Statistical analysis of offshore production sensors for failure detection applications

Análise estatística dos sensores de produção offshore para aplicações de detecção de falhas

DOI:10.34117/bjdv7n8-681

Recebimento dos originais: 31/07/2021
Aceitação para publicação: 31/08/2021

Mayara de Jesus Rocha Santos
Master in Mechanical Engineering
Universidade Federal Fluminense (PGMEC-UFF)
Endereço: Rua Passo da Pátria, 156 - Sala 213 Bloco E. Niterói - Rio de Janeiro 24210-240
E-mail: mayarajrs@id.uff.br

Antônio Orestes de Salvo Castro
Doctor in Production Engineering
Universidade do Estado do Rio de Janeiro.
Endereço: Maracanã ,20550900 - Rio de Janeiro, RJ – Brasil
E-mail: orestesh004@gmail.com

Fabiana Rodrigues Leta
Doctor in Mechanical Engineering
Universidade Federal Fluminense (PGMEC-UFF)
Endereço: Rua Passo da Pátria, 156. São Domingos, 24210240 - Niterói, RJ – Brasil
E-mail: fabianaleta@id.uff.br

João Felipe Mitre de Araujo
Doctor in Chemical Engineering
Universidade Federal Fluminense.
Endereço: Rua Passo da Pátria, 156, TEQ/EE, Bloco D, sala 307. São Domingos, 24210240 - Niterói, RJ – Brasil
E-mail: jf_mitre@id.uff.br

Geraldo de Souza Ferreira
Doctor in Production Engineering
Universidade Federal Fluminense.
Endereço: Rua Passo da Pátria, 156, TEQ/EE, Bloco D, sala 307. São Domingos, 24210240 - Niterói, RJ – Brasil
E-mail: geraldoferreira@id.uff.br

Ricardo de Araújo Santos
Master in Production Engineering
Universidade Federal Fluminense
Escola de Engenharia, Departamento de Engenharia de Produção. São Domingos, 24210240 Niterói, RJ – Brasil
E-mail: ricardosantos@id.uff.br

Cláudio Benevenuto de Campos Lima
Doctor in Management Systems
Petrobras
Endereço: Rua Henrique Valadares,15. Rio de Janeiro-Brasil
E-mail: clima@petrobras.com.br

Gilson Brito Alves Lima
Doctor in Production Engineering
Universidade Federal Fluminense (PPGEP)
Escola de Engenharia, Departamento de Engenharia de Produção. São Domingos
24210240 - Niterói, RJ – Brasil
E-mail: glima@id.uff.br

ABSTRACT
Detecting the early stages of failures is an old concern of petroleum industry. In order to tackle this problem, a novel sensor analysis methodology is proposed. The assessment of production sensors' behavior, individually or in a group, leads to a better understanding of failure modes during oil and gas production. Thus, Principal Components Analysis and Logistic Regression are incorporated as multivariate statistical modeling for studying the impact of different anomalies in production sensors. Therefore, a deep statistical analysis of these sensors can strengthen assumptions for supporting the modeling process of early fault detection systems. Based on a reliable public data set containing data from real wells, the application of the PCA approach combined with a Logistic Regression resulted in better visualization and understanding of some failures that occurred during petroleum production, such as the abrupt increase in BSW (Basic sediment and water), spurious closure of DHSV (Down hole Safety Valve), severe slugging, flow instability, productivity loss, quick restriction in PCK (production choke), scaling in PCK and hydrate formation in production lines. The two statistical approaches were used as a combined method to provide useful information regarding the failure modes in the dataset. Also, the dataset presented two classes that are important for anomaly detection in oil wells: “normal” and “abnormal”, which allow detecting when production is outside its normal condition. Then, using the production sensors analysis with failure data can help to formulate better detection algorithms. By using PCA and Logistic Regression it was possible to identify which set of variables is better for detecting a specific type of problem. The application of these techniques boosts the modeling of early detection systems in oil and gas production. Besides, the assumptions led to conclusions about how to put groups of sensors and abnormalities together and how much time a well stands in a steady normal condition. Other conclusions showed the significance of transient information for fault detection modeling and the need for individual wells analyses. Hence, using PCA for treating and transforming the data brings important contributions for early fault detection modeling, once it allowed insight into how sensors and abnormal events can be related. Consequentially, the present paper has significant novelty contribution: it raises important assumptions that help to build solid knowledge about the anomalies behavior and help researchers to implement a better modeling strategy.

Keywords: PCA, Production sensors, Logistic Regression.

RESUMO
Detectar os estágios iniciais das falhas é uma preocupação antiga da indústria petrolífera. A fim de resolver este problema, uma nova metodologia de análise de sensores é proposta.

