Scientific Research Publishing

Unsupervised Methods to Classify Real Data from Offshore Wells

Antônio Orestes De Salvo Castro[1], Mayara De Jesus Rocha Santos[2], Fabiana Rodrigues Leta[2], Cláudio Benevenuto C. Lima[3], Gilson Brito Alves Lima[2]

[1]Rio de Janeiro State University (UERJ), Rio de Janeiro, Brazil
[2]Department of Engineering of Fluminense Federal University (UFF), Niterói, Brazil
[3]Petróleo Brasileiro S.A, Rio de Janeiro, Brazil
Email: mayarajrs@id.uff.br

Abstract

In the petroleum industry, sensor data and information are valuable. It can detect, predict and help to understand processes during oil production. Offshore wells require more attention. Once workovers, maintenance, and intervention are more costly than onshore wells. Coupling data-driven methods for well-monitoring applications, two unsupervised classification methods, one statistical and one machine learning-based, are proposed to detect anomalies in well data. The novelty is presented by applying a Control Chart using a 3 standard deviations window for the Permanent Downhole Gauge Pressure sensor (P-PDG), and a Fuzzy C-means algorithm to classify data from pressure and temperature sensors in an offshore field. The main goal in structuring a classified data set is using it to train machine learning models to monitor and manage petroleum production. Modeling applications for early fault detection systems in offshore production, based on real-time data from production sensors, require classified data sets. Then, labeling two target classes: "normal" and "fault" is a key step to be implemented in order to train the machine learning models. Therefore, this paper applies two methodologies to classify a real-time data set to create a training data set divided into "normal" and "fault" classes. Thus, it is possible to visualize the abnormal events pointed out by the methodologies and compare how sensible is each method. In addition, it is proposed a random forest application to test the performance of the classified data sets from both methods. The results have shown that the control chart method presents higher sensibility than fuzzy c-means, however, the differences between are insignificant. The random forest performance displayed sensitivity and specificity values of 99.91% and 100% for the data set classified by the control chart method and 94.01% and 99.98% for the data set classified by fuzzy c-means algorithm.

1. Introduction

Brazilian oil and gas production has been increasing in the last 50 years. In 2019, Brazil was the tenth bigger petroleum producer, with an average of 2787 million barrels per day (bpd) and 12,253 million cubic meters per day of natural gas [1]. Most of the Brazilian production is centered in offshore fields, which demands more complex and expensive operations [2].

The predominant production unit established in Brazil is the FPSO (Floating, Production, Storage, and Offloading). The FPSO is built using petroleum ship structures that are attached to the process and storage plant [3]. The subsea set of offshore fields is characterized by wet Christmas trees, flowlines, riser, and other subsea equipment. Christmas trees are an assembly of valves, sensors, and connectors responsible for controlling the hydrocarbon flow from inside the well through the flowline and riser, up to the topside. The sensors and valves are designed to uphold high pressure and severe working conditions.

Developing new strategies that apply data mining, machine learning and intelligent methods have become a new trend in petroleum exploration and development [4]. The data mining process allows the extraction of knowledge from databases. Knowledge would help build decision systems, improving productivity, and reducing costs [5]. Thus, the database that will provide the required knowledge for helping the decision-making process must present some key aspects; for instance, the data set for supervised machine learning applications requires labeled data. The novelty presented in this paper provides two methods to classify real well data in "normal" and "fault" labels. Hence, with a structured method to classify well data into the desired classes, intelligent systems and supervised machine learning algorithms to detect early fault occurrence can be applied.

In oil and gas fields, the monitoring and maintenance of production systems are key points considering revenue and safety aspects. Data has become a potential resource that allows the continuous development of artificial intelligence technology. Although, smart systems require high-quality data, clear application scenarios, proper models, and other provisions as well [4]. Supervised machine learning techniques have been applied to data collected by multi-domain sensors for integrity monitoring of production components. The Knowledge Discovery in Database is the key to extract new findings and unravel patterns in databases [6]. However, for old component systems where the sensor reliability is compromised or in cases where there is no labeled data available, an unsupervised method is required [7].

Considering the different sensors disposed on the subsea and topside set and

the advance in data acquisition, storage, and process, the production data acquired by monitoring systems aligned with smart solutions can lead to improvements in well management. This paper proposes using two unsupervised approaches to classify a real data from oil and gas production to identify fault occurrence from normal production operation. The main gain in producing a classified data set, labeling normal and abnormal production, is producing a training data set for further machine learning applications. With an unsupervised method using sensor data to create a training data set, it is possible to classify new data, providing, then, a data-driven approach to manage well production information.

1.1. The Early Fault Detection Approach

To minimize costly intervention in production wells, alongside preventive measures, it is necessary to take corrective action as soon as a fault is identified. Thus, avoiding more complex issues. Several papers tackle machine learning applications for early fault detection [8] [9] [10] [11]. Besides early fault detection, data-driven methods can also perform fault diagnosis. However, in the initial stages, the features of a fault are hard to identify due to noise in signals and the unnoticeable symptoms. Thus, in order to combine fault diagnoses with early detection systems, it is important to invest in other techniques to improve data quality in data-driven based algorithms [12].

