# Preventive Leak Detection for High Pressure Gas Transmission Networks

**Rui Zhang**
zhangrui@us.ibm.com
IBM, T.J. Watson Research
1101 Kitchawan Road
Yorktown Heights, NY, 10598

**Jefferson Huang**
jh2543@cornell.edu
School of ORIE
Cornell University
Ithaca, NY 14853

**Tarun Kumar**
ktarun@us.ibm.com
IBM, T.J. Watson Research
1101 Kitchawan Road
Yorktown Heights, NY, 10598

**Abstract**

Recent developments in SCADA (Supervisory Control and Data Acquisition) systems for physical infrastructure, such as high pressure gas pipeline systems and electric grids, have generated enormous amounts of time series data. This data brings great opportunities for advanced knowledge discovery and data mining methods to identify system failures faster and earlier than operation experts. This paper presents our effort in collaboration with a utility company to solve a grand challenge; namely, to use advanced data mining methods to detect leaks on a high pressure gas transmission system. Leak detection models with unsupervised learning tasks were developed analyzing billions of data records to identify leaks of different sizes and impacts, with very low false positive rates. In particular, our solution was able to identify small leaks leading to rupture events. The model also identified small leaks not identifiable with current detection systems. Such high-fidelity early identification enables operation personnel to take preventive measures against possible catastrophic events. We then formulate several generic detection methods with models derived from time series anomaly detection methods. We show that our leak detection models are superior to the SCADA alarm system, a mass balance model and other generic time series anomaly detection models in terms of both detection accuracy and computation time.

There are 91,000 miles of natural gas transmission and distribution pipelines across the United States. Natural gas leaks pose grave threats to public safety, and can have significant environmental and economic implications. In 2011, according to (Markey 2013) gas distribution companies reported releasing 69 billion cubic feet of natural gas into the atmosphere, which is almost equal to annual gas consumption of the state of Maine and is equal to the annual carbon dioxide emissions of about six million automobiles. This gas is primarily comprised of methane, which is a greenhouse gas that is at least 21 times more potent than carbon dioxide. Americans also remain at risk from gas explosions and other safety hazards caused by leaky natural gas pipelines. From 2002 to 2012, almost 800 significant incidents occurred on gas distribution pipelines, including several hundred explosions, which resulted in 116 deaths, injured 465, and caused more than $800 million in property damage. In view of this, our objective is to use data to detect leaks in a complex high pressure natural gas transmission system in the Rocky Mountain area near Denver, Colorado. This was made possible in part by recent developments in SCADA (Supervisory Control and Data Acquisition) systems for physical infrastructure, such as high pressure gas pipeline networks and electrical grids. Such SCADA systems generate and record enormous amounts of data continuously. This data contains valuable information on the operating conditions of such physical systems, and can be used for anomaly detection.



There has been an extensive amount of effort in the realm of leak detection research. Physical models, finite element methods, and statistical models have been developed and tested on synthetic data and small-scale data. More details on existing methods can be found in the section of this paper on Related Work. We note that none of the existing methods have been tested on real world operational data at the scale we consider. In addition, there are many challenges that arise from uncertainties in real measurements and operational data: 1. The measurements may be noisy and biased, particularly those from the flow rate sensor. 2. There are constant system operation and maintenance activities that are not recorded, which should be accounted for in learning to detect leaks. 3. The gas in the transmission system is compressible, which means the effects of operations, e.g. changes in compressor pressure, take time to propagate in the system; this needs to be captured by the leak detection model. 4. The data is large-scale; there are thousands of sensors in the system, and finding the most relevant ones is challenging in terms of feature selection, automatic or manual.

The key contributions of this paper are as follows:

- We propose an innovative and effective anomaly detection solution for high pressure compressible natural gas transmission networks, which identifies small leaks preceding rupture events in a real system.
- We develop an effective approach that identifies small leaks occurring over extended periods of time in a real high pressure natural gas transmission system.
- We compare our method with a mass balance model and several temporal anomaly detection methods on time series data in terms of detection accuracy and computation time.

