Anomaly Detection For Natural Gas Regulator Stations

by

Adam Christopher Chao

B.S., Massachusetts Institute of Technology, 2008

Submitted to the MIT Sloan School of Management and the Institute for Data, Systems, and Society in partial fulfillment of the requirements for the degrees of

Master of Science in Engineering Systems

and

Master of Business Administration

in conjunction with the Leaders for Global Operations Program at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author ..... Signature redacted

MIT Sloan School of Management and the Institute for Data, Systems, and Society

May 6, 2016

Certified by ..... Signature redacted

Saurabh Amin, Thesis Supervisor

Robert N. Noyce Career Development Assistant Professor, Department of Civil and Environmental Engineering

Certified by ..... Signature redacted

Georgia Perakis, Thesis Supervisor

William F. Pounds Professor of Management Science, MIT Sloan School of Management

Approved by ..... Signature redacted

John N. Tsitsiklis

Clarence J. Lebel Professor of Electrical Engineering, IDSS Graduate Officer

Approved by ..... Signature redacted

Maura Herson

Director, MBA Program, MIT Sloan School of Management

Anomaly Detection For Natural Gas Regulator Stations

by

Adam Christopher Chao

Submitted to the MIT Sloan School of Management and the Institute for Data, Systems, and Society on May 6, 2016, in partial fulfillment of the requirements for the degrees of
Master of Science in Engineering Systems
and
Master of Business Administration

Abstract

Natural gas regulator stations control the flow of gas across PG&Es gas transmission and distribution system. Ensuring the proper functioning of these stations is critical for the safety of the natural gas system. Currently, PG&E uses sensors linked to a Supervisory Control and Data Acquisition (SCADA) system to monitor pressure and other characteristics of select regulator stations, with continuing installation of new sensor systems across the network. PG&E seeks to develop algorithms for detection and prediction of safety issues before they occur, as well as monitor performance degradation in a regulator station.

First, analysis of historical failure events was conducted to better understand the varying causes of regulator overpressure events and their corresponding downstream pressure patterns. Then, downstream pressure time-series data was collected and processed for each regulator station. Useful features from these time-series were extracted, including day-to-day changes and moving averages. Piecewise linear segmentation was also performed on the time-series to extract relevant features.

These features were then used to cluster stations by their operating characteristics, grouping stations with similar volatility and pressure patterns. Anomaly detection methods were then developed and calibrated for the station clusters. We use a variety of statistical process control techniques, including CUSUM and EWMA to detect changes in the behavior of a regulator downstream pressure time-series. Detection algorithms were then evaluated with and without clustering using ROC curves on simulated pressure anomalies.

Ultimately, we show that modified CUSUM and adaptive sliding window techniques can detect pressure anomalies in natural gas regulators with reasonable false positive rates. We also show how improvements to data handling and sharing at PG&E can facilitate better algorithms for regulator anomaly detection.

Thesis Supervisor: Saurabh Amin
Title: Robert N. Noyce Career Development Assistant Professor, Department of Civil and Environmental Engineering

Thesis Supervisor: Georgia Perakis

Title: William F. Pounds Professor of Management Science, MIT Sloan School of Management

Acknowledgments

First I would like to thank my academic advisors, Professor Georgia Perakis and Professor Saurabh Amin for their valuable insights, technical guidance and strong support throughout the project.

I would also like to thank the leadership and employees at Pacific Gas and Electric for providing such a great and rewarding experience. Specifically, Mallik Angalakudati and Melvin Christopher provided great support as project champions. Daniel Menegus and George Gaebler in GCSS were great at providing needed resources and linking me up other groups. Bryan Hennessy was fantastic at handling numerous data challenges and IT issues.