This paper aims to classify a training data set in two target classes ("normal" and "fault") for further machine learning applications focused on fault detection. The Control Chart method using 3 standard deviations limits the normal operation zone and is a hard classification. Therefore, a given instance belongs to one class or another, they are mutually exclusive. On the other hand, the Fuzzy C-means classification is a soft classification. A given instance can belong to more than one class with different values for the membership function, but with a higher value for one class than another. The Fuzzy C-means classification implements data from downhole pressure, temperature and flow sensors, while the control chart classification is built on downhole pressure data.
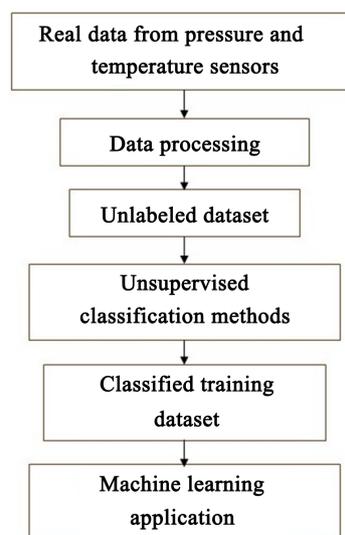
1.2. Fuzzy C-Means

The Fuzzy C-means algorithm was created by Bezdek in 1992. By default, it is necessary to specify firsthand how many clusters the data set will be divided into [13]. However, there are alternative methodologies to determine previously the number of clusters necessary, as published in [14] [15] [16]. This algorithm is characterized by a fuzzy clustering problem, where the goal is to obtain a fuzzy partition in a data set assigning each instance to a membership function value for each class. Nevertheless, the Fuzzy C-means method involves splitting the data into clusters, assigning random coefficients of cluster membership for each instance, calculating the clusters' centroid and the clusters' membership values, and repeating this process until the algorithm converges. For the Fuzzy C-means method, the R package named "e1071" was implemented.

1.3. Control Chart Using 3 Standard Deviations

The control chart approach is commonly used to monitor a process. It shows graphically the average value and the upper and lower control limits of a process [17]. The control chart method is a continuous way to monitor one or more parameters of a production process. The upper and lower limits established help to detect changes in the process. They can be categorized into two types; memory-less and memory control charts [18]. The memory-less type depends on current information to draw control limits. However, the memory type is built on current as well past samples. Fuzzy exponentially weighted moving average [17], and cumulative sum [19] are examples of memory control charts. In this paper, a memory-less control chart is proposed, and the lower and upper limits are calculated based on 3 standard deviations from the average value in the interval.

2. Methodology and Principals

The methodology carried out in this paper follows the workflow shown in **Figure 1**. The production data available by the Plant Information system go through consistency routines to avoid outliers, problems with sensor signal or instrument failures. Thus, the real data gathered from pressure and temperature sensors in a production well in Campos Basin are assembled in a database, where pressure and temperature data per minute from different sensors are synchronized. The database also included Christmas tree and topside valves status for each time-stamp. The data processing phase embraces the study of subsea and topside valve configuration to determine whether the well is in a production or maintenance operation. Thus, excluding maintenance operation data and also taking out periods after changes in valves status, the production data sets were created. Separating a data set with variables related to pressure and temperature sensors only, producing an unlabeled data set for unsupervised classification methods.


Figure 1. Methodology flowchart.

The unsupervised methods, Fuzzy C-means and Control Chart using 3 standard deviations, applied in the data set provided the classification of production data. The Fuzzy C-means algorithm classified the data set into three classes; "normal", "high fault" and "low fault" for data above and under the normal level, respectively. Meanwhile, the Control Chart divided the data set into groups "normal" and "fault", considering data with variation higher than 3 standard deviations to be a failure indication.

2.1. Performance Evaluation

The values reported by the production sensors are the input information to develop models for fault detection in oil and gas wells. Regarding the performance of the two unsupervised methods proposed in this work, a comparison metric considering the amount of a variable value that overlaps the two target classes. In other words, the amount of data that belong, at the same time, to the "normal" and "fault" classes. Figure 2 characterizes how to issue this metric. For a given pressure sensor, the orange data represents input values labeled as "normal", in blue and green is shown the advance of a faulty state. On the left is marked the range of pressure values that were labeled as "normal" and, on the other side, is the range of pressure values considered as "fault". The values that overlapped the normal and fault range were taken into account to compare the control chart and fuzzy c-means classification. Considering that a higher degree of overlapping data causes negatives impacts on model training, it is aimed for a sensor and method that present the least amount of overlapping data for better performance.