The rest of the paper is organized as follows. The Related Work section reviews current approaches to gas leak detec-

tion and anomaly detection research on time series data. The Model Setup section presents the formulation of the big leak detection and small leak detection methods. We introduce the formulation of methods for generic time series anomaly detection in the Other Methods section, and provide experimental results on synthetic data suggesting that they may be useful for our leak detection task. The real application data set for the high pressure natural gas transmission system is introduced in the Application section, where the results on leak detection using our methods and the generic anomaly detection methods are summarized. We conclude our paper with our findings and insights in the Conclusion.

## Related Work

Two types of leaks of particular interest are ruptures and small seeps. Ruptures are the most dangerous, with paramount environmental impact and grave threats to public safety that sometimes result in deaths. A number of methods have been developed to detect such events, including acoustic monitoring, flow monitoring, and model-based methods. It is fair to say that most methods will detect a rupture event, as a rupture will result in significant deviations of system characteristics, i.e., acoustic, flow, or pressure, compared to normal operating conditions. However, the false alarm rate is often high; in a real system, there were 15,058 alarms, while the true positive rate was only 0.0199%.

In addition to non real-time inspection methods (Stearns et al. 2005; Wang et al. 2001), two approaches to detecting small seeps that do not necessarily lead to rupture events are active real-time monitoring (Sivathanu 2003; Stearns et al. 2005) and model-based real-time monitoring.

Active real-time monitoring approaches include acoustic methods, optical methods, and flow/pressure monitoring. However, a large number of acoustic/optical sensors is required to monitor a large pipeline system. Additional disadvantages include the high cost of implementation and the high incidence of false alarms

Model-based methods constitute a rich research area. (Dos Santos et al. 2011) modeled the pipeline as a Linear Parameter Varying (LPV) System driven by the source node massflow with the gas inventory variation in the pipe as the scheduling parameter. It was found that the method is able to detect leaks. (Wan et al. 2011) proposed a hierarchical leak detection method with signal processing and support vector machine (SVM) based classification. The method was validated in an experimental pipeline of 25m in length and 160mm in diameter. (Isa and Rajkumar 2009) employed SVM, which was trained on a number of samples representing the presence of leaks of various sizes and locations in a lab scale experimental rig. Model-based methods mainly employ signal processing methods or Kalman filtering techniques with or without a combination of classification methods; see e.g., (Ma, Yu, and Huo 2010; Martins and Seleghim 2010; bin Md Akib, Bin Saad, and Asirvadam 2011; Kim and Lee 2009; Lay-Ekuakille, Vendramin, and Trotta 2009; Bai, Yue, and Li 2005).

All of the previous studies were validated with lab-scale setups and simulated leak events. And there is no previous work that employs a time series analysis point of view, which is the foundation of our method. In particular, we consider a well-studied framework for time series anomaly detection that involves fitting a model to the data, and then constructing an anomaly score for new data (Chandola, Banerjee, and Kumar 2009, Section 7.1.2). For example, (Qiu et al. 2012) proposed such an approach based on learning Granger causality graphs via $L_1$-penalized regression. In (Takeishi and Yairi 2014), the anomaly score is computed by examining local patterns in the series using ideas from sparse coding and natural language processing. (Laptev, Amizadeh, and Flint 2015) estimate the distribution of residuals relative to the model using kernel density estimation, quantify changes in this distribution using Kullback-Leibler divergence, and evaluate models based on moving averages, exponential smoothing, and Kalman filtering. Recently, (Jones et al. 2016) proposed an approach based on efficiently representing information about subsequences of the time series.

## Model Setup

In this section, we describe in detail the models we developed for two tasks. The first task is to detect big leaks, which usually lead to major rupture events. The second task is to detect small seeps, which are more difficult to detect; such events do not typically lead to a major rupture, but can lead to the release of large amounts of gas if left undetected.

### Big Leak Detection Model

The principal idea is to train a model using historical data, which will capture the correlations between physical measurements from connected stations under normal operating conditions. The model is used to predict the value of a chosen critical variable using other measurements at both the same station and connected stations. Then, the deviation between the predicted values and the actual measurements can be used to compute the likelihood of any anomalies, i.e., leaks. The critical variable we chose for the task of leak detection is the compressor discharge pressure, given that leaks may lead to pressure changes. Other reference measurements used were flow and compressor operation condition measurements, namely the compressor fuel consumption and compressor engine RPM (Revolutions Per Minute). In addition, the temporal delays between measurements are computed; as the fluid in the pipeline is compressible, there are significant temporal delays in the system. For example, if the compressor operates to increase the pressure, the pressure at a downstream station may not see an instantaneous pressure increase.

