Spatiotemporal Anomaly Detection Applied to Flow Measurement Points in Natural Gas Production Plants

Hadriel Toledo Lima, Flavia Bernardini

Universidade Federal Fluminense, Brazil
hadriellima@gmail.com, fcbernardini@gmail.com

Abstract. In an oil production unit, the volume of production is measured by Measurement Points distributed throughout the process plant. The distribution of these points is designed for operational monitoring of the plant, and ensure that all produced fluid is measured. In 2000, ANP and INMETRO published the Technical Regulation of Measurement (TRM), establishing requirements for measurement systems in the production units. Insurance of the measured value of oil and gas became not only legal but also operational issue. To facilitate monitoring problems in Flow Measurement Points (FMPs), this work proposes a method based on Dynamic Bayesian Networks for Detection of Anomalies in values reported by FMPs. This method explores the relationship among volumes reported by FMPs at an instant of time, and temporal relationship of the values of a Measurement Point. Two experiments were carried out with the proposed method using real data from a production plant. The first one aimed at evaluating the impact of variating parameters of the method on the predicted values, reported by a FMP. The second experiment aimed at verifying the effectiveness of the modeling in detecting anomalies, using the parametrization that obtained the best performance in the first experiment. The method is promising because it was able to identify most of the anomalies present in the used data set.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data mining; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: Anomaly Detection, Dynamic Bayes Net, Flow Measurement Points

1. INTRODUCTION

The ANP — National Agency of Petroleum, Natural Gas and Biofuels — is the Brazilian regulator of activities in oil, natural gas and biofuels industry in Brazil. In 2000, ANP published a regulation term, called Technical Regulation of Measurement — TRM —, jointly to INMETRO — National Institute of Metrology, Quality and Technology — that regulates the process of measuring oil and natural gas in Brazil [ANP and INMETRO 2000]. This regulation established administrative, technical, metrological and operational requirements to account produced and transported volumes. This control is important for guaranteeing correct collection and distribution of government transfers to states and cities. Updates on the term occurred in [ANP and INMETRO 2010] and [ANP and INMETRO 2013], though the main purpose of accurate and complete results in production measurement was maintained. Thus, after the TRM, detecting anomalies in measured values of oil and natural gas became not only operational but also a legal issue.

In a production plant, produced and transported fluids flow throw ducts. The measurement of these fluids is done by Flow Measurement Points (FMPs). These points are distributed in the production plant, in order to ensure that all produced and transported fluid is measured. So, this plant forms a network of production and distribution points, ducts and FMPs. This net may be represented by a directed graph. Figure 1 shows a directed graph representing an instance of a natural gas plant, from

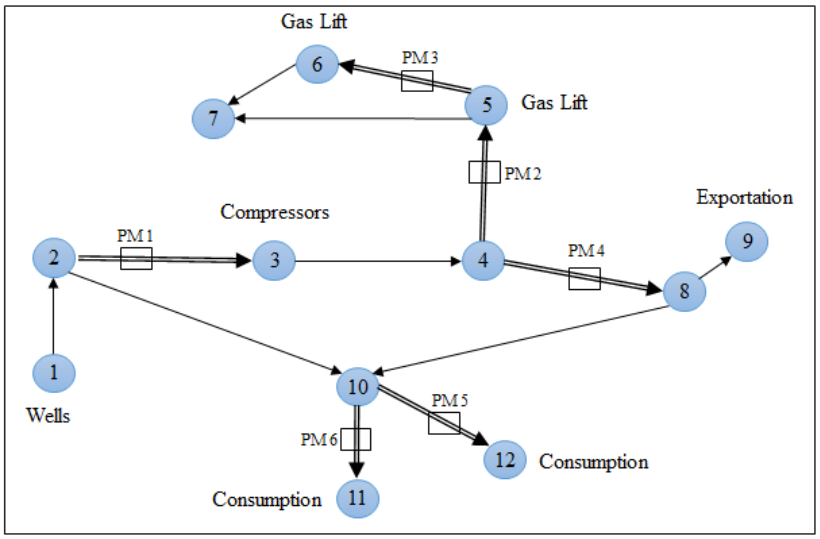

Fig. 1. A network of production and distribution points, ducts and FMPs represented by a directed graph

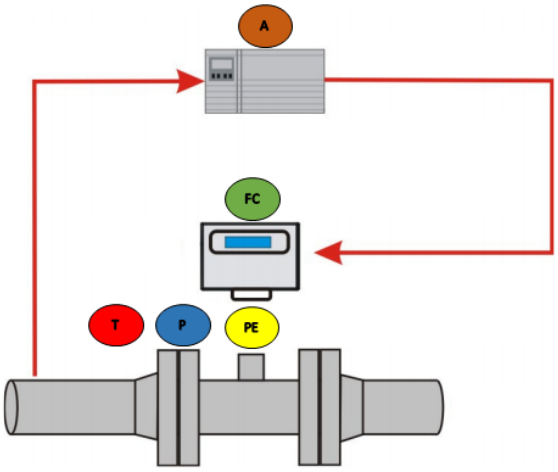

Fig. 2. Schematic of installation of a measuring point. Adapted from [Lazari et al. 2009]

which the data was collected for experimental analysis in this work. In this graph, edges represent ducts through which the fluid flows and nodes represent points of interest for measurement in the plant. There are special edges, represented by double dash, indicating that, in these ducts, there is a measuring point in each edge. Also, there are four special groups of nodes, which affects the network. These groups are called Wells, where the fluid (natural gas) is produced; Exportation, where the fluid is sent to another process plant or pipeline; Gas Lift, where the gas is used in a method called artificial lift; and Consumption, where the gas is consumed in the production plant itself, such as in power generators, among other equipments. The net shown in Fig. 1 contains one node of the type Wells, represented by (1); one of Exportation, represented by (9); two of Gas Lift, represented by (5) and (6); and two of Consumption, represented by (11) and (12). The other nodes are of the type bifurcation or junction of the fluid. Nodes represented by (2), (4) and (10) are bifurcations, and (7) and (8) are junctions.