A avaliação do comportamento dos sensores de produção, individualmente ou em grupo, leva a uma melhor compreensão dos modos de falha durante a produção de petróleo e gás. Assim, a Análise de Componentes Principais e Regressão Logística são incorporadas como modelos estatísticos multivariados para estudar o impacto de diferentes anomalias nos sensores de produção. Portanto, uma análise estatística profunda desses sensores pode reforçar as suposições para apoiar o processo de modelagem de sistemas de detecção precoce de falhas. Com base em um conjunto de dados públicos confiáveis contendo dados de poços reais, a aplicação da abordagem PCA combinada com uma Regressão Logística resultou em melhor visualização e compreensão de algumas falhas ocorridas durante a produção de petróleo, como o aumento abrupto do BSW (Sedimento básico e água), fechamento espúrio do DHSV (Válvula de Segurança Down hole Safety Valve), graves slugging, instabilidade do fluxo, perda de produtividade, restrição rápida no PCK (estrangulamento da produção), escalonamento no PCK e formação de hidratos nas linhas de produção. As duas abordagens estatísticas foram usadas como um método combinado para fornecer informações úteis a respeito dos modos de falha no conjunto de dados. Além disso, o conjunto de dados apresentou duas classes que são importantes para a detecção de anomalias em poços de petróleo: "normal" e "anormal", que permitem detectar quando a produção está fora de seu estado normal. Então, usando a análise dos sensores de produção com dados de falha pode ajudar a formular melhores algoritmos de detecção. Usando PCA e Regressão Logística, foi possível identificar qual conjunto de variáveis é melhor para detectar um tipo específico de problema. A aplicação destas técnicas impulsiona a modelagem de sistemas de detecção precoce na produção de petróleo e gás. Além disso, as suposições levaram a conclusões sobre como colocar grupos de sensores e anormalidades juntos e quanto tempo um poço está em condição normal e estável. Outras conclusões mostraram a importância de informações transitórias para a modelagem da detecção de falhas e a necessidade de análises individuais dos poços. Assim, o uso do PCA para tratar e transformar os dados traz importantes contribuições para a modelagem precoce da detecção de falhas, uma vez que permitiu uma visão de como os sensores e os eventos anormais podem ser relacionados. Conseqüentemente, o presente documento tem uma contribuição inovadora significativa: ele levanta importantes suposições que ajudam a construir um conhecimento sólido sobre o comportamento das anomalias e ajudam os pesquisadores a implementar uma melhor estratégia de modelagem.

Palavras-Chave: PCA, Sensores de produção, Regressão logística.

1 INTRODUCTION

Preventing failures in oil and gas systems is a big interest for many companies. Detection systems and algorithms can be built to identify when something in the process or equipment is getting away from normal behavior. An effort to improve failure detection with machine learning and artificial intelligence approach is being done through the last decade; the digital evolution of oil fields offers more ways to gather information in real-time (Zang, et al, 2014).

Pressure gauges and temperature sensors are a key part when developing control and optimizing strategies. Data provided by these sensors gives important information about downhole, subsea, or topside conditions, depending on their location (Li and Zhu, 2011). Therefore, linking the output data from pressure and temperature sensors, among others available, with machine learning and artificial intelligence techniques could result in effective detection systems.

Also, oil and gas production requires a good functioning production system. Some failures can occur during this phase, such as flow assurance issues and mechanical problems. Flow instability, severe slug, scaling and hydrate formation are flow assurance concerned, for example. Spurious closure of DHSV and quick restriction in production choke can be caused by mechanical or hydraulic reasons. These anomalies occurrence impact the oil and gas production rate, in more severe cases they can lead to a production shutdown (Luna-Ortiz et al, 2008). So, preventing these failures to happen can save money and time from well intervention and workover operations. One way to preventing a failure is understanding the failure mechanisms and the variables that are most impacted while an anomaly situation.

Information regarding faults and anomalies keeps monitoring and maintenance personal better informed and more capable of taking good plans to solve problems during production operations. The literature shows that sensor data are used to enhance detection systems (Cai *et al.*, 2021; Giro *et al.*, 2021). However, most sensor data are used to diagnose sensor and process faults only, as proposed by Salahshoor, Mosallaei, and Bayat (2008).

Therefore, the present work aims to provide an analysis of production pressure and temperature sensors under fault and normal conditions. The novelty proposed is analyzing subsea and topside sensors' data under production anomalies occurrences to detect when an offshore well leaves its normal production condition and starts to get into a failure mode. In order to improve monitoring systems using information from production sensors, a mixed application of PCA and logistic regression is proposed to study the production sensors' behavior when the production system is under failure mode. Thus, the sensor analysis can support detection algorithms. Combining these two techniques for different types of failures that occur during oil production can provide useful insights about sensors' performance and normal condition identification. The outcomes of the study will help well operations, IT specialists, and oil & gas operations

managers to design a system capable of alarming early signs of anomalous behavior is challenging.

2 PRINCIPLES AND METHODOLOGY

2.1 PRINCIPAL COMPONENTS ANALYSIS (PCA)

PCA is an important tool for extracting features from a dataset and combined with classification methods it can perform pattern recognition and detection tasks (Zhou et al, 2014). PCA also enhances classifier algorithm accuracy, being a very useful technique to be applied with machine learning classifiers (Zhu et al, 2019). Furthermore, PCA was applied in this work to improve visualization and understanding of the data by reducing the variable's dimension. Instead of using all sensors together, they were also combined in subgroups of topside and subsea sensors. In this work, up to 8 production variables were available depending on the type of anomaly studied. The PCA reduced these variables in 2 main variables, combining different sensors. Also, using PCA features it is possible to identify correlations between the data from different sensors.