Nevertheless, as for the performance of the classified data set produced by the unsupervised methods in the random forest application, two training data sets regarding the same production interval and the same number of labeled classes ("high fault" and "low fault" labeled by fuzzy c-means algorithm were considered as "fault" class) were implemented. In this study, the Random Forest
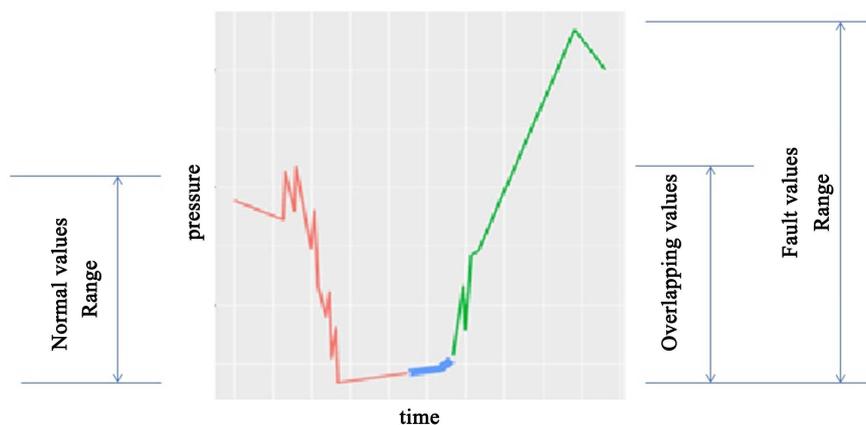
Figure 2. Comparison parameter proposed to compare the unsupervised methods. The red printed segment represents the values labeled as "normal", the blue and green segments are the values assigned as "fault".

algorithm from the library "random Forest" available in Rstudio, with 1000 trees settled as input parameter, was applied to test the classified data sets. Moreover, specificity and sensitivity metrics compared the model's performance for using the classified data set by Fuzzy C-means and Control Chart method.

2.2. Pressure and Temperature Sensors

With the advance of Digital Transformation in the oil and gas sector, production units with adequate instrumentation systems produce reliable information within short intervals. The temperature, pressure, and flow rate sensors generate direct information in real-time for offshore wells [20]. These sensors are located in different positions, collecting data in distinct points during the production flow. In the present paper, the data set contained 6 production variables related to the sensors available, they are displayed in Table 1.

2.3. Problems during Oil and Gas Production

Oil fields have a long production life, reaching many years or decades. Thus, assuring production maintenance at profitable levels requires efficient management. To fail in controlling and managing the production operation is to reduce the field's expected production life [21]. The detection of an abnormal behavior during the production flow is based on an individual or group of monitored production variables. Therefore, depending on the variable's behavior, decisions and measures to mitigate the production loss can be made in time.

Anomalies during oil and gas production can be due to flow assurance, mechanical, and integrity problems. Restriction in the diameter available to the production flow, caused by the fluid composition, solid deposition, or related to pressure and temperature fluctuations, leads to flow assurance issues. Hydrate formation, scaling, and slugging are some examples [22] [23] [24]. Also, changes in the reservoir static pressure can increase BSW (Basic Sediments and Water) rate, modifying the viscosity of the production fluid. On the other hand, mechanical or integrity failures in oil wells lead to fluid leaking. Some equipment and structures prone to these types of faults are DHSV (Downhole Safety Valve) valves, production and casing tubing, production packers, Christmas Tree, wellhead, flowlines, gas lift, and choke valves. In addition, there are also issues related to

Table 1. Production variables from subsea and topside sensors.

Variable	Description
P-PDG	Pressure at the Permanent Downhole Gauge sensor
T-PDG	Temperature at the Permanent Downhole Gauge sensor
P-PCK	Pressure at the topside production choke sensor
T-PCK	Temperature at the topside production choke sensor
P-GL	Pressure at the gas lift injection sensor
Q-GL	Flowrate at the gas lift injection sensor

the reservoir as water and gas cones, and sediments production. In Brazilian offshore production, hydrate formation, scaling, restriction in the production choke valve, increase in BSW, DHSV failure, severe slugging, rapid production loss, and flow instability are the most common undesired events that occur during offshore petroleum production [25].

2.4. Data Processing

In the present work, data from real wells were provided by the petroleum company which manages the offshore field in Campos Basin that is responsible for a large contribution to the national oil and gas production. The data were extracted from P-PDG, T-PDG, P-PCK, T-PCK, Q-GL, and P-GL once these sensors were provided with higher quality signals than others.

The well data refer to 4 periods of analysis, selected between January and December in 2012. Thus, the data processing proceeded by following the steps bellow:

1) Remove data that follow a valve closing or opening and data during maintenance periods. Data-driven models for fault detection use the well normal condition as a parameter to establish whether the operation is at fault or not. Knowing if the valves are assembled for operation or maintenance purposes, enables operators to separate intervals where the hydrocarbon production is steady, without oscillations due to valve closing and opening or due to maintenance implications.

2) Then, after selecting the target intervals with only operation data, withdrawing transient data related to change in valves' status. The well data were partitioned into 11 intervals, where 4 of them were chosen as input for the unsupervised classification process. For each chosen interval, step 3 was carried on.

3) First, apply Fuzzy C-means classifier in P-PDG, T-PDG, and T-PCK, grouping 3 classes; "normal", "high fault" (when the target data is above the normal class), and "low fault" (when the target data is under the normal class). Afterward, the "low fault" and "high fault" classes will be reclassified as "fault" class. This subdivision was done to help in the clustering process, once signals can be related with system pressurization or depressurization. Then, plot and analysis the classes graphs and their membership function for the "normal", "high fault" and "low fault" classes. After identifying the "normal" class membership diagram, consider as normal operation the data with 95% of chance to belong to "normal" class and consider as fault occurrence everything else. On the other hand, the control chart approach classifies the data set directly into "normal" and "fault" classes.