We now provide a detailed description of the method:

For a high pressure gas pipeline system there are $n$ stations; the set of all stations is denoted by $\mathcal{S} = \{S_1, \cdots, S_n\}$. For station $S_i$, $i = 1, \cdots, n$, we take four types of measurements, namely the pressure $S_{i1}$, flow $S_{i2}$, compressor operation measurements $S_{i3}$ (i.e., compressor RPM or compressor fuel consumption), and the air temperature $T_i$. We denote the set of stations downstream from $S_i$ by $\mathcal{A}_i$, as is shown in Figure 1.

As explained in the above section, there are delays (lag) in response between upstream stations and downstream sta-

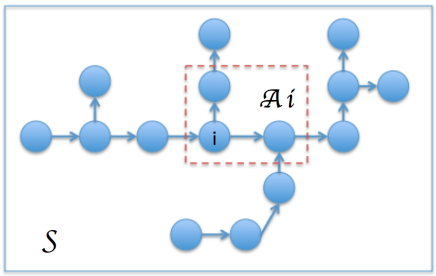


Figure 1: An illustrative gas pipeline system. The dashed box encloses the stations that are downstream from station $S_i$.

tions. We first find the lags between connected stations and critical variable

that maximizes the cross correlation function (CCF) between the time series corresponding to the critical variable $S_{i1}(t)$, and the time series for the measurements $S_{jk}(t)$ $(k = 1, 2, 3,\ j \in \mathcal{A}_i)$ as follows:

$$L_{ijk} = \underset{l}{\text{argmax}}(\rho_{S_{i1}S_{jk}}(l)),$$

where:

$$\rho_{S_{i1}S_{jk}}(l) = \frac{E[(S_{i1}(t) - \mu_{S_{i1}})(S_{jk}(t-l) - \mu_{S_{jk}})]}{\sigma_{S_{i1}(t)}\sigma_{S_{jk}(t-l)}}$$

The CCF is computed by using the fact that:

$$\rho_{S_{i1}S_{jk}}(l) = [IFFT(FFT(S_{i1}) \times FFT^*(S_{jk}))]_l$$

where FFT denotes the Fast Fourier Transform, IFFT denotes the Inverse Fast Fourier Transform, and the asterisk denotes the complex conjugate. For $S_{jk}$ $(k = 1, 2, 3,\ j \in \mathcal{A}_i)$, let $S_{jk}^{L_{ijk}} = [S_{jk}(t - L_{ijk}), \ldots, S_{jk}(t_0 - L_{ijk})]$. The formulation of the leak detection model using a linear model is as follows:

$$S_{i1} = [\ S_{i2} \quad S_{i3} \quad S_{jk}^{L_{ijk}} \quad T_i\ ] \cdot \beta \qquad (1)$$

where:

$k = 1, 2, 3, \quad j \in \mathcal{A}_i,$

$\beta$ denotes the vector of regression coefficients.

The model is trained with a one-month moving window of operational data. Then the model is used to infer the critical variable of pressure measurement as $\widehat{S_{i1}^{t+h}}$, where $h$ is the time window for the new data. Typically, $h$ can be the next 10 minutes, 30 minutes, or 1 hour, depending on the frequency for leak detection tasks. The deviation between the predicted and measured pressure is calculated as: $r_i^{t+h} = S_{i1}^{t+h} - \widehat{S_{i1}^{t+h}}$. A multiple of the standard deviation $\sigma_r$ of the deviations is used as the event detection threshold. The choice of the threshold is chosen based on the event history of the location. In this study, the threshold was chosen based on two leak events. The threshold values can vary from station to station, and were generally at least $6 \times \sigma_r$.

### Small Leak Detection Model

The objective of the small leak detection model is to detect leaks that do not cause significant changes in pressure, and hence may not be discovered by SCADA alarms or even the big leak detection model. We have developed a small leak detection model based on the model described in Section Big Leak Detection Model.