According to [Lazari et al. 2009], a FMP is composed by a set of instruments, shown in Figure 2. The Primary Element, represented by PE, is located directly in the duct, having direct participation in flow rate measurement. When PE is of "Orifice Plate" or "V-cone" type, for instance, PE calculates flow rate in the duct through differential pressure. Analyzers, represented by A, or samplers, are instruments that influence the final measurement result, calculating the percentage of water and sediment in relation to the total volume of fluid produced. Pressure and temperature measurements are done by instruments represented, respectively, by P and T. Flow Computer, represented by FC, calculates flow rate and volume of the fluid from values provided by all the instruments that compose the system, *i.e.*, PE, A, P and T.

Although most of the time FMPs properly work, sometimes they may present some common problems, regarding to variation in measured flow: (i) the measured value remains in zero; (ii) the measured value "freezes", that is, remains in a fixed value; and (iii) the measured value does not represent the real value, and so a deviation is presented in the measurement. Problems (i) and (ii) are easily identifiable, but (iii) is more difficult to detect, especially when the variation is small. So, a computational tool able to identify anomalies from FMPs, and to alert those responsible for its treatment, is useful to avoid non-compliance with TRM. However, there are three major challenges for implementing this tool, that are (i) temporal distance of the flow among the FMPs (a fluid takes time to traverse the production plant), and temporal dependence among data of various FMPs; (ii) variation of the stock in the duct; and (iii) imprecision of the instruments.

In literature, the type of problem tackled by this work is called Anomaly Detection. It includes problems related to finding patterns that depart from normal (expected) behavior. In many domains, anomalies in data translate real anomalies [Chandola et al. 2009]. Anomaly Detection in measuring networks is one of these domains. One commonly used approach for Anomaly Detection is regarded to using Machine Learning algorithms to recognize normal and abnormal patterns of operation [Russell et al. 2003]. Problems where instances present in the available data are labeled with abnormal or normal operation, supervised learning algorithms for binary problems can be used. In the available data, an anomaly is registered per day, but the exact time that the anomaly occurred is not available, turning impossible to use binary classifiers. The method should detect anomaly in periods shorter than 24 hours, and this is another feature that reinforces the infeasibility of this kind of supervised learning. On the other hand, problems where only negative (normal) instances are presented, due to rarity of anomaly, one-class classification algorithms and methods can be used [Tax and Duin 2001]. Initially, the problem was modeled in a way that each node was represented by a feature. In this case, all domain features are the same, and one classifier would be constructed for each FMP. In this case, each dataset to construct each classifier would only change the class values. Also, one prerequisite to use classical supervised machine learning is that the domain features are statistically independents. This fact does not seem to be the case in the tackled problem. One characteristic of algorithms to find the structure of Bayesian Networks is finding the probabilistic relations between pairs of nodes [Murphy 2002], and this motivated its use, regardless of the difficult of modeling and implementing the solution, as there are not tools that facilitate this task, on the contrary of using supervised learning algorithms for binary and one-class classification. This was specially strengthened for the fact that no work was found that modeled the problem of detecting anomalies in FMPs, specially on natural gas production plants, and the most similar tackled problems founded are [Dereszynski and Dietterich 2011] and [Wang et al. 2008]. Both use Bayesian Networks to model the anomaly detection problem, but the data used by these two works present less variation than the data of gas flow rate. The collected data allowed to observe that considerable variations in flow rate can be found in every minute. So, the key research question that this work answered is related to the efficacy of using Bayesian Networks to model anomaly detection in this kind of data, as other approaches using machine learning are not easy to be used.

This article aims to propose a method for identifying anomalies in natural gas measurement points in production plants using Hybrid Bayesian Networks (HBN). The proposed method learns the re-

lationships between measurement points from data and so does not require the knowledge of any specialist in the operation of the plant to be built and can be adaptable to any natural gas production plant. Section 2 presents a literature review on anomaly detection, and described the nature of data, which deeper explains the decision to use Bayesian Networks. This section is important to allow comprehending the method proposed in Section 4. For matter of computational optimization, an adaptation was made in traditional HBN modeling in our problem, which is considered a contribution of this work, showing its feasibility even being simpler than the traditional method used by [Dereszynski and Dietterich 2011] and [Wang et al. 2008]. Real data from natural gas measurement points were used for experimental analysis. Such plants present a large number of FMPs, which turns monitoring more difficult, and even more failures may occur. Experimental analysis and the obtained results are shown in Section 5. Finally, Section 6 concludes this work, discuss some fragile points of the proposed method, and presents future work. It should be observed that, although not tested in the scope of this article, it is expected that this modeling can also be used at petroleum measurement points, because petroleum flow data present less variance than gas flow.

2. NATURE OF DATA, CONCEPTS AND LITERATURE REVIEW

2.1 Nature of Data in FMPs

The Flow Computer (FC) of a FMP, previously described, can be configured to provide various pieces of information about the flow at a FMP. One of them is the instantaneous flow rate, measured at a time instant t. Other information reported by the Flow Computer is the volume accumulated in a period of time $k = \Delta(t)$, which is obtained from the sum of flow rate in k.

Petroleum and Natural Gas are fluids that variation in pressure and temperature can cause variation in volume. Therefore, volume information without indication of pressure and temperature at the moment of measurement is deficient. So, Corrected Volume value refer to the volume of fluid converted to normalized conditions, i.e., temperature $20°C$ and pressure $101.325kPa$. In this article, volume value with no reference to pressure and temperature values refers to the corrected volume.

The FC also reports Corrected Volume value, i.e, an accumulator that increases the volume passed by the FMP every second. From this value it is possible to extract several accumulated volume intervals. However, for developing this work, there was no indication from specialists of the appropriate interval to use in our method. So, this interval k is the accumulated Corrected Volume interval, in minutes.