PCA captures dimensions that present high variance and turn them into fewer dimensions (Yang et al, 2019). At the same time, PCA handles the useless and redundant data from the dataset (Poornima and Paramasivan, 2020). The originals variables that belong to a shaft system X with p dimension are transformed into a new shaft system Z, which is a linear combination of the variables in X as represented by the equation 1 (Matloff, 2017):

$$Z_i=a_{i1}X_1+a_{i2}X_2+...+a_{ip}X_p(X,X) \quad \text{.....(1)}$$

The principal components of shaft system Z are obtained from the eigenvalues of a covariance sample matrix. The covariance matrix is symmetric and is described as follow:

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{np} \end{bmatrix}$$

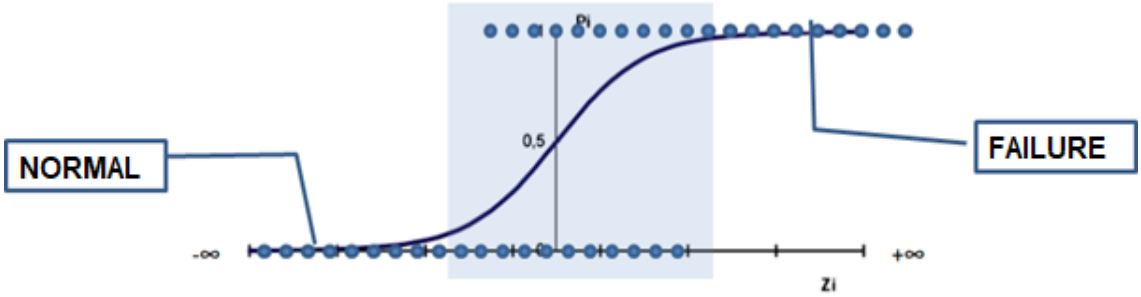
Where the Cjj elements on the main diagonal are the variance of Xi (var(Xi)) and the Cij on the secondary diagonal represent the covariance between variables Xi and Xj (cov (Xi, Xj)).

The PCA simplified the dataset visualization. The abnormalities under study can be developed over time, such as scaling and hydrate production, or they can be spontaneous like spurious closure of DHSV. Flow instabilities and severe slugging depend on the flow conditions. An abrupt increase in BSW occurs when the injection water reaches the production well or when the aquifer reaches the pay zone. Well production depends on many factors, such as reservoir static pressure, BSW content, fluid viscosity, production tubing and flowline diameter, and so on (Devold, 2013). Therefore, alterations in these properties can lead to a lower flow rate and even to a production shutdown.

2.2 REGRESSION LOGISTIC

Another statistical approach often used in classification problems is Logistic Regression (LR), once it can be used for determining probability and for classification purposes (Zhou et al, 2014). The LR aims to explain the occurrence of events when the output variable is binary such as “male” and “female”, “yes” and “no” or “high” and “low” (Nardi et al., 2019). The "normal" and "abnormal" classification are adopted in the present paper by checking the probability of an event occurrence using a logistic function. The predictive variable takes values between 0 and 1, representing which class one unit belongs to. For any Z value between −∞ and +∞, the logistic function shows what is the probability of one unit be part of class 0 (normal) or 1 (abnormal), as shown in Figure 1.

Fig. 1— Classes distribution in a Logistic Regression



2.3 METHODOLOGICAL APPROACH

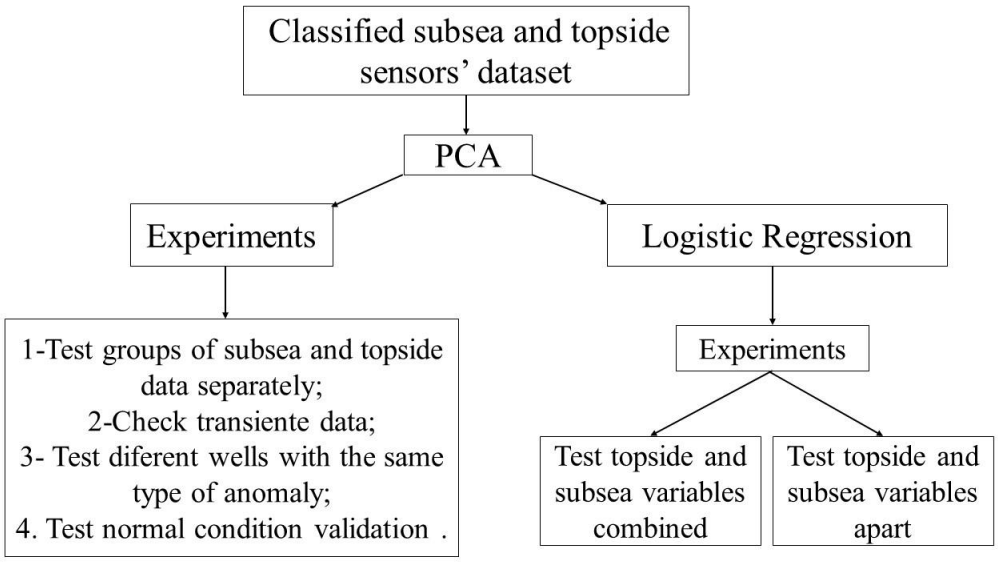
Statistical tools as PCA and logistic regression were applied in a dataset published by Vargas et al (2019) which contains real data from offshore oil wells and is divided by anomaly type. There are seven types of anomalies; abrupt increase in BSW (Basic Sediments and Water), spurious closure of the Downhole Safety Valve (DHSV), severe slugging, flow instability, rapid production loss, quick restriction in the production choke, scaling in the production choke and hydrate in production line; in addition, there is also a file with normal condition data for the different wells available in the dataset. Moreover, depending on the type of anomaly, different sensors could be arranged to better identify the change from normal production condition to an abnormal regime.