Contents

1 Introduction and Background **10**

1.1 Company Overview 10

1.2 Overview of PG&E Natural Gas System and Regulator Stations 10

1.3 Problem Statement and Goals 13

1.4 Contributions and Key Findings 13

1.5 Thesis Overview 14

2 Literature Review **15**

2.1 Overview of Time-Series Anomaly Detection 15

2.2 Data Mining and Feature Extraction 16

2.3 Statistical Process Control Techniques 17

2.4 Model Based 19

2.5 Time-Series Clustering Methods 20

2.6 Predictive Maintenance 21

3 Methodology **22**

3.1 Overview of Methodology 22

3.2 Data Collection 22

3.3 Analysis of Historical Overpressure Events 23

3.4 Feature Extraction 25

3.4.1 Daily Average Pressure Features 26

3.4.2 Intra-day Pressure Features 33

3.5 Clustering of Stations 34

3.6 Detection Techniques 40
3.6.1 CUSUM-EWMA 40
3.6.2 Local Regression 40
3.6.3 Adaptive Window 43
3.6.4 Conclusion 44

4 Testing and Results **45**
4.1 Overview of Testing and Results 45
4.2 Testing and Evaluation of Detection Methods 45
4.2.1 Simulation of Overpressure Events 45
4.2.2 Detection Evaluation 46
4.3 Identification and Classification of Stations 50

5 Recommendations and Conclusions **55**
5.1 Recommendations for Data Handling and Sensor Installation 55
5.2 Ideal Model for Anomaly Detection 57
5.3 Applications in Other Areas 58
5.4 Conclusion 60

List of Figures

1-1 Diagram of Natural Gas Pressure Regulator 11

1-2 Distribution regulator undergoing maintenance 12

3-1 Methodology of Thesis 22

3-2 Hours-Scale Overpressure Events 25

3-3 Days-Scale Overpressure Events 25

3-4 Hourly and Daily Average Downstream Pressure 26

3-5 Histogram and Density Plot of Daily Change in Pressure 27

3-6 Histogram and Density Plot of Moving Average Difference 28

3-7 Histogram and Density Plot of EWMA Difference 29

3-8 Histogram and Density Plot of Sliding Window Slope Coefficient 30

3-9 Time Series and its Piecewise Linear Approximation 31

3-11 Within-group Sum of Squares Vs Number of Clusters 36

3-12 Box-plots of Daily Average Change Across Clusters: Method 1 36

3-13 Box-plots of EWMA Difference Across Clusters: Method 1, $\lambda = 0.1$ 37

3-14 Box-plots of Local Regression Slope Coefficients Across Clusters: Method 1 . 37

3-15 Box-plots of Daily Average Change Across Clusters: Method 2 39

3-16 Box-plots of EWMA Difference Across Clusters: Method 2, $\lambda = 0.1$ 39

3-17 Box-plots of Local Regression Slope Coefficients Across Clusters: Method 2 . 39

3-18 CUSUM Detection Technique for Station 108 41

3-19 Local Regression Detection Technique for Station Q 42

3-20 Slope Coefficient Quantiles By Segment Length 43

4-1 Simulated Spike in Time-Series 46

4-2 Performance of CUSUM-EWMA on 2014 Dataset, Varying λ, K = 0 47
4-3 Performance of CUSUM-EWMA on 2015 Dataset, Varying λ, K = 0 47
4-4 Performance of CUSUM-EWMA on 2014 Dataset, Varying K, λ = 0.1 48
4-5 Performance of CUSUM-EWMA on 2015 Dataset, Varying K, λ = 0.1 48
4-6 Performance of Local Regression Method on 2014 Dataset, Varying W 49
4-7 Performance of Adaptive Window Method on 2014 Dataset, Varying β . . . 49
4-8 Performance of Adaptive Window Method on 2015 Dataset, Varying β . . . 50
4-9 Performance of Detection Methods on 2014 Dataset, By Cluster For Cluster Method 1 51
4-10 Performance of Detection Methods on 2014 Dataset, By Cluster For Cluster Method 2 52
4-11 Sawtooth Pattern 53
4-12 Set-point Changes 53
4-13 Winter Droop 53
4-14 Weekend-Weekday Swings 54
4-15 Other Anomalous Pressure Patterns 54

5-1 Downstream Pressure vs HDD 58
5-2 Example of Flow Versus Downstream Pressure Variations 59
5-3 Illustration of Ideal Model 59

List of Tables

3.1 MOP/HiHi Events Jan 2013-Jan 2016, By Cause 24
3.2 Time Scale of Equipment Failures 24

Chapter 1

Introduction and Background

1.1 Company Overview

Pacific Gas & Electric is a large investor-owned utility operating in northern and central California, handling transmission and distribution of both electricity and natural gas. PG&E has about 4.3 million natural gas customer accounts. Transporting this gas, PG&E owns 42,000 miles of distribution pipeline and 6,400 miles of transmission pipeline. The California Public Utilities Commission (CPUC) regulates the natural gas system in California.