4) Compare the classification obtained by the Fuzzy C-means algorithm and the control chart method, in a machine learning application such as Random Forest, using sensitivity and specificity metrics.

For Control Chart method it is important that the data is normally distributed. The algorithm loses its definition when the input data is no longer a normal distribution and the results obtained cannot be trusted. Therefore, before applying

the control chart methodology, the unclassified data set was tested to verify if the data were normally distributed. The Shapiro-Wilk normality test showed a p-value of 0.01749 indicating that the normality hypothesis was not rejected using a significance level of 1%. The kurtosis coefficient was 1.5602, indicating the data distribution has a more flattened aspect than a normal distribution curve. In addition, the asymmetric coefficient value obtained was -0.29512 implicating that the curve is slightly shifted to the left. The visual data representation of the P-PDG sensor is displayed in Figure 3.

3. Results

3.1. Data Classification

Using Fuzzy C-means, the data set was divided into 3 clusters. The clusters were colored in red, blue, and orange displayed in the first graphic in Figure 4. To determine which cluster accounts for the normal condition data, an analysis of the membership function was carried out. Figure 5 and Figure 6 show the membership function of cluster 1 and 2. The cluster which may represent the normal condition must be the one that better divides the probability of a given data be "normal" or "fault" class. Indeed, cluster number 3, showed in more detail in Figure 7, presents a membership function that provides a clearer visualization of the "normal" and "fault" class ("0" and "1", respectively). Then, with a significance level of 0.5%, data with a probability assigned by the membership function of cluster blue with values lower than 0.005 would be classified as "normal" and otherwise "fault".
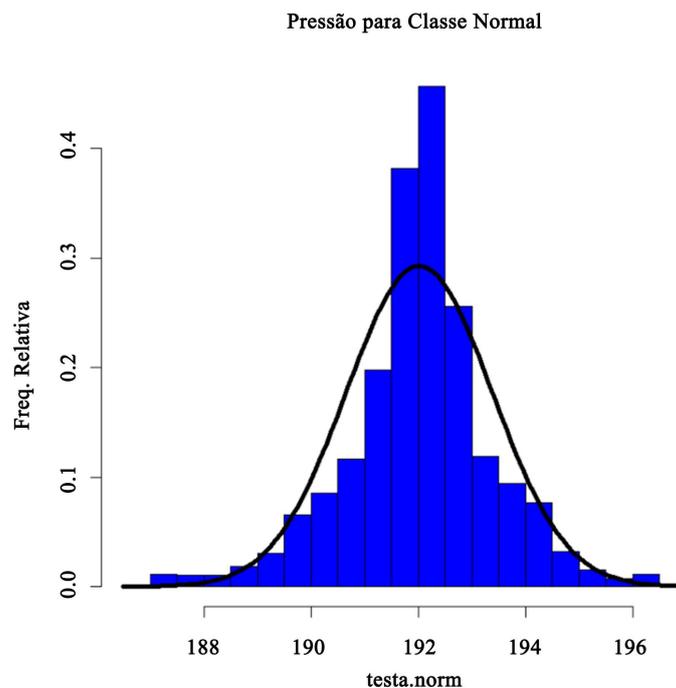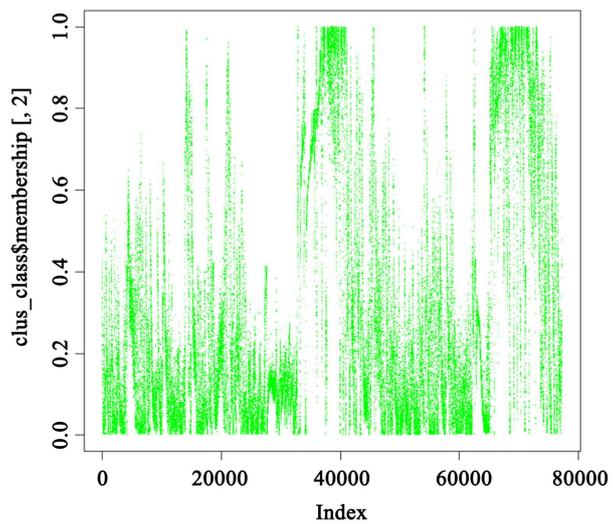