It is worth to note that, in our problem, anomaly records are associated only to the day they occurred, and there is no relation to the exact time instant. Flow rate data is collected in seconds (and minutes), but the anomaly record is related to periods of 24 hours. In other words, an anomaly is registered per day, but the exact time that the anomaly occurred is not available. Also, an important issue of our domain is the imbalance among (i) time that a FMP works with anomaly and (ii) time of normal operation. For instance, in the dataset considered in our experiments, in only 11 of the 180 consecutive days of data collected, anomaly occurred. Thus, it is observed that the measurement points works properly most of the time, and the anomalies are exceptions, as usual in this kind of problem. So, there are (many more) instances of normal operation of the FMPs than anomalies.

The dataset used to experimental analysis in this work was collected from a natural gas production plant with 6 Measurement Points, PM1, PM2, PM3, PM4, PM5 and PM6. Figure 1 shows a representation of this plant. Spearman Correlation Coefficient [Hauke and Kossowski 2011] was calculated in order to discover if there is high correlation between the Measurement Points in the collected dataset. Spearman Correlation Coefficient values are presented on Table I. Only correlation between PM1 and PM2 is high (greater than 0.7), while the others are low. Independence was also tested by d-variable Hilbert Schmidt Independence Criterion, a nonparametric measure of dependence between continuous variables. The independence was rejected by this test, leading to use Bayesian Networks in this case.

Table I. Spearman Correlation Coefficient among Measurement Points dataset

	PM2	PM3	PM4	PM5	PM6
PM1	0,90	0,46	0,63	-0,12	-0,31
PM2		0,47	0,37	-0,19	-0,24
PM3			0,27	-0,15	-0,07
PM4				-0,04	-0,20
PM5					0,02

2.2 Spatiotemporal Anomaly Detection

In the problem tackled in this work, it is difficult to determine a period of time which represents normal behavior in the FMPs. One key difficult is determining elapsed time in the passage of fluid through the FMPs. This issue is similar to problem domains presenting spatiotemporal characteristics. In other words, these domains belong to spatiotemporal anomaly detection category. Spatiotemporal Anomaly Detection is applied in domains whose dataset has a spatial component, which represents a spatial reference for all domain variables, and a temporal component, which is composed by a time series where the data is consecutive and equally spaced in time [Das et al. 2009]. The most important characteristics of spatiotemporal datasets are (i) autocorrelation or non-independence; (ii) heteroscedasticity; (iii) spatiotemporal relationship, i. e., relationship that combines the spatial and temporal components; and (iv) large data volume [Das et al. 2009].

[Paschalidis and Smaragdakis 2009] worked on an anomaly detection problem with spatiotemporal data. Their approach is based on evaluating large deviations from normal operations to empirical measurements. However, the deviations of values at FMPs cannot be large, which made this work less relevant for our work. [Kut and Birant 2006] used a three-step approach in a data domain with spatiotemporal characteristics. The authors used clustering, spatial neighbors checking, and temporal neighbors checking. The goal is to find small groups of data objects that are exceptional, when compared to most data. The purpose of them is to find individual anomalies, so this approach was also not interesting for our work. On the other hand, there are some articles that only use the temporal relation of the data for anomaly detection. [Ma and Perkins 2003a] and [Ma and Perkins 2003b] model the temporal sequence using a regression algorithm with support vectors to detect novelties in temporal sequences. [Akouemo and Povinelli 2016] used a probabilistic approach to detect anomalies in natural gas time series. Different from the problem tackled in this article, they had more detailed information about the anomalies data and, therefore, it was possible to use a Bayesian classification algorithm for model learning.

[Dereszynski and Dietterich 2011] and [Wang et al. 2008] tackled problems with characteristics similar to ours. The first one proposed a quality control system for air temperature sensor data. The second one aims to detect flaws in toxic gas sensors in a coal mine. Both have in common the use of Bayesian Networks to construct the model, but the data of the two works present less variation than the data of gas flow rate, tackled in this work, which presents considerable variations in every minute. The model proposed by [Dereszynski and Dietterich 2011] uses two main components: (i) the spatial component, which represents the relation among the measurement points at an instant of time; and (ii) the temporal component, which represents the transition from one instant of time to another. The spatial component was built using a learning algorithm for learning Bayesian Networks, using Hill Climbing algorithm and the BGe score metric. The temporal component was built by relating each variable that represents the reading of the current time instant of a sensor, with a variable that represents the reading at the previous time instant (lag variable). [Wang et al. 2008] proposed a model that considers an anomalous event in the time series of data, a value discrepancy of the time series likelihood values. The approach used by these authors to their tackled problem is similar to the

one made by [Dereszynski and Dietterich 2011], distinguished mainly by, instead of using a learning algorithm for learning the structure of the Bayesian Networks, the structure was built by an expert domain. Due to the similarity between the domain studied in this work and the domains presented in [Dereszynski and Dietterich 2011] and [Wang et al. 2008], and also , for the other works previously described, Bayesian Networks (BN) was the better indicated approach to detect anomalies in FMPs of natural gas.

2.3 Bayesian Networks

BNs (or Bayesian Belief Networks, Causal Networks or Probabilistic Networks) are used to represent uncertain knowledge [Guo and Hsu 2002]. Russel and Norvig [Russell et al. 2003] define RBs as a directed graph, where:

—Nodes V represent random variables X.
—Direct Arcs (X, Y) represent the intuitive meaning of an arc from X to Y where X is a parent node and Y is a child node, and X directly influences Y.
—Each node has a Conditional Probability Table (CPT), which stores how much parent nodes influence the child node.
—The graphs have cycles not directed, but do not have directed cycles, and for this reason it is an Directed Acyclic Graph — DAG.