The hypothesis for the small leak detection model is that a small leak will result in cumulative small changes in the critical variable of pressure. Therefore, we employed the CUSUM (CUmulative SUM) model to detect changes in the calculated deviation of the critical pressure. CUSUM is commonly used in quality control to detect deviations of variables from benchmark values (Basseville, Nikiforov, and others 1993). In our task, the variable is the critical variable of pressure, and the benchmark values of the variable are the predicted values $\widehat{S_{i1}^{t+h}}$ computed from (1). The differences between the measurement and the benchmark values are cumulatively summed up. If there is no anomaly in the system, the measurements do not deviate significantly from the benchmark, so the measurements greater than the benchmark and those less than the benchmark average each other out, and the CUSUM value should remain around the benchmark level. If a continuous anomaly occurs, the measurement continuously deviates from the benchmark in the same direction, resulting in a larger CUSUM value. The CUSUM is defined by:

$$\text{Upper Control Limit:} \quad C_t^+ = \max(0, S_{i1}^t - \widehat{S_{i1}^t} + C_{t-1}^+)$$

$$\text{Lower Control Limit:} \quad C_t^- = \max(0, \widehat{S_{i1}^t} - S_{i1}^t + C_{t-1}^-)$$

A small leak event is defined as time period when the CUSUM values $C_t^+$ or $C_t^-$ exceed threshold values for more than five days. The threshold values can vary from station to station, but in general it should be larger than $6 \times \log(|C|)$, where $C$ is the control limit.

## Other Methods

The leak detection problem can be generalized into a time series anomaly detection problem. Here we formulate our problem in this framework, without using expert-picked features. Temporal dependencies between time series can often be captured by a sparse Granger Causality model; see e.g. (Arnold, Liu, and Abe 2007). However, several challenges exist for the system we consider:

1. The number of time series in the system is very large, and there are many non-informative and highly correlated measurements, so being able to find the most relevant time series among many is important.

2. The volume of the incoming data is huge. In most SCADA systems, the data is sampled every 30 seconds, which results in 2,880 data points per day, or 1.05 million per year.

3. The nature of the physical system dictates non-linear and lagged dependencies between the measurements. These characteristics pose great challenges for effective anomaly detection model development and on-line real-time detection.

In this section, the model training follows the same principle as in the big leak detection model. We define the pressure measurement as the critical variable, and all time series measurement from the same station and all the downstream stations are used as features. We consider several non-linear modeling approaches. The effectiveness of these methods for detecting dependencies between time series is evaluated in Section Experiment,and in Section Results for Other Methods.we compare them to the method proposed in Section Big Leak Detection Model.

## Formulation

We consider the following model formulation. For station $S_i$ $(i = 1, \ldots, n)$, let $K_i$ denote the number of time series available at station $S_i$, and let $K_j$ denote the total number of time series from station $S_j$ $(j \in \mathcal{A}_i)$. The value of the critical variable $S_{i1}(t)$ for station $S_i$ at time $t$ is modeled as a possibly non-linear function $\mathcal{F}$ of the values at time $t$ of the time series $S_{j1}(t), \ldots, S_{j,K_j}(t)$ available at station $S_j$ $(j \in \mathcal{A}_i)$ and the lagged values

$$S_j^{t,L} := [S_{j,1}(t-L), \ldots, S_{j,1}(t-1), \ldots, S_{j,K_j}(t-L), \ldots, S_{j,K_j}(t-1)]$$

of the time series at the station $S_j, j \in \mathcal{A}_1$ connected to $S_i$, for a specified maximum lag $L$. Letting

$$X_i(t) := (S_{i,2}(t), \ldots, S_{i,K_i}(t), S_j^{t,L}) \quad (j \in \mathcal{A}_i)$$

it follows that we are modeling $S_{i1}(t)$ with $\mathcal{F}(X_i(t))$.

In this study, we used four methods, namely: 1. linear regression with $L_1$ penalty, i.e., LASSO, 2. Elastic Net with both $L_1$ and $L_2$ penalties; 3. Gradient Boosting Machines; 4. Gradient Boosting Machines with Huber loss; 5. Random Forests.

## Anomaly Detection Model

The models produced by each of the methods described in this section can also be used for anomaly detection in the way described at the end of Section Big Leak Detection Model.First, for each station $i$ a model $\mathcal{F}(X_i(t))$ is fit to the critical variable $S_{i1}(t)$ using one month of data, and the residuals $S_{i1}(t) - \mathcal{F}(X_i(t))$ are computed. A normal distribution is then fit to the residuals, and the threshold for event detection is taken to be a multiple of the standard deviation of this fitted distribution. The model can then be re-trained as new data becomes available, according to a specified time window $h$.