Figure 3. Data distribution of pressure values from P-PDG sensor and the normal distribution hypotheses.
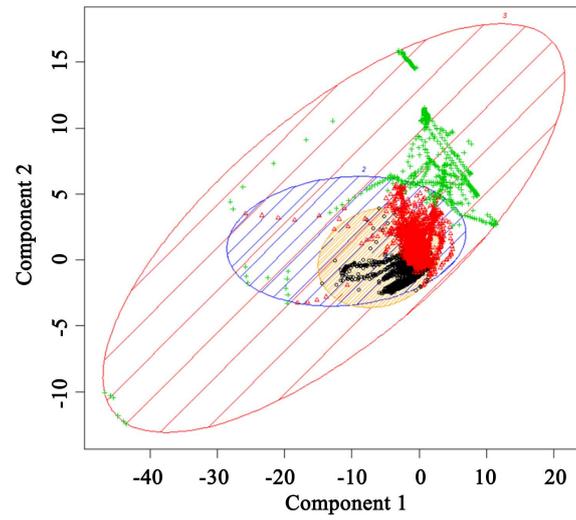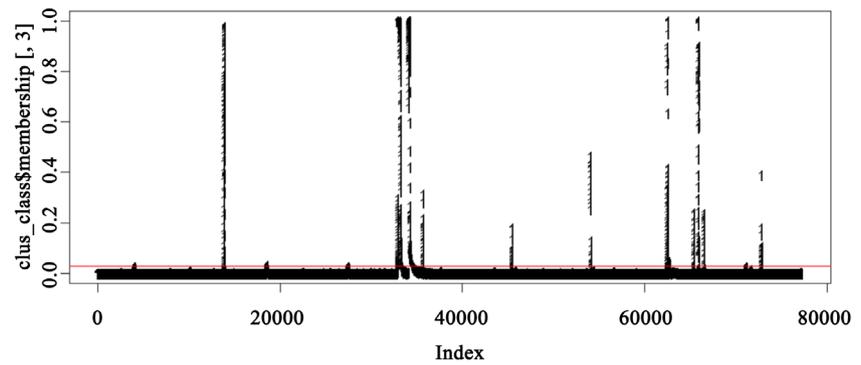
Figure 4. Membership function for each class cluster. Fuzzy C-means data classification.
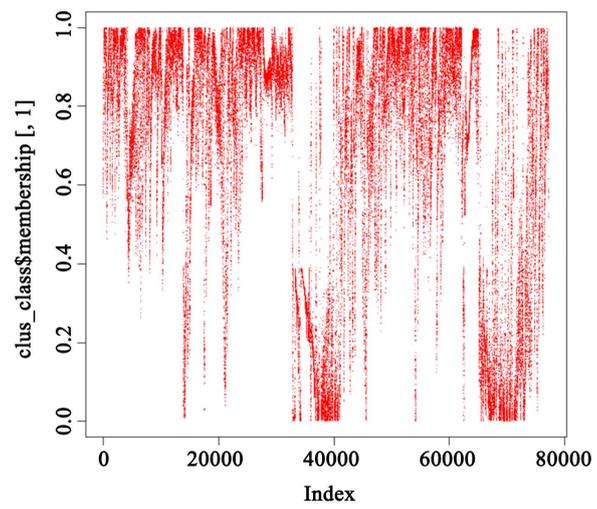
Figure 5. Membership function for cluster 1.

**Figure 6.** Membership function for cluster 2.

235

Figure 7. Membership function for cluster 3, the cluster with normal operation data.

The Control Chart method already divides the data set into two groups, one group limited by the normal condition zone specified within a variation of 3 standard deviations and another containing the data gathered outside this zone. From the 77,307 instances in the interval chosen, 72,033 were classified as normal condition and 5274 as fault events for this method. Moreover, for the Fuzzy C-means algorithm, 73,303 instances were classified as "normal" and 4004 as "fault".

The resulting classification for both methodologies applied is presented in Figure 8. For the P-PDG sensor, it is shown the normal range condition for this sensor considering both techniques. In blue is represented the data considered as normal operation during oil and gas production, and in red is presented abnormal behavior during well production that was not caused by a change in valves status or maintenance procedures.

3.2. Comparing Fuzzy C-Means and Control Chart Method

In order to compare the unsupervised methods proposed a metric, considering the percentage of data that overlay two the target classes, is proposed. Table 2 displays the results obtained for each production sensor available in this study. The "Overlapping 3SD" column stands for the percentage of overlapping data in the control chart method using 3 standard deviations, and "Overlapping FUZ" for the fuzzy c-means algorithm. The lower the percentage of overlapping data the better the classified data set to develop fault detection models. Thus, among the sensors available, P-PDG and P-PCK are the ones with higher potential to deliver a better classification model. Meanwhile, the columns named "FAULT" indicate the percentage of the total sensor's range that was classified as "fault". In the cases where sensors presented 100% of their range as fault class states that all range of values labeled as normal are overlaying the values assigned as fault. Moreover, the "NORMAL" columns display the percentage of the sensor's range that is classified as "normal". Thus, as stated in Table 2, the P-PDG sensor presents a lower amount of overlapping data. For the production sensor available, the control chart method showed the least degree of overlaying data, however, the difference compared with the results from fuzzy c-means is negligible.
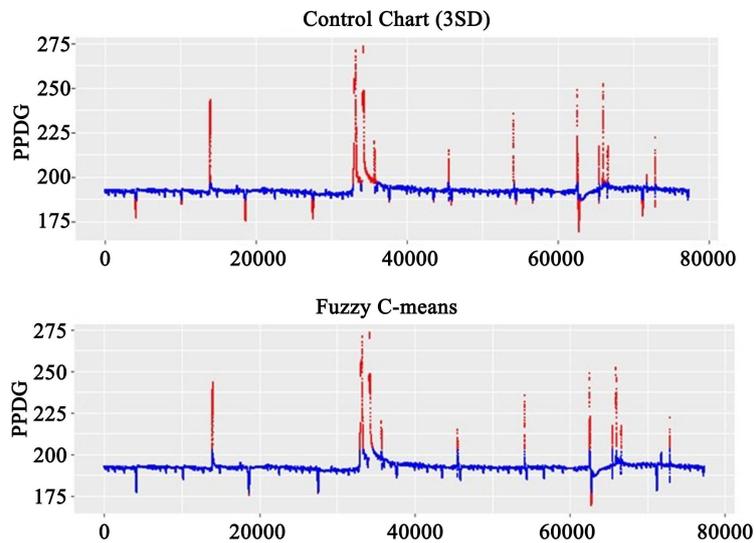
Figure 8. Classified data set applying Control Chart (upper graph) and Fuzzy C-means algorithms. Where the normal condition is shown in blue and anomalies in red. The y-axis presents the pressure range values while the x-axis accounts for the instances classified.