Thus, using a single file concatenating all wells with the same type of anomaly were carried out. The experiments also focused on the difference between subsea and topside sensors, transient data evaluation, and well behavior under specific failure modes. After that, according to the best sensor's combination based on PCA and logistic regression algorithms, some assumptions were raised and tested relating to how much time a well stands in a steady normal condition and how wells that contain the same type of anomaly works together. Then, the benefits of using PCA and Logistic Regression for increasing detection capability are discussed. Here, the main goal is understanding when and how pressure and temperature sensors indicate that a well is not producing in its normal condition. For this reason, the transient condition is also regarded as a fault condition because it shows early signals that sensors are out of their normal behavior. The following chart displays the methodological structure of this paper. Using the dataset published by Vargas et al (2019), two sets of experiments were fulfilled. The first regards PCA application and the second combines PCA and Logistic Regression.

The experiments were divided into: sensors analysis, transient data evaluation, analysis of how different wells behavior under the same abnormality type, normal condition validation, and logistic regression application. Starting with sensor analysis, it is shown that there is a possibility of an existing difference between topside and subsea sensors' timing to perceive some anomalies. The evaluation of transient data meant to identify the significance of the transition stage between the normal and abnormal situation in an attempt to identify if there is typical transient behavior for each failure mode. After a better understanding of variables and the influence of transient data, the study proceeded with the analysis about wells' behavior considering each type of anomaly. And, to finish the experiments, an R script executed the logistic regression in the dataset considering all

classes (normal and not normal behavior) to better detect when the system goes outside the normal situation region.

Fig. 2— Methodological flowchart



3 DATA ORGANIZATION

All data used to run an analysis with the Statistica software are from the work of Vargas et al (2019). The paper brought a realistic dataset with rare undesirable real events that happen during the production of oil wells. Vargas et al. (2019) generated a dataset for applications in machine learning algorithms, providing more data available for abnormal events management.

The variables are measured in sensors located in the production tubing (P-PDG), on the subsea Christmas tree (P-TPT and T-TPT), production line (P-MON-CKP and T-JUS-CKP), and gas lift line (P-JUS-CKGL, T-JUS-CKGL, and QGL). These variables correspond to pressure at the Pressure Downhole Gauge, pressure and temperature at Temperature and Pressure Transducer, pressure upstream the PCK, temperature downstream the PCK, pressure downstream gas lift choke, temperature downstream, and gas lift flow, respectively. The authors published a dataset about 8 rare undesirable events that occur in oil wells such as abrupt increase of BSW, spurious closure of DHSV, severe slugging, flow instability, rapid productivity loss, quick restriction in PCK, scaling in PCK, and hydrate in a production line. The total number is 21 of real and 939 for simulated wells from OLGA considering that a different simulation needs to be done for each problem (Vargas *et al.*, 2019).

4 EXPERIMENTS

As shown in Figure 2, the experiments carried out were designed to test the following hypotheses:1- Are subsea sensors able to detect anomalies, such as hydrate formation, first than topside sensors? 2- Does the transient information, data between normal and abnormal status, provide useful knowledge about the anomaly that is taking place? 3- How different wells behavior when presenting the same type of anomaly? 4- The normal condition stays valid for how much time?. Then, the Logistic Regression was applied in the PCA dataset with both topside and subsea sensors. In the attempt to improve the result obtained by the model, subsea and topside sensors combined by PCA were also used separately.

5 RESULTS

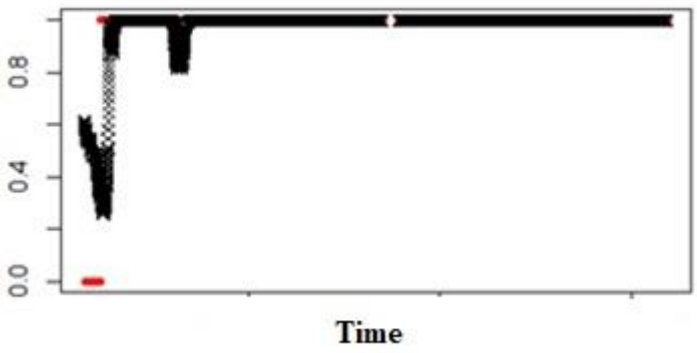
The results obtained by following the workflow shown in Figure 2 are displayed in the next five sections. The first section addresses the analysis of subsea and topside sensor's reaction time using logistic regression. The second section shows the usefulness of transient data, it presents a PCA application for two wells with hydrate formation problem. In the next section, using datasets composed by different wells with the same production anomaly, it verified if these wells will present close clusters of data, showing then similarities, that would enable to characterize the type of failure based on well information. Moreover, in order to check how much the well's normal standard changes during the production time, one section discussing the validity period of the normal condition data for a given well is also presented. Furthermore, combining the datasets transformed by PCA and applying the logistic regression algorithm, the results show the impact in logistic regression classification using all sensors mixed and separating them in groups.