## Experiments

We conducted numerical experiments on synthetic data with known dependencies to compare the effectiveness of the methods introduced in Section Other Methods in detecting sparse non-linear temporal dependencies.

**Synthetic Dataset** To assess the extent to which our approach identifies time series that are actually relevant, we applied each method to randomly generated synthetic data. This data was generated as follows. Each "critical variable" $y(t)$ is taken to be a function of $J$ time series corresponding to normalized stock data; $J$ was chosen randomly, and the $J$ stocks were chosen randomly out of a set of 10 stocks. The functional form of the dependence between $y(t)$ and these $J$ time series is then taken to be a sum of products of transformations of lagged variables of the $J$ time series, plus Gaussian noise with mean zero and standard deviation 0.25. The number of terms in the sum and the number of factors in each product were each chosen randomly from the set $\{1, 2, 3\}$, and each term in the sum has a coefficient that was chosen randomly from $\{1, \ldots, 5\}$. Also, the time series used in each factor was chosen randomly from $\{1, \ldots, J\}$, its lag was chosen randomly from $\{1, \ldots, 20\}$, and each transformation was chosen randomly from the set of mappings $\{x \mapsto x, x \mapsto x^2, x \mapsto \sin(x), x \mapsto \cos(x), x \mapsto \ln(|x|)\}$. For example, one of the generated series $y(t)$ is given by

$$y(t) = 3x_4(t-15)\ln(|x_4(t-14)|)x_8(t-2) + 4\cos(x_1(t-4)) + \epsilon(t), \tag{2}$$

where $x_1, x_4, x_8$ are time series of normalized stock data and $\epsilon(t)$ denotes the Gaussian noise. Finally, 1000 observations of $y(t)$ are generated.

**Performance evaluation** We evaluated each of the methods described in Section Other Methodsusing its average $F_1$-score over 20 randomly generated critical variables $y(t)$. A given time series $x_j$ is actually relevant to $y(t)$ if $x_j$ appears in the equation defining the evolution of $y(t)$; for example, if $y(t)$ is defined by (2) then the series $x_1, x_4, x_8$ are actually relevant to $y(t)$. Then for each method, a model is fit to the data for $y(t)$, where the regressors are the lagged variables for all 10 stocks up to a maximum lag of 20. Next, a given time series $x_j$ is deemed to be relevant by the method if $\alpha_j \geq \underline{\alpha} := 0.2$. For a given method, letting $tp$ denote the number of true positives (i.e. the number of time series deemed relevant by the method that are actually relevant), letting $fp$ denote the number of false positives, and letting $fn$ denote the number of false negatives, the *precision* of the method is $P := \frac{tp}{tp+fp}$, and its *recall* is $R := \frac{tp}{tp+fn}$.

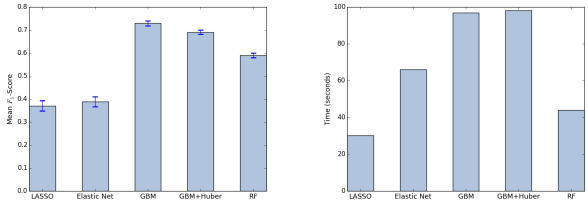
A method's $F_1$-score is the harmonic mean of its precision and recall, i.e., $F_1 := 2 \cdot \frac{PR}{P+R}$.

**Results** The average $F_1$-scores for each method, and their corresponding standard errors, over 20 generated critical variables are shown in Figure 2. The LASSO and Elastic Net parameters were obtained using 3-fold cross-validation. The results for gradient boosting with regression trees were obtained with both the squared (GBM) and Huber (GBM+Huber) loss functions.

As expected, the more flexible non-parametric methods significantly outperformed the linear methods on the highly non-linear data. In addition, only slightly more time was needed compared to the linear methods; see Figure 2. Overall, gradient boosting exhibited an attractive balance between performance and time requirements, and random forests achieved much better performance than the linear methods while using a similar amount of time.