Table 2. Amount of overlapping data regarding control chart and fuzzy c-means classification for each sensor.

| SENSOR	Overlapping 3SD	Overlapping FUZ	Fault 3SD	Fault FUZ	Normal 3SD	Normal FUZ
P. PDG	0	0.05	0.91	0.96	0.09	0.1
T. PDG	0.52	0.52	1	1	0.52	0.52
P. PCK	0.23	0.09	1	1	0.23	0.09
T. PCK	0.32	0.32	1	1	0.32	0.32
QGL	0.65	0.65	1	1	0.65	0.65

3.3. Random Forest Application

The random forest algorithm using 1000 trees was applied in the two data sets classified by Control Chart and Fuzzy C-means techniques. The data sets contained 77,307 instances, where 50,000 were set as the training data set and the remaining 27,307 as the testing data set. The training data set size was established based on the number of faults occurred. Within the 50,000 instances, both training data set presented the same number of "fault" labels, which allowed a better comparison between the two methods.

Furthermore, sensitivity and specificity, metrics based on the number of true positives and the number of true negatives over total instances, respectively, were assigned to compare the result obtained by a random forest classifier.

The data set classified by the Control Chart method divided the training data set into 45,923 "normal" class and 4077 "fault" class. Meanwhile, the Fuzzy C-means training data set had 46,848 instances of "normal" class and 3142 "fault" class. The inconsistency between the two classified training data sets numbered

925 instances. Graphically, as presented in **Figure 9**, the differences between the Control Chart and Fuzzy C-means training data sets are very smooth. The resulting classification from the random forest algorithm using the two different data sets is shown in **Figure 10**. The classification based on the training data set using the Control Chart method presented a sensitivity of 99.91% and 100% of specificity. Nevertheless, the result from the training data set classified by Fuzzy C-means showed 94.01% of sensitivity and 99.98% of specificity. The results from both methods confirm that the training data set size was enough to achieve a satisfactory classification.
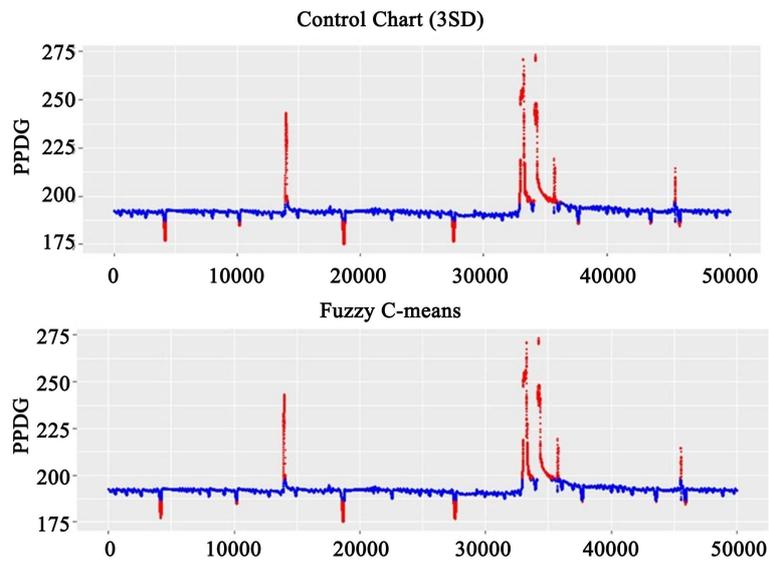
Figure 9. Classified training data sets for control chart and fuzz c-means methods used as input in the random forest model.
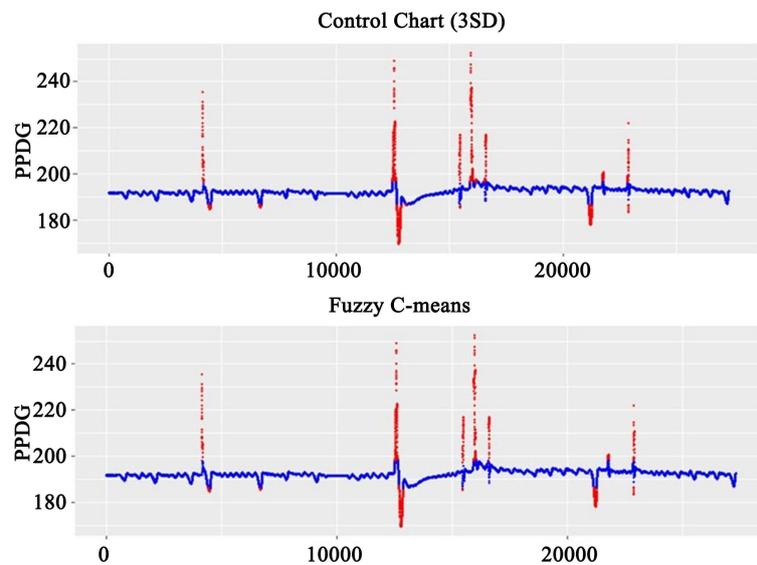
Figure 10. Resulting classified data set from random forest application using control chart and fuzzy c-means training data sets. Blue data represents "normal" condition data while red printed data displays "fault" occurrence.