1.2 Overview of PG&E Natural Gas System and Regulator Stations

PG&E obtains natural gas from wells in Canada, the Rockies, the Southwest, as well as California. The natural gas, after processing to remove impurities, flows through the transmission system. Functionally, transmission pipelines operate at pressures above 60 PSIG, while distribution lines operate below 60 PSIG. Large backbone pipelines branch off into local transmission pipelines. Compressor stations increase the gas pressure in the lines to ensure flow. For purposes of maintenance, construction and control, PG&E's gas system is divided into 18 regional divisions.

Regulator stations control the pressure of gas in the transmission and distribution lines,

thereby ensuring adequate flow of gas to customers. Transmission regulator stations lower the pressure from one section of transmission pipe to another, while distribution regulator stations link transmission and distribution lines. PG&E owns 2,407 distribution regulator stations and 631 transmission regulator stations.

Natural gas regulator stations use a mechanical controller to the alter and control the flow of gas through the network. The basic regulator is composed of a diaphragm and spring. The force the spring exerts on the diaphragm is controlled by a piece of equipment called the pilot. During maintenance, technicians can set the pressure of the regulator by adjusting the spring. Regulators typically have a secondary, backup regulator called a monitor ahead of the main equipment. This is set to a slightly higher pressure than the main regulator, so that in the event of a main regulator failure, the backup will take over.

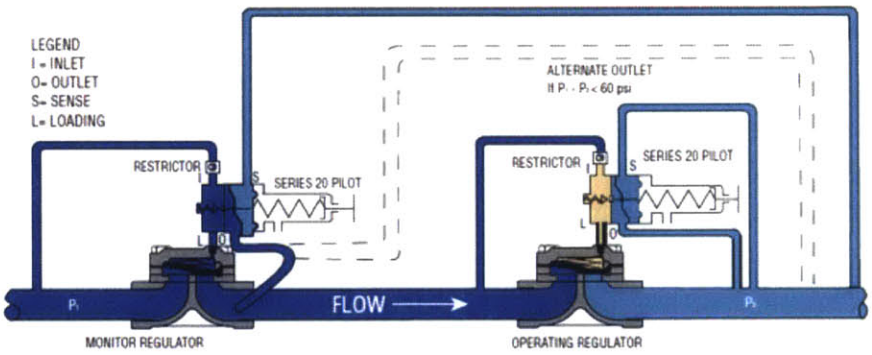


Figure 1-1: Diagram of Natural Gas Pressure Regulator

Safety of regulators is of paramount importance, given that failure of a regulator can result in high-pressure gas entering a low-pressure pipe. Overpressure events can be extremely dangerous and can result in large leakages of natural gas into a populated area. Typical failure modes involve the diaphragm not being able to fully close, due to contaminants, thereby causing high pressure gas leakage. Another area of failure is contaminant build-up in the pilot, which causes the regulator to be unable to maintain the desired pressure.

As part of increasing safety and improving operations, PG&E has been installing sensors on many regulator stations, linking them into the Supervisory Control and Data Acquisition (SCADA) system. The SCADA system allows for real-time monitoring of stations. Currently, about 445 transmission stations have real-time monitoring of SCADA feeds (RTUs,

Figure 1-2: Distribution regulator undergoing maintenance

Remote Terminal Units). About 214 distribution stations have either RTUs or are linked with ERX sensors, which report in remotely at given intervals of time. PG&E plans to continue expansion of the SCADA system. These SCADA sensor feeds are viewed in the Gas Control Center in San Ramon, CA.