5.1 SENSOR'S ANALYSIS

The sensors studied by this work are from topside sensors (QGL, P-MON-CKP, T-JUS-CKP, P-JUS-CKGL, T-JUS-CKGL) and subsea sensor (PDG, P-TPT, T-TPT). For hydrate formation problems, the accumulation of crystals occurs in low temperature and high-pressure environments. During offshore oil and gas production, hydrate formation is observed in the subsea region. Then, applying LR in the complete dataset it is possible to check when the initials signs of hydrate blockage are detected. Figure 3 shows the data variability for topside and subsea sensors applied in logistic regression,

the vertical label presents the empirical probability of an input data belongs to normal (0) or abnormal (1) conditions. In the temporal manner, the firsts evidences of hydrate formation are registered by the subsea sensors as depicted by the first graphically anomaly. The second anomaly represents the impact of hydrate blockage in topside sensors. The logistic regression model in red is mostly overlayed by the data itself. Initially, the model identify the well normal condition and, after an anomaly is registered in the subsea sensors, the LR model assigns the remaining data as abnormal condition.

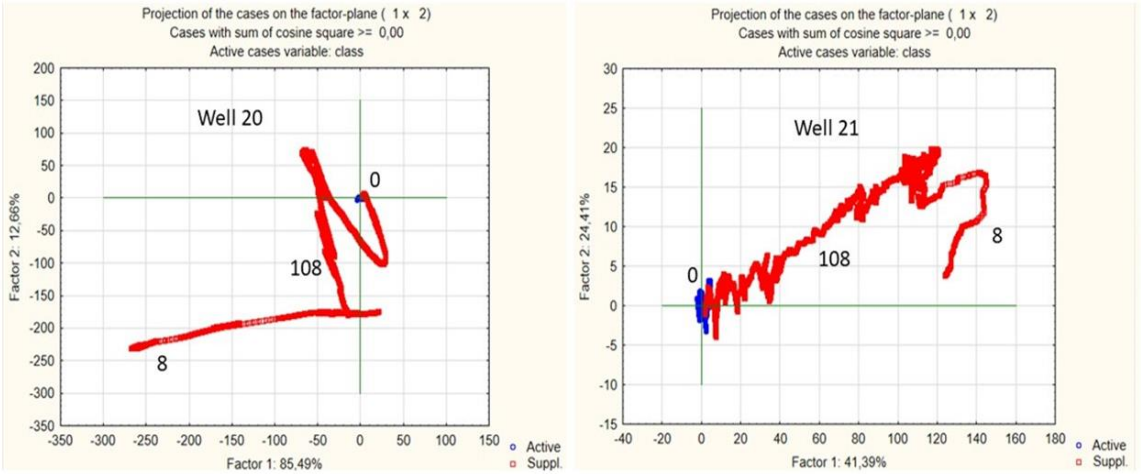
Fig 3– Subsea and topside variables behavior with time from hydrate formation data. The y-axis represents the empirical probability provided by logistic regression, while the x-axis shows the time recording.



5.2 TRANSIENT DATA EVALUATION

Splitting sensors in topside and subsea groups or using all variables together, according to each type of abnormality, it is possible to evaluate the influence of transient data. Figure 4 relates the transient data of wells 20 and 21, two wells containing hydrate formation, with all variables (QGL, P-MON-CKP, T-JUS-CKP, P-JUS-CKGL, T-JUS-CKGL, PDG, P-TPT, T-TPT) put together. The figure axes display factors 1 and 2 from the PCA combination of variables, these two factors are responsible for explaining most of the data variability. As it is shown, the transient data do not display a pattern. The same result can be seen while separating subsea and topside variables from these two wells. Also, it can be implied that well 20 shows more variability than well 21. The scale for these two wells is very different, some hypotheses are that they have different geometry, production fluid, production time or they are in different reservoirs. However, the dataset published by Vargas et al (2019) does not include specific information about the wells used.

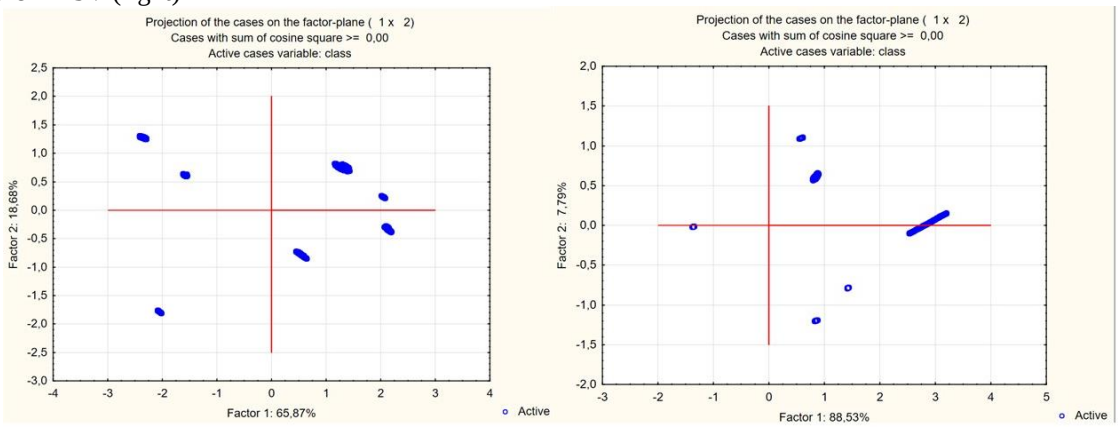
Fig. 4 -Transient data comparison between two wells containing hydrate problem with all sensors data together. The axes display factor 1 and 2 from PCA combination of variables.