BNs provide a compact representation of the Joint Probability Distribution (JPD) of a set of variables [Guo and Hsu 2002]. JPD is a way of obtaining the probabilities for each possible combination in a set of variables [Trifonova et al. 2015]. So, a Bayesian Network is considered a visual representation of the JPD, including all the variables that involve the problem. Also, BN can be considered a probabilistic knowledge base, represented by the network topology and the CPT of each node [Guo and Hsu 2002]. The main function of a knowledge base is to compute a response on the domain, i.e., to inference on the domain. Inference in a BN, named Bayesian Inference, is the process of determining information based on the conditional probabilities of any hypothesis given the available data, through application of the Bayes rule [Alves et al. 2015]. Bayesian Inference algorithms are divided into two groups [Guo and Hsu 2002]: (i) **Exact Inference**, consisting of calculating the exact probability values of the complete a posteriori distribution [Alves et al. 2015] — in general, this inference presents exponential execution time to the width of graph, and so in many cases the exact inference is not feasible [Russell et al. 2003] —; and (ii) **Approximate Inference**, used in cases which exact inference is not feasible. In this case, accuracy is sacrificed, and replaced by an approximate description of the posterior distribution. This work uses approximate inference.

Bayesian networks were originally constructed to deal with problems described by features which domains are discrete. However, there are problems like the one of this work where feature are continuous random variables or combination of continuous and discrete features, named Hybrid Bayesian Networks [Driver and Morrell 1995]. Some approaches transform continuous variables to discrete ones, to treat them as discrete variables [Maldonado et al. 2015]. However, these approaches causes loss of information and may adversely affect the results of the model. There are other proposals to work with continuous variables, such as Mix of Polynomials (MoP) [Shenoy and West 2011], Mix of Truncated Base Functions (MoTBFs) [Langseth et al. 2012], and others. Also, time series modeling by BN is possible by a BN extension known as Dynamic Bayesian Networks (DBNs). The nodes represent the variables at given time intervals [Trifonova et al. 2015]. A Dynamic BN means that it is used in dynamic systems, and it is not necessary that the network changes with each instant of time [Murphy 2002].

In this work, the approach used to treat continuous variables was the same used by [Dereszynski and Dietterich 2011], named Conditional Linear-Gaussian Network (CLG), which considers that all

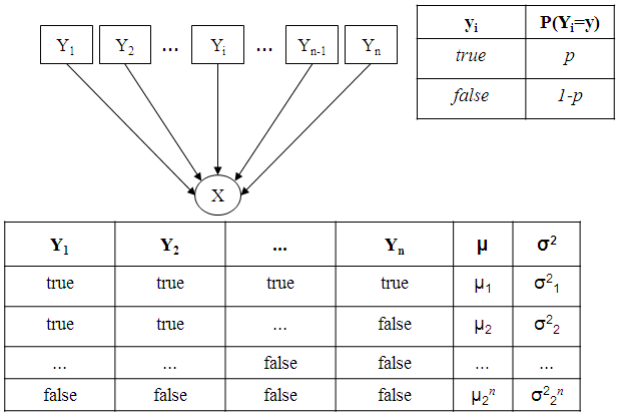

Fig. 3. An example of Bayesian Network with a continuous variable with a set of discrete parents

continuous variables are modeled as Gaussian. The choice is based on the similarity between the problem addressed in this work and the problem study in [Dereszynski and Dietterich 2011]. In this type of Bayesian Network, the probability distribution of the random variables has the following parameterization [Dereszynski and Dietterich 2011]:

(1) **Continuous variables with discrete parents**
Consider a continuous random variable X. For each possible combination of the discrete parents of X, X assumes a Gaussian distribution with separate values of μ and σ^2. Let $Y = \{Y_1, Y_2, ..., Y_n\}$ the set of discrete parents of a continuous variable X, then $|Y| = |Y_1| \times |Y_2| \times \cdots \times |Y_n|$ as the number of possible combinations of Y. The CPT of X will be a vector with dimension $|Y|$, $\vec{\mu} = \langle \mu_1, \mu_2, \cdots, \mu_{|Y|} \rangle$, and, similarly for the variance set, the vector $\vec{\sigma^2} = \langle \sigma_1^2, \sigma_2^2, \cdots, \sigma_{|Y|}^2 \rangle$. Figure 3 shows an example of a continuous variable X, with a set of discrete boolean variables Y_i, parents of X.

(2) **Continuous variables with continuous parents**
Consider a continuous variable X, with m continuous parents $Z_i \in \{Z_1, Z_2, \cdots, Z_m\}$. For each parent variable Z_i, X has a weight w_i, such that the mean of X is calculated by:

$$\mu_x = \epsilon + \sum_{i=1}^{m} w_i z_i$$

where z_i is the value of the parent variable Z_i, and ϵ is the intercept term of X in the linear regression formula. The variance of X, by computational questions, is given by only a parameter σ^2 and is not conditional to the parents.

(3) **Continuous variables with continuous and discrete parents**
When a continuous variable has continuous and discrete parents, a Linear Gaussian distribution is used for each combination of discrete parent values. This model is known as Linear-Gaussian Conditional Network. Given that X is a continuous variable with a set of discrete parents Y and a set of continuous parents Z. X has a mean, variance and a set of regression weights, $w_{|Y|,|Z|}$. The average variable is given by the vector $\vec{\mu} = \langle \mu_1, \mu_2, \cdots, \mu_{|Y|} \rangle$, and the variance by the vector $\vec{\sigma^2} = \langle \sigma_1^2, \sigma_2^2, \cdots, \sigma_{|Y|}^2 \rangle$ and the regression matrix $|Y| \times |Z|$:

$$\begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,|Z|} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,|Z|} \\ w_{...,1} & w_{...,2} & \cdots & w_{...,|Z|} \\ w_{|Y|,1} & w_{|Y|,2} & \cdots & w_{|Y|,|Z|} \end{bmatrix}$$