The stations connected to SCADA typically measure the pressure downstream of the regulator. A select number of stations have pressure sensors upstream and between the monitor and regulator. Additionally, a handful of stations have sensors measuring the flow volume through the regulator. The output of a sensor is stored as a time-series of 20 second intervals. Currently, PG&E Gas Control sets upper and lower limits on the pressure signal. These are composed of the LowLow limit for underpressure events, and the HiHi and MOP limits for overpressure events. Pressures outside the limits will trigger an alarm at Gas Control, who will send a maintenance crew to respond. Typical maintenance response times

are about 20 minutes.

1.3 Problem Statement and Goals

While the fixed alarm limits have been adequate, PG&E is seeking to build a more intelligent, adaptive system to predict overpressure and underpressure events before they occur. Overpressure events have many causes; including human error, facility design issues, and equipment failures. Equipment failures include valve failures, debris contamination, sulfur build-up and liquids contamination. In particular, debris or sulfur can build up over time in the pilot or diaphragm of the regulator, preventing full closure of the regulator, and causing high-pressure gas to leak into the downstream pipeline. Therefore, the goal of this thesis is the generation and implementation of algorithms for predicting these unsafe conditions as well as detecting performance degradation in regulator stations. In addition, recommendations for improvements in data processing and sensor installation will be made, to improve future expansion of the SCADA system for predictive algorithms. For this thesis, I limited the scope to distribution regulator stations; transmission stations present an area for future research.

1.4 Contributions and Key Findings

This thesis contributes to the fields of time-series anomaly detection and predictive maintenance by modifying traditional statistical process control (SPC) techniques for use with volatile time-series with changing means. Techniques such as exponentially weighted moving averages, piecewise linear approximations, and local regressions are used to extract features for processing with these SPC methods. Therefore, even without knowing the target mean of a process, one can identify deviations from expected behavior. Another contribution is the use of characteristic-based time series clustering for grouping stations with similar pressure behavior.

Altogether, we find that intelligent feature extraction combined with traditional SPC methods can identify potential safety issues in a natural gas regulator, with acceptable levels

of false alarms. These techniques can be improved with further data, to reduce the false alarm rate. This emphasizes the importance of standardized information collection at PG&E in implementing predictive maintenance algorithms.

1.5 Thesis Overview

A review of time-series anomaly detection and clustering techniques is presented in Chapter 2, covering statistical process control, time-series models, and other approaches. In addition, current research regarding condition-based maintenance is reviewed.

The methodology for predictive analytics for natural gas regulator stations is presented in Chapter 3. The methodology covers data collection of the regulator station sensor time-series, as well as station characteristics. An analysis of historical overpressure events is conducted, showing the types of failures and corresponding pressure patterns for regulator stations. The chapter then delves into feature extraction of the downstream pressure time-series, using multiple methods to extract and process useful features from these data sets. We then show how these features are used to cluster stations, creating groups of stations with similar operating characteristics. The detection algorithms are then presented, with parameters calibrated for each cluster.

Chapter 4 shows the testing and results of the clustering and detection algorithms. In particular, we used simulated anomalies to estimate algorithm performance. We also analyze stations with abnormal pressure patterns that are susceptible to creating false alarms.

Chapter 5 presents conclusions and recommendations for improving predictive analytics at PG&E. We also present potential avenues for further developing predictive algorithms for regulator stations.

Chapter 2

Literature Review

2.1 Overview of Time-Series Anomaly Detection

In this thesis, we seek to identify regulator pressure observations which do not behave according to expected patterns. We are therefore concerned with the general problem of detection of anomalies or outliers in a time-series. Time-series anomaly detection is a large field, with substantial applications for financial analysis, medical diagnostics, computer network intrusion detection, and industrial monitoring.

A large number of surveys and reviews of anomaly detection and time-series analysis techniques have been published. Chandola et al [5] gives an overview of anomaly detection techniques, distinguishing between three types of anomalies. Point anomalies are individual points that deviate from the rest of the sample, whereas collective anomalies are collections of points (each of which could be non-anomalous alone) which deviate from the sample as a whole. Likewise, Chandola defines contextual anomalies as data points that deviate from their specific context given by the structure of the data-set. All three types of anomalies are of interest in analyzing regulator pressure time-series. There can be single points substantially outside the normal bounds of pressure, as well as abnormal pressures for a given time of day. Collective anomalies are also significant; ramps in downstream pressure due to sulfur or debris build-up can involve pressure readings which are not individually abnormal, but are collectively so.