5.3 DIFFERENT WELLS UNDER THE SAME ABNORMALITY

For the analysis of different wells with the same type of problem, it was considered just the "abnormal" class denominated by the problem number. The scaling in PCK dataset did not have the "abnormal" class, just "normal" and "transient" for more than one well, so it is not possible to compare two wells under this condition. Excluding the problem of scaling in PCK, six experiments showed that even for different wells, disregarding the normal condition for each of them, the abnormal class does not vary significantly as displayed in Figure 5. On the left side, it shows seven wells, represented by the blue cluster, with flow instability issues, and, on the right, six wells showing problems in the DHSV.

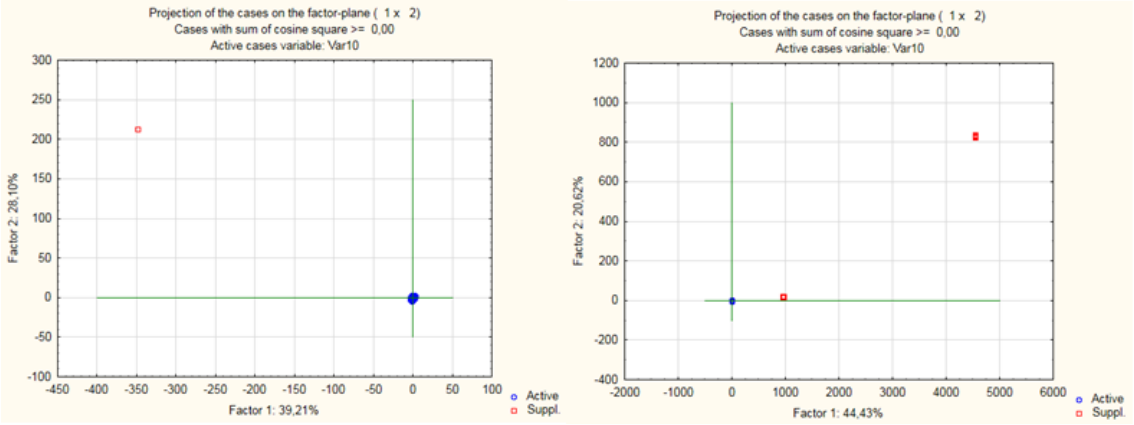
Fig. 5 – Seven different wells with flow instability problem (left) and six different wells with a problem on the DHSV (right)



5.4 NORMAL CONDITION VALIDATION

Oil and gas reservoirs change their pressure and production rate during production time. With more fluids out the reservoir less static pressure it presents, decreasing the production rate (Rosa et al, 2009). Because of this change in pressure level, it is important to update periodically the normal condition status. Figure 6 shows how the normal condition changes in months of production. The first plot shows the same well with data from almost one month apart, clusters blue (reference) and red, presenting modification in its normal condition. The second plot adds more data from another month. As the x-axis indicates the difference between these three clusters of the normal condition is growing apart in distance.

Fig.6 – Normal condition shifting through production time.



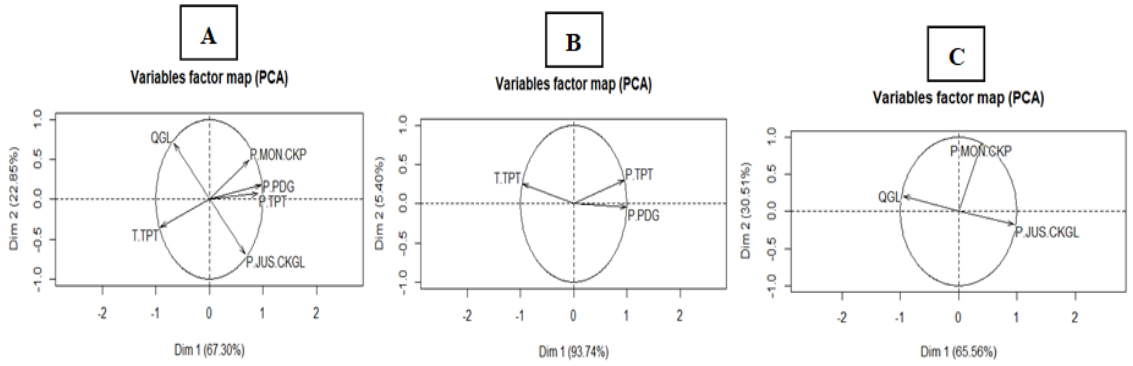
5.5 LOGISTIC REGRESSION APPLICATION RESULTS

The scope of this work is to present assumptions that can improve significantly the modeling strategy for anomaly detection in oil wells. The dataset used here includes three classes “normal”, “transient” and “abnormal”. The main goal of a detection algorithm must be alarm when the data step out of the normal condition of a given operation. Therefore, the "transient" and "abnormal" classes were mixture together, leaving the algorithm to detect everything that is not a "normal" class. The Logistic Regression is a technique applied for binary classification as shown in figure 1; it provides the probability of a given point being part of a specific cluster or category. The categories for our analyses, after combining "transient" and "abnormal" classes in one are "normal" and "abnormal". Logistic regression was applied to each abnormality that presents the classes of interest.

