lecouat, B., Manek, G., Chandrasekhar, V.R., 2018. Efficient gan-based anomaly detection. arXiv preprint arXiv:1802.06222

.
Zhong, Z., Sun, A.Y., Yang, Q., Ouyang, Q., 2019. A deep learning approach to anomaly detection in geological carbon sequestration sites using pressure measurements. Journal of hydrology 573, 885–894.

Appendix

A. Isolation Forest Strategy for 1 evaluated point

Given the values from a time interval (a matrix of multivariate values), the point to be evaluated and the user-defined cut-off value, the algorithm returns True if the evaluated point is considered anomaly and otherwise False.

Algorithm 1 Isolation Forest Strategy for 1 evaluated point.

Input:

1: **X_interval:** values from the time interval
2: **x:** the evaluated point
3: **t:** the user-defined cut-off value

Output: True if the evaluated value is an anomaly or False otherwise.

4: Calculate **c(n)** (Eq. 5 in the manuscript), using n as the number of instances of the time interval
5: Build the 100 isolation trees (**iTrees**).
6: **for** $k \leftarrow 1$ to 100 **do**
7: For each variable, identify the minimum and the maximum value
8: Choose a variable randomly
9: Choose a random value in the variable range
10: Repeat the steps 8. and 9. until reaching the maximum depth.
11: **end for**
12: For each iTree, calculate the $h(x)$.
13: **for** $k \leftarrow 1$ to 100 **do**
14: Calculate the number of edges that x traverses in the iTree $h(x)$.
15: **end for**
16: Calculate $E(h(x)) = \frac{\sum_{i=1}^{100} h_i(x)}{100}$.
17: Calculate the score $s(x, n)$ (Eq. 6 in the manuscript).
18: **if** $(s(x, n) \geq t)$ **then**
19: **return** False
20: **else**
21: **return** True
22: **end if**