Chandola goes on to distinguish between several different categories of anomaly detection

techniques. Classification-based techniques classify observations as normal or anomalous, a train a model on the data, such as a neural network or support vector machine to assign a particular observation to a given class. Distance-based techniques involve measuring the distance or similarity of an observation with other observations, such as examining the distance of a point with its kth nearest neighbors. Statistical techniques involve estimating the probability distribution of a given feature, with abnormal observations being those with low probability. Statistical techniques include both parametric and non-parametric models for estimating the probability distributions.

In addition to Chandola, both [23] and [13] give reviews of novelty detection methods, similarly distinguishing between different classes of detection techniques. Gupta [9] gives a review of time-series outlier detection, surveying different types of time-series data and different types of anomalies. In particular, there is a distinction between techniques that use static (offline) time-series data, versus streaming data, which is processed in an online manner. The RTUs of regulator stations report real-time data streams, and we are interested in algorithms for detecting anomalies in real-time.

2.2 Data Mining and Feature Extraction

The first step for outlier and anomaly detection of time-series is collecting the data and pre-processing it. Pre-processing time-series data allows extraction of relevant features from it; this enables smaller data storage requirements, faster computation times and better detection of relevant characteristics. Aggarwal [1], covers a large number of feature extraction and data mining techniques for time-series. With natural gas regulators, the RTUs report pressure readings at 20 second intervals; this results in significant amounts of data to store and process; pre-processing and feature extraction is necessary to condense these time-series for easier storage and anomaly detection.

A frequent method for time-series feature extraction involves extracting a subsequence of data from a sliding window [17]. For instance, we might extract the last M periods from a given time-series point. The subsequence can be compared using a given distance metric to other windows [3]. We can also extract features such as median or mean from a given

window, and use those to detect local outliers within that window [6] [2]. For a natural gas regulator, we can extract a sliding window, for instance, the last 7 days of pressure readings, and process that subsequence to detect local patterns.

Sometimes it is more effective to convert a numeric time-series into a sequence of discrete elements. Symbolic methods such as these allow for the use of techniques such as Markov chains and hidden Markov models for modeling. Lin and Keogh [19] created the Symbolic Aggregation approXimation to facilitate clustering and anomaly detection in time series. Georgoulas [8] uses the SAX method for fault detection in rolling element bearings.

For periodic time-series, discrete Fourier transform (DFT) is frequently used. By representing a time-series in frequency-space, we can analyze the spectrum and find anomalous frequencies. A non-periodic time-series can be converted into time-frequency space using the discrete wavelet transform (DWT). This effectively represents a time-series as a set of averaged differences. Analyzing these wavelet coefficients is frequently used for time-series anomaly detection [24].

2.3 Statistical Process Control Techniques

Once we have extracted our relevant features from the time-series, there are a large number of techniques for identifying anomalies. Outlier detection for industrial processes have traditionally centered around statistical process control techniques. Shewhart charts set upper and lower control limits, typically at $\mu \pm 3\sigma$, where μ is the process mean [21]. Data points above or below the control limits are considered outliers and trigger an alarm. These types of control charts typically assume that the time-series is stationary and not autocorrelated. PG&E currently uses this general method in monitoring regulator pressures, setting MOP/HiHi alarm levels to detect overpressures and LowLow levels to detect underpressures. The weakness of this method is due to volatility in the mean of the time-series. Whereas traditional Shewhart charts have a fixed process mean to measure deviations from, the natural gas regulators have pressure set-points that vary depending on maintenance conditions. Moreover, there is drift over time away from the set-point. Deviations from the mean are often highly correlated, as well, with ramps in pressure, rather than single outlying points.

For detecting shifts in the mean between the control limits, cumulative sum (CUSUM) or exponentially weighted moving average (EWMA) methods are often effective. CUSUM techniques accumulate a running total of deviations from the target mean μ_0; positive and negative deviations are accumulated separately using running statistics C^+ and C^-.

$$C_i^+ = max[0, x_i - (\mu_0 + K) + C_{i-1}^+] \tag{2.1}$$

$$C_i^- = max[0, (\mu_0 - K) - x_i + C_{i-1}^-] \tag{2.2}$$

The variable K is a reference value, or allowance for deviations. CUSUM can be shown to derived from generalized likelihood ratios, where the likelihood ratio of a given number of successive deviations is evaluated.