5.6 LOGISTIC REGRESSION APPLIED TO PCA: SUBSEA AND TOPSIDE SENSORS EXPERIMENT

The variables combined by PCA were used as an input into the Logistic Regression model. Figure 7 shows three variables factor maps regarding hydrate problems. These maps provide a projection view of the observed variables into the plane spanned by the first two principal components. The projection view highlights the structural relationship between variables and the components; it allows one to read directly from the map the correlation between the variable and the component. The first map (Fig7.A) takes all variables available and plots their correlation with the first two components. Thus, one can see that the first principal component explains about 67% of the total variation, and the second principal component an additional 23%. The first two principal components explain about 90% of the total variance. The first component correlates almost perfectly with the variables P.PDG and P.TPT. Meanwhile, the second component correlates directly with Q.GL and inversely with P.JUS.PCK; suggesting that there is a correlation with these groups of variables. Then, after splitting the variables, one set of subsea variables was plotted in a variables factor map, as shown in Figure 7.B. The new variable set resulted in a different combination of components, now the first component explains around 94% and the second component explains about 5%, totaling 99%. This shows that using this new set of variables covers better the data variability than using all data available. On the other hand, figure 7.C presents the group of topside variables and the correlation of the principal components. The total variance explained was about 96%, also a better performance than using all variables together.

Fig. 7- PCA Variables map for; (A) all variables, (B) subsea variables, and (C) topside variables regarding hydrate formation



After the PCA analysis, the variables were tested with logistic regression to see if splitting variables in groups would impact the classification. The use of all variables combined by PCA did not improve the model, indicating that at least one sensor is not contributing positively in the modelling. Indeed, using all sensor together caused undesirable fluctuation. The model created by combining all sensors showed when the normal condition is over and starts to achieve the abnormal level, as displayed in Fig. 1, however, at some short and nonconsecutive intervals, it states the operation is under normal condition while it is showing failures evidence. This behavior in a detection system generates false negative alarms. Therefore, a dataset separating subsea sensors from topside sensors were used in an attempt to improve the modelling. When using just subsea sensors, the model classifies better normal and abnormal conditions, without false negative intervals, for hydrate problems. Furthermore, splitting the variables into topside and subsea variables can be an interesting strategy for building fault detection systems regarding hydrates problems, for example. The workflow presented in Figure 2 regarding the LR application can be tested for others anomalies and verify whether combining every sensor available will increase the model accuracy or not. Moreover, testes separating smaller groups of sensors can be also carried out to isolate sensor causing undesirable disturbances.

Considering just topside variables, the model does not show two distinct levels, normal and failure, with false positive or negative intervals, it presents random behavior. This suggests that the topside variables included in the PCA dataset may have a factor that disturbs the classification. Splitting topside variables in pars and testing the LR model, the gas lift rate and pressure shown to be good variables for the classification model. Other tests have shown that P.MON.CKP is not a good variable for hydrate detection, maybe it presents an erratic behavior because of unstable flow conditions through the sensor.

Different outcomes were observed when applying the same methodology for high BSW and DHSV failure data. The PCA variables map for BSW shows a different correlation between the variables available. In this case, there is no direct correlation within subsea or topside variables that suggest they should be break into two groups. The first component combines subsea and topside sensors of temperature and pressure, explaining 68% of the total variance, while the P.JUS.CKGL stands alone for the second component which explains 20% of data variability. Then, using the LR to classify BSW failures with the PCA variables as input presented good results. Similar to Fig. 1, the

normal region migrated to the fault region without disturbance. For the BSW case, it was also tested the possibility to work with subsea and topside variables apart, the results exhibit that even using sub-groups of subsea or topside variables or mixing them the LR classification still shows good performance. Also, for DHSV failure data variables from topside and subsea sensors worked together in the LR model. PCA variables' map for this type of anomaly has shown that the first principal component takes pressure and temperature variables from both locations and the second component accounts only for temperature in the production choke.

6 DISCUSSION

According to Figure 3, the change in data status for subsea sensors occur first than the topside sensors for hydrate formation problems, as represented by the first drawdown. This fact can be related to the physical nature of the abnormality, some of them can take place near the subsea ground condition and subsea sensor can get them first. For this reason, further analyses were carried out dividing the dataset into topside and subsea variables and testing the impact on using two groups of sensors. Subsea and topside variables drift in time; depending on the anomaly type, it can be first sensed in one set of variables and after some time the other set would get the fault's signal. Therefore, for hydrate detection would be interesting to build a warning system with two independent approaches: an alarm based on subsea variables and another based on topside variables to complement each other.

In Fig.4, the data variability is plotted by time does not go in a single direction. The curve plotted goes forwards and backward in an unpredictable matter. Because of this behavior, it is not easy to predict which abnormality is going to take place during the oil production just using the transient data from wells.

Different wells under the same type of problem present similarities, shown by the distance between each cluster printed in Figure 5. It displays low variability between several wells; the distance between clusters is very low compared to the scale seeing before. Also, all the other abnormalities show the same behavior. The maximum variability can be observed at the production loss problem; however, this distance is just a few units higher than presented in Fig. 5.