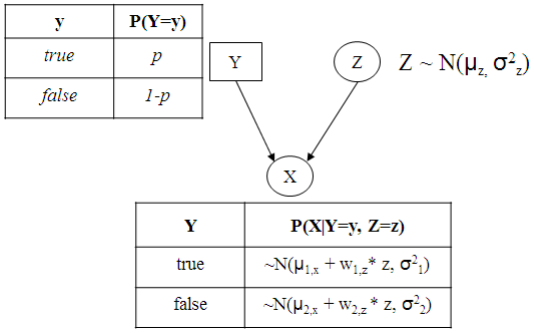

Fig. 4. An example of a BN with one continuous and one discrete variable

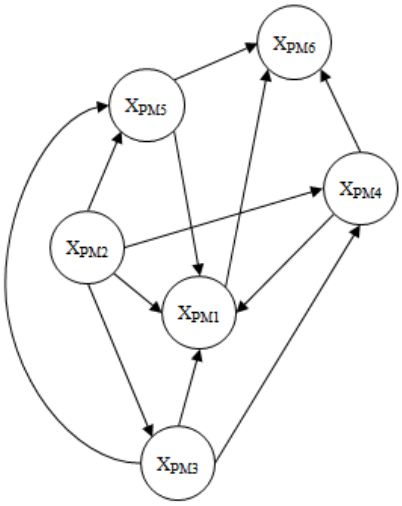

Fig. 5. Spatial Component of the Bayesian Network, constructed using Hill Climbing algorithm using BGe

Figure 4, adapted from [Dereszynski and Dietterich 2011], shows an example of a Linear-Gaussian Conditional Network, where a continuous variable X has a discrete boolean parent Y and a continuous parent Z.

3. THE PROPOSED METHOD

Inspired on [Dereszynski and Dietterich 2011], the proposed method in this work firstly divide the Bayesian Network to be constructed into two components: (i) the spatial component, modeling the relation between each pair of FMP at an instant time; and (ii) the temporal component, representing the transition from one time instant to another.

To build the spatial component, Bayesian Network structure algorithm learning was used to learn the set of spatial relationships among the measurement points. This approach allows the model to adapt to each plant where the model is used, without the need of prior knowledge of these relationships. The Hill Climbing algorithm was used to find the best spatial structure using BGe (Bayesian Metric for Gaussian) scoring metric. To use BGe, it was assumed that the data set corresponds to a multivariate Gaussian distribution. The obtained network structure is shown in Figure 5. Each node in the network, labeled X_i, represents the measurement point $i \in \{PM1, PM2, PM3, PM4, PM5, PM6\}$.

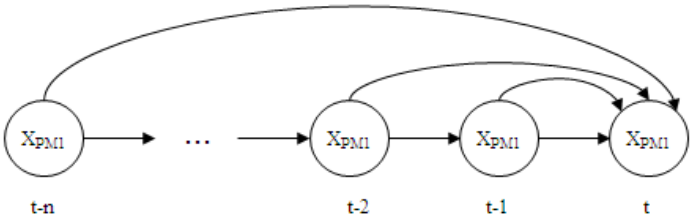

Fig. 6. Example of a Temporal Component added to FMP PM1 with n lag variables

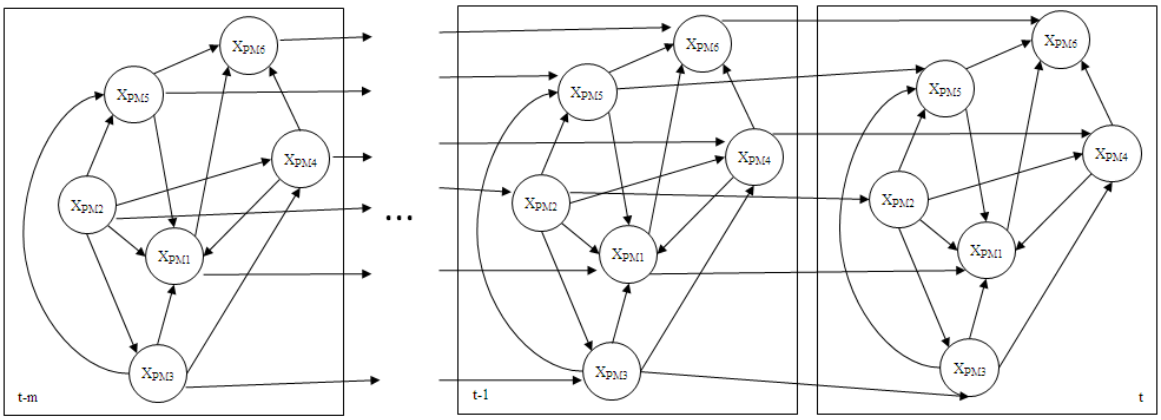

Fig. 7. Time transitions from time instant m

To build the temporal component, each variable that represents the reading of the current time instant of a sensor is related to a lag variable. These nodes are also called lag variables. Figure 6 shows the relationship between the measurement point PM1 and the lag variables up to time n. The addition of these variables is responsible for transforming a Bayesian Network from Static to Dynamic.

The next step of our method is estimating the value of the parameter, *i.e.*, μ_i (mean of the variable i), σ_i^2 (variance of the variable i) and w_i (weight of the variable i) using the Maximum Likelihood Estimates (MLE) approach. This estimation is a multiple linear regression problem [Russell et al. 2003].

After constructing the model, containing spatial and temporal components, and estimating the parameters, it is possible to predict the value measured by a measurement point, based on the parent nodes in the Bayesian Network. The prediction process is iterative: at each instant of time, the value read in a time instant is used to predict the value of the subsequent time. Figure 7 shows a model, with spatial component and temporal component with 1 lag variable, after m time instants.

To use this model to diagnose anomaly in FMP, an inference component was added to the model. It is composed by a node representing the current observation of the FMP O_i and a variable representing the state of this point S_i, where i represents the measurement point. The nodes S_i are discrete, and can store values working or fault. The parameters for these nodes were assigned manually. For S_i the ratio between the number of days with failures and the data period lifted was used (in the specific experimental analysis of this work, the rate was 11 in 180 days). It was assigned to the broken value 0.1 of failure probability to make the model a little more likely to alarm a fault. The working state has been assigned the complement of this value. For O, the following main idea was used: when a measuring point is working, the predicted value x_i must be very close to the observed value, and when the measuring point is broken a larger variation. The parameterization is shown below, and Figure 8 shows the structure of the Hybrid Bayesian Network.