EWMA methods allow detection of small, long term shifts in the process by smoothing out variability with a moving average. In particular, an EWMA is given by:

$$z_i = \lambda x_i + (1 - \lambda) z_{i-1} \tag{2.3}$$

Where λ is a smoothing parameter which determines how much previous observations are weighted in the moving average. Upper and lower bounds can then be set on the EWMA, allowing detection of small shifts in the mean of the process. Similar to the Shewhart charts, traditional CUSUM and EWMA methods are difficult to implement with the natural gas regulator time-series due to the changing, frequently drifting mean.

Variations of these techniques have been subject of many studies. [15] uses an Adaptive CUSUM technique, where the reference value μ is given by a modified EWMA. This allows detection of both large and small shifts in the time-series. [30] uses a similar CUSUM method for detecting denial of service attacks. The CUSCORE technique can be viewed as a generalization of CUSUM, where a general fault signature f(t, δ, τ) is to be detected.

$$C_t = max[0, (x_t - \mu + k_t) f(t, \delta, \tau) + C_{t-1}] \tag{2.4}$$

When the fault signature is a step-function, the CUSCORE becomes the normal CUSUM

statistic. [27] compares CUSCORE techniques with another class of techniques using generalized likelihood ratios. [11] also compares a set of CUSCORE, CUSUM and GLR methods for detecting mean-shifts. Another method, involving change-point detection with t-statistics is used for SPC in [12]. Sequential t-tests for change-point detection have also been used in estimating shifts in climate data [26].

2.4 Model Based

For non-stationary and/or autocorrelated time-series, anomaly detection can be performed by fitting a parametric model to the data. In particular, Autoregressive Integrated Moving Average (ARIMA) models are frequently used for modeling time-series, and analysis of the resulting residuals can be used for anomaly detection. [4] uses an ARMA model for structural health monitoring, while [7] uses piecewise AR models for estimating change-points in a time-series. [29] uses a CUSCORE method on the residuals of an ARMA process to estimate mean shifts.

Other anomaly or change-point detection methods involve fitting two separate models around a given point, and analyzing how the model parameters vary. For instance, a two phase linear regression can be fitted around point c:

$$x_t = \begin{cases} \mu_1 + \beta_1 t + \epsilon_t & 1 \leq \text{x} \leq \text{c} \\ \mu_2 + \beta_2 t + \epsilon_t & \text{c} + 1 \leq \text{x} \leq \text{n} \end{cases} \tag{2.5}$$

The values of β_1 and β_2 can be compared to detect a change-point, or the two phase regression can be evaluated against a single regression. These methods are frequently used for estimating change-points in climate data [25]. These methods can be used for detecting chnage-points in the regulator pressure time-series; however, they don't indicate whether the new time-series regime is abnormal. Given the volatility in the time-series, structural break-points are frequently detected; other methods are needed for ascertaining whether the new behavior after the break-point is a potential safety concern.

For general time-series anomaly detection, more targeted models can be generated for specific circumstances, for example, models for detecting possible epidemic outbreaks often

use a variety of social, demographic and medical data [28]; residuals then are tested using Shewhart or CUSUM methods. Ideally, for natural gas regulators, we'd use a series of sensor feeds (such as flow and upstream pressure) as inputs into a predictive model for downstream pressure. However, only a handful of stations have sensors beyond downstream pressure ones.

2.5 Time-Series Clustering Methods

In addition to anomaly detection, we are also interested in grouping similar time-series together. This is useful both for tailoring models to a specific cluster, and for identifying how time-series vary with each other. Liao [18] gives an extensive overview of time-series clustering methods. There are many algorithms for the general problem of clustering; hierarchical methods cluster data objects into a tree, k-means groups objects into k distinct clusters, and other model-based approaches assume some type of underlying structure to given clusters. In particular, we use k-means for grouping time-series in this thesis. This algorithm seeks to minimize the sum of distances between observations and their respective cluster centers, for a given k clusters. The objective function is:

$$MinJ(U,V) = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik} \|x_i - v_k\|^2 \tag{2.6}$$

Where u_{ik} is the cluster assignment of point i to cluster k, and $\|x_i - v_k\|^2$ is the distance between cluster center v_k and the observation x_i. While Euclidean distance is most common, time-series clustering also can make use of other distance-metrics, such as correlation coefficient or dynamic time warping (DTW) distance.