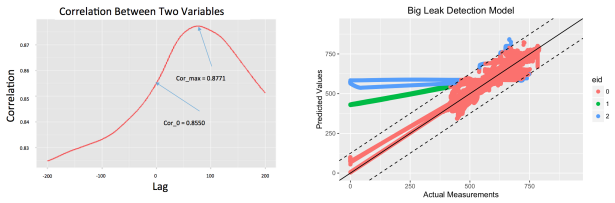
# Application

The system under investigation consists of 24,000 miles of natural gas pipeline, including 21,242 miles of distribution




(a) (b)

Figure 2: (A) Mean $F_1$-scores over 20 runs, with standard errors. (B) Total times (in seconds) needed to obtain the results.




(a) (b)

Figure 3: (A) The CCF between two stations. (B) The big leak model output for the location 1.

and 2,301 miles of transmission pipelines, and about 40 compressor stations and distribution centers. The system has extended lengths of pipelines exposed to harsh environments in the mountains. In particular, they are subject to damage from corrosion and landslides, which are the major causes of leaks. The SCADA system contains 65,000 measurement points (referred to as tags); the sampling frequency of the data is 30 seconds. The total volume of data for the period of 7 years is about 1.7 TB. We performed several data scoping and processing procedures before applying the gas leak detection model including smoothing, outlier removal, and short spike removal on the raw data.

## Big Leak Detection Results

The total number of leaks during the seven-year period was three. The identities of the locations and leak incidence times were not known by the authors before the development of the detection models. 5 most important locations was defined and the leak detection model was applied to these locations first. We present the detailed results for one location (L2) and summarize the results for the three locations with past leak events in this section.

Figure 3 illustrates the temporal delay between measurements from two stations. The CCF peaks at a lag of 76, which corresponds to 38 minutes. The big leak detection model output for location 2 (L2) is shown in Figure 3. As is shown, the model correctly captured the leak event, with only 1 false positive event, whereas the false positive rate of SCADA system is 4148, and 9 from the mass balance model. We'd like to point out that the identified false positive event

Table 1: The event detection summaries of the big leak detection model.

| Location | | L1 | L2 | L3 |
|---|---|---|---|---|
| # of event | | 1 | 1 | 0 |
| SCADA | LT (hour) | 0 | 0 | - |
| | $F_1$ | 0.001 | 0.0004 | - |
| | FP | 1664 | 4148 | 4775 |
| Mass-Balance | LT (hour) | 0.9 | 17 | - |
| | $F_1$ | 0.11 | 0.18 | - |
| | FP | 16 | 9 | 4 |
| Our Model | LT (hour) | 9 | 21 | - |
| | $F_1$ | 0.18 | 0.67 | - |
| | FP | 9 | 1 | 0 |

Note: LT – Lead Time, FP – False Positive. It is estimated that most of the false positives correspond to maintenance activities. Since no accurate records of such activities were available, no validation is provided.

was a valve replacement (eid=1). Thus in a real application system, if such planned maintenance events were known to the model, the false positive rate would be even smaller The actual leak event (eid=2) that occurred on November 11th, 2012. And most importantly, the detection time of the event was 21 hours before the rupture occurred. Shortly after midnight on Nov.11th, the compressors operated to ramp up the pressure, and the model identified that the pressure was not at the level where it should be. In fact, the residual between the actual pressure and predicted pressure was 14 standard deviations away from its normal value, which only happened twice in 7 years. Thus it is a very strong signal of a severe anomaly in the system. 21 hours later, the leak escalated to rupture, causing a rapid loss of pressure which was caught by the SCADA alarm at 9:59PM. The rupture event not only released a huge amount of gas into the atmosphere, but also impacted thousands of customers during the cold winter weather. Such a rupture may have been prevented if the leak was detected earlier. In addition to this incidence, we summarize the event detection lead times and $F_1$-scores for our model, the SCADA alarms, and a mass-balance model in Table 1. We want to point out that such a big event can be detected by most models, even simple ones such as those based on pressure monitoring or mass balance. However, the value of our model lies in the sensitivity of the detection model. Our model has shown superior capability in detecting medium leaks leading to the rupture event.