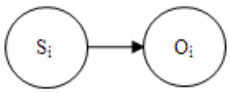

Fig. 8. Inference Component

$$P(S_i = working) = 0.90$$
$$P(S_i = fault) = 0.10$$
$$P(O_i|S_i = working, X_i = x_i) \sim N(x_i, 0.10)$$
$$P(O_i|S_i = fault, X_i = x_i) \sim N(0.0001x_i, 10000)$$

So, the inference process is done in two steps. In the first step, the prediction of the value of a FMP X_i is made based on the nodes, related to it, each of them with spatial and temporal components. Then, the predicted value is set in O_i, based on the observed value of this measurement point.

The method was implemented using R language [R Core Team 2016] and the bnlearn package [Scutari 2010], which provides some algorithms for learning structure, parameter estimation and inference in Bayesian Networks[1] .

4. EXPERIMENTS AND RESULTS

The production plant used in this work was presented in Figure 1. As mentioned before, there are six FMPs: PM1, PM2, PM3, PM4, PM5 and PM6. The Corrected Volume data were collected for $k = \{10, 20, 60\}$ minute. The dataset (Corrected Volume values from the six FMPs) was collected from November of 2015 to April to 2016. It was divided into training set, containing data from November to December, and test set, containing data from January to April. Spurious data and communication problems were removed, and the Corrected Volume values were normalized to the range [0; 1].

Another decision was regarded to how many lag variables should be tested. For performance issues, a test was performed with up to two lag variables ($m = 2$), that is, tests were made using three variations of the model, called (i) SC, composed only by the Spatial Component, without considering the temporal component; (ii) ST1C, composed by spatial and temporal components, with 1 lag variable; and (iii) ST2C, composed by spatial and temporal components, with 2 lag variables.

The first carried out test aims at verifying the impact of k and variation of the model in prediction performance of the value, reported by a FMP. For this end, nine BN models were constructed, one for each combination of model variation (SC, ST1C and ST2C) and k value 3. This test consists in comparing the value predicted by the model with the observed value using the mean squared error metric (MSE). Table II shows the results of this test. It is important to observe that MAPE (Mean Absolute Percentage Error) is commonly used in regression problems. However, the data used in this work present some zero values, which turns impossible to use MAPE.

Statistical Tests: The Friedman test was used to verify if the combination of time interval k and variation of the model presents significant statistical difference [Demsar 2006]. For this end, the first random variable v_1 is relative to the results obtained for $K = 10$ and Model SC; V_2 is relative to the results obtained for $k = 10$ and Model ST1C; V_3 is relative to the results obtained for $K = 10$ and Model ST2C; V_4 is relative to the results obtained for $K = 20$ and Model SC; V_5 is relative to the results obtained for $K = 20$ and Model ST1C; V_6, for $K = 20$ and Model ST2C; V_7, for $K = 30$ and Model CE; V_8, for $K = 30$ and Model ST1C; And finally, v_9, to $K = 60$ and Model ST2C. According to the test, the null hypothesis was rejected. So, there is a significant difference with 95% of confidence

[1]The source code of the Bayesian Network used in this work is available at `http://github.com/hadriellima/Anomaly-Detection-Bayesian-Network`.

Table II. Comparison between Models Using MSE (Mean Squared Error)

Measurement Point	k = 10			k = 20			k = 60		
	Model SC	Model ST1C	Model ST2C	Model SC	Model ST1C	Model ST2C	Model SC	Model ST1C	Model ST2C
PM1	0,003	0,001	0,001	0,004	0,002	0,002	0,010	0,003	0,003
PM2	0,055	0,001	0,001	0,061	0,002	0,002	0,061	0,003	0,003
PM3	0,039	0,001	0,001	0,027	0,001	0,001	0,034	0,005	0,006
PM4	0,142	0,012	0,010	0,041	0,007	0,006	0,032	0,006	0,006
PM5	0,081	0,010	0,042	0,060	0,011	0,055	0,067	0,030	0,028
PM6	0,394	0,050	0,010	0,498	0,072	0,010	0,222	0,043	0,038
Mean	**0,119**	**0,012**	**0,011**	**0,115**	**0,016**	**0,013**	**0,071**	**0,015**	**0,014**

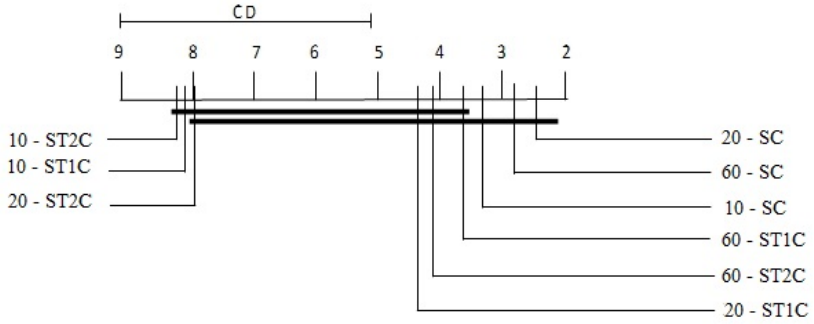

Fig. 9. Graphical representation of the pos-test of each combination of experiment scenario as a random variable

degree in the results obtained. Hence, the Nemenyi post-test was performed. In Figure 9 the test result is displayed, where the statistical significance between two groups of results can be verified.

The Friedman test was also performed to verify if there is statistical difference in results when k value changes. The first random variable v_1 is equivalent to $k = 10$; V_2 is equivalent to $k = 20$; And v_3 is equivalent to $k = 60$. The values observed were: v_{1_1} — MSE for point PM1 and Model SC; v_{1_2} — MSE for point PM1 and Model ST1C; v_{1_3} — MSE for point PM1 and Model ST2C; v_{1_4} — MSE for point PM2 and Model CE; And so on, up to $v_{1_{18}}$ — MSE for PM6 and Model CET2. The test did not reject the null hypothesis, meaning that the results are statistically similar for different values of k. So, $k = 10$ is preferred, as lower the value of k, lower the granularity of prediction time.