To build a model for anomaly detection it necessary to know what is the normal behavior of a given well. Suppose the model used to detect anomalies has not to update on well normal condition, the clusters of the normal condition can be shifted and, then,

mistaken with an anomaly occurrence, when the case is that the well has new normal condition parameters, as pressure and temperature levels. Updating the well's normal condition within the model means keeping its ability to identify failures at any time of well production life.

After testing the hypotheses above, Logistic Regression was applied to each type of anomaly that presents the classes of interest. As output, the model detected, whether the system was in normal condition or it was showing signs of abnormal behavior. The first step to implement LR model was to understand how topside and subsea variables were connected in each case by using the PCA variables' map. For hydrate formation problems, the best approach was splitting the dataset to improve LR classification. For BSW and DHSV problems, for example, the use of both groups did not interfere in LR outcome.

7 CONCLUSIONS

The methodology presented in this paper showed that PCA can be used to determine distances from real-time data to a specific failure cluster. In some real cases, clusters with failure data may not be available. Mostly because some wells will take a long time until developing an anomaly, scaling for example, or for new wells there will be no failure information available to train models. To overcome this obstacle, it is recommended to train models with the normal class only. For every well, engineers and operators will have in hands information about the well normal condition. Also, it is important to update this cluster periodically, once normal conditions in a well can change with reservoir depletion.

The PCA coupled with Logistic Regression can provide an outlook about which is the best set of variables for detection and classification purposes. For BSW and DHSV failures it showed that subsea and topside variables could work together without disturbing the LR outcome. On the other hand, for hydrate formation cases PCA showed correlations within subsea topside variables, endorsing that working with these two groups apart could be a better alternative. In this case, an advantage of using the separated alarm for the subsea and topside variables is that mature and new wells will both have at least topside variables. For old wells, subsea variables will not be reliable or available in some cases. On the other hand, the drawback is that instead of just a model, the operator will have two models to update, training and test.

ACKNOWLEDGMENTS

The authors are grateful for all the support given by Petrobras through the R&D Project nº 21354-6, CAPES, PPGEP, PPSIG, and PGMEC-UFF.

REFERENCES

Cai, B.; Hao, K.; Wang, Z.; Yang, C.; Kong, X.; Liu, Z.; Ji, R.; Liu, Y. (2021). Data-driven early fault diagnostic methodology of permanent magnet synchronous motor. Expert Systems With Application 177, 115000.

Devold, H. Oil and gas production handbook: An introduction to oil and gas production, transport, refining, and petrochemical industry. 3rd Edition. Oslo, 2013.

Giro, R. A.; Bernasconi, G.; Giunta, G.; Cesari, S. (2021). A data-driven pipeline pressure procedure for remote monitoring of centrifugal pumps. Journal of Petroleum Science and Engineering 205, 108845.

Li, Z., Zhu, D. Optimization of production performance with ICVs by using temperature data feedback in horizontal wells. SPE Production and Operation, 253-261, 2011.

Luna-Ortiz, E., Lawrence, P., Pantelides, C. C., Adjiman, C. S., Immanuel, C. D. An integrated framework for model-based flow assurance in deepwater oil and gas production. ESCAPE 18th, 2008.

Martloff, N. Statistical regression, and classification from linear models to machine learning classification. CRC Press, 2017.

Mask, G., Wu, X., Ling, K. An improved model for gas-liquid flow pattern prediction based on machine learning. Journal of Petroleum Science and Engineering 183, 2019.

Nardi, I. R.; Nascimento, S. C. C.; Costa, J. B. C.; Souza, A. M. (2019) The development of a mathematical model for the forecast on approval in calculus 1 using logistic regression. Brazilian Journal of Development vol. 5, n. 10, 22245-22256. DOI:10.34117/bjdv5n10-352.

Poornima, G. A., Paramasivan, B. Anomaly detection in wireless sensor network using machine learning algorithm. Computer Communications 151, 2020.

Rosa, A., Carvalho, R., Xavier, J. Engenharia de reservatórios de petróleo. Editora Interciência, 2009.

Salahshoor, K., Mosallaei, M., Bayat, M. Centralized and decentralized process and sensor fault monitoring using data fusion based on adaptative extended Kalman filter algorithm. Measurement 41, 2008.

Vargas, R. E. V., Munaro, C. J., Ciarelli, P. M., Medeiros, A. G., Amaral, B. G., Burrionuevo, D. C., Araújo, J. C. D., Ribeiro, J. L., Magalhães, L. P. A realistic and public dataset with rare undesirable real events in oil wells. Journal of Petroleum Science and Engineering, v. 181, 2019.

Yang, K., Hu, B., Malekian, R., Li, Z. An improved control-limit-based principal component analysis method for condition monitoring of marine turbine generators. Journal of Marine Engineering and Technology, 2019.

Zang, H., Li, J., Khor, S. H. Integrated management strategy of flow assurance for digital fields. SPE – Annual Technical Conference and Exhibition. Amsterdam, 2014.

Zhou, C., Wang, L., Zhang, Q., Wei, X. Face recognition based on PCA and logistic regression analysis. Optik 125, 2014.

Zhu, C., Idemudia, C. U., Feng, W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked 17, 2019.