4. Discussion

Real data from offshore well does not always present good quality. When working with sensors data, it is important to structure and synchronizing the information available in a way to avoid the loss of data. The data set structured with production data presented the P-PDG variable as the one with more quality and higher frequency, also it was the sensor with fewer overlapping classes as shown by Table 2, hence the best to be applied in the Control Chart methodology. Otherwise, the Fuzzy C-means algorithm could be applied in any sensor, the best outcome was led by the P.PCK sensor but for comparison reasons, the P-PDG was chosen to be displayed. For this method, the key to identifying the "normal" class is studying the membership function, as plotted in Figure 7. Furthermore, Figure 8 infers that both methods resulted in similar classifications. The Control Chart technique classified more instances as "fault" than Fuzzy C-means. Nevertheless, the training data sets used for the random forest application also presented similar classification. Once more the control chart data set showed few more "fault" instances; however, the training data sets also led to analogous classification. The random forest sensitivity and specificity metrics have shown that both methodologies are qualified to build classified training data sets for supervised learning applications for monitoring and managing well data.

5. Conclusions

To apply intelligent solutions and reduce costs with well interventions, structured and labeled databases are necessary. This work described two methods to classify well data from subsea and topside sensors of an offshore field. The Control Chart Method used the P-PDG sensor to divide the data set into "normal" and "fault" labels. Meanwhile, the Fuzzy C-means approach could be applied to the other sensors as well. The classification performed by Control Chart and Fuzzy C-means showed that both methods generate similar results. A normal distribution of sensor data is a key point for the Control Chart method, its implementation is simpler than the fuzzy c-means algorithm, which needs a membership function and established limits to classify the data set.

A random forest application was also proposed to show how the labeled data set from the Control Chart and Fuzzy C-means algorithms would perform as training data sets for supervised machine learning-based systems. The training data sets were similar and, thus, were the resulting classifieds data sets using random forest. Nevertheless, the data set from the Control Chart method produced a model with higher sensitivity and specificity.

Acknowledgements

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Agencia Nacional de Petróleo, Gás Natural e Biocombustíveis. Anuário Estatístico 2020.
https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos/anuario-estatistico-2020-dados-abertos

[2] Agencia Nacional de Petróleo, Gás Natural e Biocombustíveis. Boletim da Produção de Petróleo e Gás Natural 2021.
https://www.gov.br/anp/pt-br/centrais-de-conteudo/publicacoes/boletins-anp/bmp/2021/2021-01-boletim.pdf

[3] D'almeida, A.L. (2015) Indústria do Petróleo no Brasil e no Mundo: Formação, Desenvolvimento e Ambiência Atual. Edgard Blucher, São Paulo.

[4] Linchun, K., He, L., Yili, R., Kai, L., Mingyu, S., Jian, S. and Xin, L. (2021) Application and Development Trend of Artificial Intelligence in Petroleum Exploration and Development. *Petroleum Exploration and Development*, **48**, 1-14.
https://doi.org/10.1016/S1876-3804(21)60001-0

[5] Monedero, I., Biscarri, F., León, C., Guerrero, J.I., Gonzáles, R. and Pérez-Lombard, L. (2012) Decision System Based on Neural Networks to Optimize the Energy Efficiency of Petrochemical Plant. *Expert Systems with Application*, **39**, 9860-9867.
https://doi.org/10.1016/j.eswa.2012.02.165

[6] Akinola, S.O. and Oyabugbe, O.J. (2015) Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study. *Journal of Software Engineering and Applications*, **8**, 470-477.
https://doi.org/10.4236/jsea.2015.89045

[7] Giro, R.A., Bernasconi, G., Giunta, G. and Cesari, S. (2021) A Data-Driven Pipeline Pressure Procedure for Remote Monitoring of Centrifugal Pumps. *Journal of Petroleum Science and Engineering*, **205**, Article ID: 108845.
https://doi.org/10.1016/j.petrol.2021.108845

[8] Khodnenko, I., Ivanov, S., Perets, D. and Simonov, M. (2019) Detection of Lost Circulation in Drilling Wells Employing Sensor Data Using Machine Learning Technique. *Procedia Computer Science*, **156**, 300-307.
https://doi.org/10.1016/j.procs.2019.08.206