Finally, the Friedman test was performed to verify if there is statistical difference in the results when there is a variation of the model (SC, ST1C or ST2C). The first random variable v_1 is equivalent to the SC model; V_2 is equivalent to model ST1C; And v_3 is equivalent to the ST2C model. The observed values were: v_{1_1} — MSE for point PM1 and $k = 10$; v_{1_2} — MSE for point PM1 and $k = 20$; v_{1_3} — MSE for point PM1 and $k = 60$; v_{1_4} — MSE for point PM2 and $k = 10$; and so on, up to $v_{1_{18}}$ — MSE to point PM 6 and $k = 60$. The test rejected the null hypothesis, that is, there is a significant difference in the results obtained. Hence, Nemenyi post-test was performed. Figure 10 shows the test result, and it is verified that the statistical significance is between two groups of results, meaning that temporal component affects the model behavior.

Analysis of the best model: Due to the results obtained, it can be observed that the best model obtained was considering model ST2C and $k = 10$. Thus, the constructed model was chosen

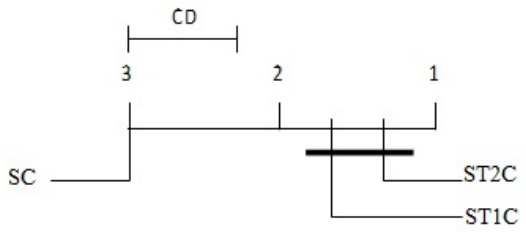

Fig. 10. Graphical representation of the post test of each Bayesian network model as a random variable

Table III. Result of anomaly detection reported by Spatiotemporal Model

		Model Prediction		Accuracy	Precision	Recall
		Working	Fault			
PM1	**Working**	120	0	1	1	1
	Fault	0	0			
PM2	**Working**	120	0	1	1	1
	Fault	0	0			
PM3	**Working**	120	0	1	1	1
	Fault	0	0			
PM4	**Working**	115	3	0.975	1	1
	Fault	0	2			
PM5	**Working**	109	6	0.942	0.991	1
	Fault	1	4			
PM6	**Working**	82	32	0.733	1	1
	Fault	0	6			

to evaluate the anomaly diagnosis. This test consists of verifying the effectiveness of the model in detecting measurement point failure (or anomaly). All samples from the test set were tested one at a time. When the probability of failure of an example was greater than 50%, the test of the following example used only the spatial component, since a value with potential to be incorrect as a delay variable would impact the prediction of the next value.

A summary of the modeling behavior is presented in Table III. As reported failures are associated only to the day they occurred, the model adherence analysis was built in days. Lines 3 and 4 of this table present a confusion matrix for the measuring point PM1; lines 5 and 6, a confusion matrix for the measuring point PM2; lines 7 and 8, a confusion matrix for measuring point PM3; lines 9 and 10, a confusion matrix for measuring point PM4; lines 11 and 12, a confusion matrix for the measuring point PM5; and finally lines 13 and 14, a confusion matrix for the measuring point PM6. Only one fault, occurred in PM5, was not detected by the model. This failure occurred along with the failure in PM6 that was diagnosed.

The occurrence of false positives in FMPs PM4, PM5 and especially in PM6, with 32 false positives, were shown in Table III. Further, the occurrence of false positives increased when the time of the read value is further from the training data, i.e., they occurred in March and April. Figure 11 shows a graph illustrating this fact. Measurement Points PM4 and PM5 had no more than 2 occurrences of false positives per month throughout the training period, while PM6 Measurement Point, which occupied the same range in January and February, jumped to 12 in March And 17 in April. The closure of a well, diverse of production, and increase of consumption for operational reasons are possible occurrence facts that can alter the characteristic of the fluid flow (gas) through the plant.

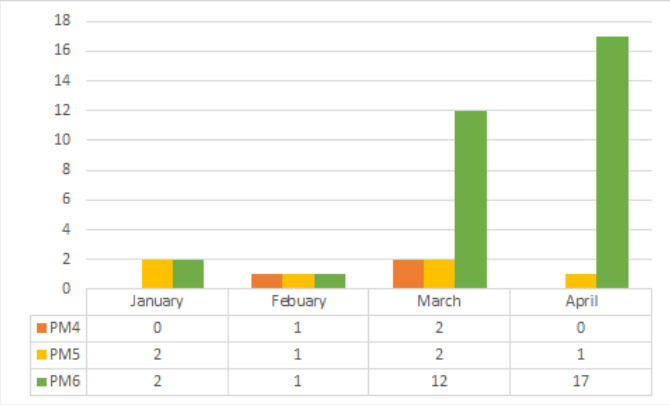

Fig. 11. Graph showing the evolution of the occurrence of false positives grouped by month

Analyzing PM6 measurement data in March and April, it is possible to note differences from training data, probably caused by a change in operational condition.

5. CONCLUSIONS

This article proposes a method based on Dynamic Bayesian Networks to construct a spatiotemporal model for Anomaly Detection in a Natural Gas production plant. The proposed method has the advantage of being adaptable to any process plant, and does not require the knowledge of any specialist in the operation of the plant to be implanted. The model learns the relationships between measurement points from data, making it easily scalable.

The method was implemented using R language. Two tests were conducted in experimental analysis. The first one aimed to verify the impact on forecast performance of the value, reported by a FMP, varying k (time interval considered in the data collection) and model variation. The k values were 10, 20 or 60 minutes, and the considered models were Bayesian Networks with (i) only spatial component (SC); (ii) spatial and temporal component with a delay variable (ST1C); and (iii) spatial component With two delay variables (ST2C). This test showed that the lowest value of k, $k = 10$, and the ST2C model obtained the best performance. This was the alternative used in the second phase of the tests, aimed to verify the modeling effectiveness in detecting anomalies. Modeling the problem to use Dynamic Bayesian Networks led to identify most of the anomalies present in the used dataset used, failing to identify only one. On the other hand, it could be observed a high number of false positives at one of the measurement points. The highest number of false positives occurred in the last months of the test period, indicating that the model has to be adapted over time. Thus, using a sliding window in the dataset appears to be a good alternative to continue the work and should be explored in the future.