## Small Leak Detection Results

In this section, we present detailed results for the location where there was a leak. The small leak was discovered by the operation team during their annual line inspection in late June 2010, when a helicopter was used to visually examine the pipelines for problems. It was found that the pipeline was damaged by landslides, but since the damage did not cause a significant pressure drop, it was left unnoticed for a long period of time. The operation personnel then examined the SCADA records and estimated that the leak may

Table 2: The event detection summaries of the small leak detection model.

| Location | | L1 | L2 | L3 |
|---|---|---|---|---|
| # of event | | 0 | 0 | 1 |
| SCADA | LT (day) | - | - | 0 |
| | $F_1$ | - | - | 0 |
| | FP | 1665 | 4149 | 4775 |
| Mass-Balance | LT (day) | - | - | 80 |
| | $F_1$ | - | - | 0.4 |
| | FP | 17 | 10 | 3 |
| Our Model | LT (day) | - | - | 73 |
| | $F_1$ | - | - | 1 |
| | FP | 0 | 0 | 0 |

Note: LT – Lead Time, FP – False Positive. It is estimated that most of the false positives are some maintanence activities. Since no accurate records of such activities were available, no validation was provided.

have been present since April, which was later confirmed by the CUSUM model. As is shown in Figure 4, starting on April 17th 2010, the lower CUSUM value dropped below a value that has a probability $1.38 \times 10^{-24}$, which clearly indicates the presence of an event in the system . The threshold was chosen as $\log(C^-) < 6$ (i.e., $C^- < -403$). Using the threshold, the model identified the true event correctly. The $F_1$-scores and false positive rates for event detection for the three locations are summarized in Table 2.

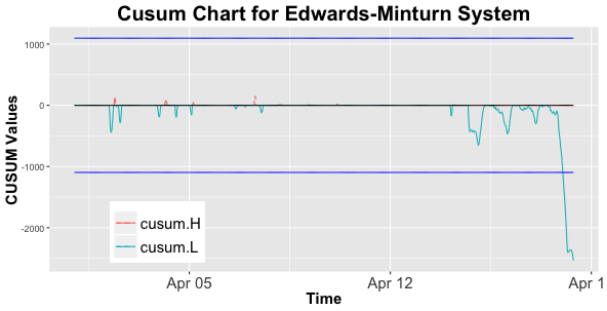



Figure 4: The lower CUSUM and upper CUSUM chart for Location 3. It is shown that the lower CUSUM has exceeded the threshold level (the lowest horizontal line).

## Results for Other Methods

The other more generic anomaly detection methods described in Section Other Methodswere applied location 1. Location 1 is a more complex system, with more false alarms reported by the big leak detection model. The data for the year of 2008, which is when the leak event happened, was used for this study. For location 1, there are 330 raw measurements from the SCADA system. Some of the data have near zero variance, and were filtered out. In the end, 52 measurements were used for the generic anomaly detection methods. The maximum lag used in the models was 50, which resulted in a total of 2,600 features used.

Table 3: The event detection accuracy summary between other method and big leak model.

| Method | Big Leak | LASSO | Elastic Net | GBM | GBM+ Huber | RF |
|---|---|---|---|---|---|---|
| $F_1$-score | 0.33 | 0.5 | 0.5 | 0.4 | 0.4 | 1 |
| Lead Time(h) | 8.85 | 8.68 | 8.31 | -18.74 | -0.56 | -19.72 |
| computation Time (s) | 30.79 | 148.45 | 306.91 | 807.28 | 830.93 | 1130.37 |

We summarize the leak detection performance in terms of lead time and $F_1$ -score in Table 3. It was found that the LASSO and Elastic Net perform the best in terms of lead time. They provided about the same lead time that we obtained from the big leak detection model. We also compared the features selected by the generic methods and the features we selected in the big leak detection model; it was found that about 3 out of 5 of the features in the big leak detection model were also selected by the generic methods along with other common features selected by most of the generic methods that were not used in the expert model. However, the computation time for such generic models are much longer than the proposed expert model, as is shown in Table 3.

## Conclusion

In this paper, we presented two innovative and effective leak detection models: the big leak detection model that identifies medium size leaks preceding big rupture events, and the small leak detection model that identifies small leaks lasting for extended periods of time. The models were tested on a real high pressure natural gas transmission system. It was found that the big leak detection models outperform the SCADA alarms by over 180 times in terms of $F_1$-scores. And more importantly, the model provides a vital opportunity for the prevention of catastrophic rupture events, by detecting small / medium leaks preceding the rupture events. Both the big leak and small leak detection models were developed using carefully selected features. We then introduced other methods that use all of the measurement data in the SCADA system, and considered penalized regression and non-linear methods for the task of big leak detection. It was found that the LASSO and Elastic Net perform the best in terms of detection lead times. However, the computation time for the other methods are much higher compared to our big leak detection method. To conclude, we developed a novel approach to addressing the task of detecting leaks in high pressure gas transmission networks. The approach does not require additional instrumentation or new sensing systems other than the existing SCADA system. The approach has been proven better than the SCADA alarm system, a mass balance model and other generic time series anomaly detection models in terms of detection accuracy and computation time. Furthermore, while it is computationally more expensive, the proposed time series based method is less dependent on expert knowledge for model development, and has less dependency on specific sensor instrumentation, and could be useful in some circumstances.