[9] Shafiee, M., Animah, I., Alkali, B. and Baglee, D. (2019) Decision Support Methods and Applications in the Upstream Oil and Gas Sector. *Journal of Petroleum Science and Engineering*, **173**, 1172-1186. https://doi.org/10.1016/j.petrol.2018.10.050

[10] Mayani, M.G., Rommetveit, R., Helgeland, S. and Oedegaard, S.I. (2019) Advanced Real-Time Monitoring Provides Early Detection and Prevention of Costly Well Problems. *SPE/IADC International Drilling Conference and Exhibition*, The Hague, 5-7 March 2019, 2. https://doi.org/10.2118/194178-MS

[11] Singh, A., Sankaran, S. and Ambre, S. (2019) Improving Deepwater Facility Uptime Using Machine Learning Approach. *SPE Annual Technical Conference and Exhibition*, Calgary, 30 September-2 October 2019, 7-11.
https://doi.org/10.2118/195875-MS

[12] Cai, B., Hao, K., Wang, Z., Yang, C., Kong, X., Liu, Z., Ji, R. and Liu, Y. (2021) Data-Driven Early Fault Diagnostic Methodology of Permanent Magnet Synchronous Motor. *Expert Systems with Application*, **177**, Article ID: 115000.
https://doi.org/10.1016/j.eswa.2021.115000

[13] Bezdek, C.J. and Pal, S.K. (1992) Fuzzy Models for Pattern Recognition. IEEE Press, New York.

[14] Gath, I. and Geva, B. (1992) Unsupervised Optimal Fuzzy Clustering. In: Bezdek,

J.C. and Pal, S.K., Eds., *Fuzzy Models for Pattern Recognition*, IEEE Press, New York, 211-218.

[15] Tamura, S., Higuchi, S. and Tanaka, K. (1971) Pattern Classification Based on Fuzzy Relations. *IEEE Transactions on Systems, Man and Cybernetics*, **SMC-1**, 61-66. https://doi.org/10.1109/TSMC.1971.5408605

[16] Zahid, N., Abouelala, O., Limouri, M. and Essaid, A. (2001) Fuzzy Clustering Based on K-Nearest Neighbours Rule. *Fuzzy Sets and Systems*, **120**, 239-247. https://doi.org/10.1016/S0165-0114(99)00074-3

[17] Senturk, S., Erginel, N., Kaya, I. and Kahraman, C. (2014) Fuzzy Exponentially Weighted Moving Average Control Chart for Univariate Data with a Real Case Application. *Applied Soft Computing*, **22**, 1-10. https://doi.org/10.1016/j.asoc.2014.04.022

[18] Zafar, R.F., Mahmood, T., Abbas, N., Riaz, M. and Hussain, Z. (2018) A Progressive Approach to Joint Monitoring of Process Parameters. *Computer & Industrial Engineering*, **155**, 253-268. https://doi.org/10.1016/j.cie.2017.11.015

[19] Zaman, B., Lee, M.H., Riaz, M. and Abujiya, M.R. (2020) An Improved Process Monitoring BY Mixed Multivariate Memory Control Charts: An Application in Wind Turbine Field. *Computer & Industrial Engineering*, **142**, Article ID: 106343. https://doi.org/10.1016/j.cie.2020.106343

[20] Tian, C. and Hones, R.N. (2019) Applying Machine Learning Techniques to Interpret Flow-Rate, Pressure, and Temperature Data from Permanent Downhole Gauges. *Reservoir Evaluation & Engineering*, **22**, 386-401. https://doi.org/10.2118/174034-PA

[21] Mohammadpoor, M. and Torabi, F. (2018) Big Data Analytics in Oil and Gas Industry: An Emerging Trend. *Petroleum*, **6**, 321-328. https://doi.org/10.1016/j.petlm.2018.11.001

[22] Rebai, N., Hadjadj, A., Benmounah, A. and Berrouk, A.S. (2019) Prediction of Natural Gas Hydrates Formation Using a Combination of Thermodynamic and Neural Network Modeling. *Journal of Petroleum Science and Engineering*, **182**, Article ID: 106270. https://doi.org/10.1016/j.petrol.2019.106270

[23] Zhang, P., Liu, Y., Kuok, S.C., Tan, A.T. and Tomson, M.B. (2019) Development of Modeling Approaches to Describe Mineral Scale Deposition Kinetics in Porous Medium and Pipe Flow System. *Journal of Petroleum Science and Engineering*, **178**, 594-601. https://doi.org/10.1016/j.petrol.2019.03.070

[24] Guerrero-Sarabia, I. and Fairuzov, Y.V. (2013) Linear and Non-Linear Analysis of Flow Instability in Gas-Lift Wells. *Journal of Petroleum Science and Engineering*, **108**, 162-171. https://doi.org/10.1016/j.petrol.2013.01.012

[25] Vargas, R.E., Munaro, C.J., Ciarelli, P.M., Medeiros, A.G., Amaral, B.G., Burrionuevo, D.C., *et al.* (2019) A Realistic and Public Data Set with Rare Undesirable Real Events in Oil Well. *Journal of Petroleum Science and Engineering*, **181**, Article ID: 106223. https://doi.org/10.1016/j.petrol.2019.106223