It is worth to notice that the method proposed in this work for anomaly detection in FMPs firstly predicts the expected value, and after compares this value to the real value. This composes a different way of modeling a typical classification problem, which is considered a contribution of this work. Another contribution is turning easier to compute the values predicted by the Bayes Network than the one presented by [Dereszynski and Dietterich 2011], which is a desirable feature in problems where many components are connected, and even that good results were achieved.

A weakness of this work is the use of a Gaussian-Linear Conditional Network to handle continuous variables. The decision was based on the similarity of the problem studied in this work with that studied in [Dereszynski and Dietterich 2011]. However, other approaches to handle continuous variables can be explored. Also, other algorithms for learning the spatial component can also be explored by

evaluating alternative search algorithms to Hill Climbing algorithm, and other scoring metrics. Also, we plan to test the method using sliding window in other natural gas production plants, as well as petroleum production plants. It is important to observe that the method presented in this article is able to be applied in other types of sensor networks, however the nature of data must be studied. So, other future work is related to constructing a framework for using Bayes Net to anomaly detection in any sensor nets, facilitating the use of these algorithms, and considering what data characteristics are more suitable to this type of solution.

REFERENCES

Akouemo, H. N. and Povinelli, R. J. Probabilistic anomaly detection in natural gas time series data. *International Journal of Forecasting* 32 (3): 948–956, 2016.

Alves, J., Ferreira, J., Lobo, J., and Dias, J. Brief survey on computational solutions for bayesian inference. *Workshop on Unconventional computing for Bayesian inference (UCBI), International Conference on Intelligent Robots and Systems (IROS)*, 2015.

ANP and INMETRO. Joint ordinance ANP and INMETRO, nº. 1, 2000. provides the technical regulation of oil and natural gas measurement (*in Portuguese*). Tech. Rep. 1, Official Journal of the Federative Republic of Brazil (*in Portuguese*), Brasília, DF, Brazil, 2000.

ANP and INMETRO. Joint ordinance ANP and INMETRO, nº. 1, 2010. provides the technical regulation of oil and natural gas measurement (*in Portuguese*). Tech. Rep. 1, Official Journal of the Federative Republic of Brazil (*in Portuguese*), Brasília, DF, Brazil, 2010.

ANP and INMETRO. Joint ordinance ANP and INMETRO, nº. 1, 2013. provides the technical regulation of oil and natural gas measurement (*in Portuguese*). Tech. Rep. 1, Official Journal of the Federative Republic of Brazil (*in Portuguese*), Brasília, DF, Brazil, 2013.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys*, 2009.

Das, M., Parthasarathy, S., and Agrawal, G. *Spatio-temporal Anomaly Detection.* M.S. thesis, The Ohio State University, 2009.

Demsar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning Research* vol. 7, 2006.

Dereszynski, E. W. and Dietterich, T. G. Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns. *ACM Transactions on Sensor Networks (TOSN)* 8 (1): 3, 2011.

Driver, E. and Morrell, D. Implementation of continuous bayesian networks using sums of weighted gaussians. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc, Montreal, Quebec, Canada, pp. 134–140, 1995.

Guo, H. and Hsu, W. A survey of algorithms for real-time bayesian network inference. In *AAAI/KDD/UAI02 Joint Workshop on Real-Time Decision Support and Diagnosis Systems*. Edmonton, Canada, 2002.

Hauke, J. and Kossowski, T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae* 30 (2): 87–93, 2011.

Kut, A. and Birant, D. Spatio-temporal outlier detection in large databases. *CIT. Journal of computing and information technology* 14 (4): 291–297, 2006.

Langseth, H., Nielsen, T. D., Rumi, R., and Salmerón, A. Mixtures of truncated basis functions. *International Journal of Approximate Reasoning* 53 (2): 212–227, 2012.

Lazari, R. F., Souza, F. S. d., Dias, B. G., and Hartmann, V. N. Vision about the new version of the ordinance ANP/Inmetro nº 01/2000 (*in Portuguese*), 2009.

Ma, J. and Perkins, S. Online novelty detection on temporal sequences. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Washington, DC, pp. 613–618, 2003a.

Ma, J. and Perkins, S. Time-series novelty detection using one-class support vector machines. In *Proceedings of the International Joint Conference on Neural Networks*. Vol. 3. IEEE, Portland, Oregon, pp. 1741–1745, 2003b.

Maldonado, A., Aguilera, P., and Salmerón, A. Continuous bayesian networks for probabilistic environmental risk mapping. *Stochastic Environmental Research and Risk Assessment*, 2015.

Murphy, K. P. *Dynamic bayesian networks: representation, inference and learning.* Ph.D. thesis, University of California, Berkeley, 2002.

Paschalidis, I. C. and Smaragdakis, G. Spatio-temporal network anomaly detection by assessing deviations of empirical measures. *IEEE/ACM Transactions on Networking (TON)* 17 (3): 685–697, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., and Edwards, D. D. *Artificial intelligence: a modern approach*. Prentice hall Upper Saddle River, 2003.

Scutari, M. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software* 35 (3): 1–22, 2010.

Shenoy, P. P. and West, J. C. Inference in hybrid bayesian networks using mixtures of polynomials. *International Journal of Approximate Reasoning* 52 (5): 641–657, 2011.

Tax, D. and Duin, R. Uniform object generation for optimizing one-class classifiers. *J. Machine Learning Research* vol. 2, pp. 155–173, 2001.

Trifonova, N., Kenny, A., Maxwell, D., Duplisea, D., Fernandes, J., and Tucker, A. Spatio-temporal bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. *Ecological Informatics* vol. 30, pp. 142–158, 2015.

Wang, X. R., Lizier, J. T., Obst, O., Prokopenko, M., and Wang, P. Spatiotemporal anomaly detection in gas monitoring sensor networks. In *Wireless Sensor Networks*. Springer, pp. 90–105, 2008.