## References


Arnold, A.; Liu, Y.; and Abe, N. 2007. Temporal causal modeling with graphical Granger methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, 66–75. New York, NY, USA: ACM.

Bai, L.; Yue, Q.; and Li, H. 2005. Pipeline fluid monitoring and leak location based on hydraulic transient and extended Kalman filter. *Jisuan Lixue Xuebao* 22:739–744.

Basseville, M.; Nikiforov, I. V.; et al. 1993. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs.

bin Md Akib, A.; Bin Saad, N.; and Asirvadam, V. 2011. Pressure point analysis for early detection system. In *Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on*, 103–107. IEEE.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM Comput. Surv.* 41(3):15:1–15:58.

Dos Santos, P. L.; Azevedo-Perdicoűlis, T.-P.; Ramos, J.; De Carvalho, J. M.; Jank, G.; Milhinhos, J.; et al. 2011. An LPV modeling and identification approach to leakage detection in high pressure natural gas transportation networks. *Control Systems Technology, IEEE Transactions on* 19(1):77–92.

Isa, D., and Rajkumar, R. 2009. Pipeline defect prediction using support vector machines. *Applied Artificial Intelligence* 23(8):758–771.

Jones, M.; Nikovski, D.; Imamura, M.; and Hirata, T. 2016. Exemplar learning for extremely efficient anomaly detection in real-valued time series. *Data Mining and Knowledge Discovery* 1–28.

Kim, M.-S., and Lee, S.-K. 2009. Detection of leak acoustic signal in buried gas pipe based on the time–frequency analysis. *Journal of loss Prevention in the Process Industries* 22(6):990–994.

Laptev, N.; Amizadeh, S.; and Flint, I. 2015. Generic and scalable framework for automated time-series anomaly detection. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 1939–1947. New York, NY, USA: ACM.

Lay-Ekuakille, A.; Vendramin, G.; and Trotta, A. 2009. Spectral analysis of leak detection in a zigzag pipeline: A filter diagonalization method-based algorithm application. *Measurement* 42(3):358–367.

Ma, C.; Yu, S.; and Huo, J. 2010. Negative pressure wave-flow testing gas pipeline leak based on wavelet transform. In *Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010 International Conference on*, volume 5, 306–308. IEEE.

Markey, S. E. J. 2013. America pays for gas leaks – natural gas pipeline leaks cost consumers billions. Technical report, The House Natural Resources Committee.

Martins, J. C., and Seleghim, P. 2010. Assessment of the performance of acoustic and mass balance methods for leak detection in pipelines for transporting liquids. *Journal of Fluids Engineering* 132(1):011401.

Qiu, H.; Liu, Y.; Subrahmanya, N. A.; and Li, W. 2012. Granger causality for time-series anomaly detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, 1074–1079. IEEE.

Sivathanu, Y. 2003. Natural gas leak detection in pipelines. *Technology Status Report, En'Urga Inc., West Lafayette, IN*.

Stearns, S. V.; Lines, R. T.; Grund, C. J.; and Philbrick, C. R. 2005. Active remote detection of natural gas pipeline leaks. *US Department of Energy National Energy Technology Laboratory Technology Status Report, viewed July* 18:2008.

Takeishi, N., and Yairi, T. 2014. Anomaly detection from multivariate time-series with sparse representation. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, 2651–2656. IEEE.

Wan, J.; Yu, Y.; Wu, Y.; Feng, R.; and Yu, N. 2011. Hierarchical leak detection and localization method in natural gas pipeline monitoring sensor networks. *Sensors* 12(1):189–214.

Wang, X.-J.; Lambert, M. F.; Simpson, A. R.; Vitkovsky, J. P.; et al. 2001. Leak detection in pipelines and pipe networks: a review.