Clustering algorithms can use the raw time-series data, or can use extracted features from a time-series to cluster. In particular, for long time series, the high dimensionality means that we're often more interested in grouping time-series by overall characteristics, not by Euclidean distance. Several studies show how to extract features which identify the characteristic behavior of the time-series, and then use hierarchical or k-mean clustering on those features [31] [20]. This characteristic-based clustering is important for this project;

the high-level structure of the pressure time-series varies mainly due to the maintenance set-points of the regulators. We're interested in grouping stations by overall behavior, including volatility, occurrence of certain short-term pressure patterns, and other characteristics.

2.6 Predictive Maintenance

We also examine how these different anomaly detection techniques are applied to the field of condition-based maintenance (CBM). Jardine et al [14] gives an extensive review, defining CBM as a "maintenance program that recommends maintenance actions based on the information collected through condition monitoring." Giving increased pervasiveness of sensors, condition-based maintenance is a growing field. Sensor networks are used to measure performance of wind turbines [10], electric motors [22] and other industrial structures. Many of these studies use spectrum analysis, given that the machine under monitoring is often rotating.

Chapter 3

Methodology

3.1 Overview of Methodology

The overall methodology of this thesis is shown in Figure 3-1. Time-series and station data is collected from PGE's databases. Time-series, however, have high dimensionality, requiring feature extraction; the time-series need to be transformed into alternate representations to better understand the underlying patterns and behavior. Given the wide variability in station behavior, we next cluster stations which have similar behavior, allowing for more targeted anomaly detection. Finally, the actual anomaly detection methods are used to identify outliers based on the features extracted and the clusters formed. Time-series feature extraction, clustering and detection algorithms were implemented in R.

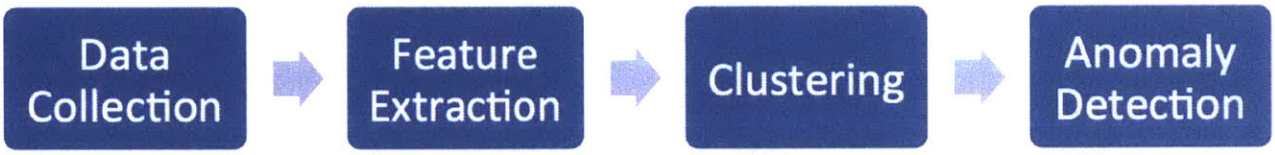


Figure 3-1: Methodology of Thesis

3.2 Data Collection

Several data sources were used in the development of the anomaly detection methods. Time-series composed of downstream pressure readings for 230 distribution regulator stations were

collected. The raw data is recorded at about 20 second intervals, but for this project was interpolated to 1 hour intervals. The time-series data is accessed through OSIsoft PI, a commonly used system for storing large sets of time-series data. The length of each time-series varies depending on the sensor installation date, but roughly half the stations have downstream pressure time-series since January 2014, and nearly all stations have pressure data from January 2015. Subsets of stations have time-series recording the pressure differential across the filter and/or the flow rate through the regulator station. This thesis is primarily focused on downstream pressure, but discussion of potential applications for other time-series is discussed in Chapter 5.

Additional station characteristics were compiled, showing the PG&E regional division, location, and overpressure limits for each regulator station. For overpressure events, Gas Operations has an Overpressure Elimination Team which compiles reports on both specific events and maintains a list of overpressure events since 2013. These reports contain descriptions of the event, and investigations into the causes.

3.3 Analysis of Historical Overpressure Events

Looking at historical overpressure events, we categorize by cause and time-scale of the pressure deviations. PG&E's current alarm system sets limits for Maximum Operating Pressure (MOP) and HiHi alarms. MOP levels are set by the design limits of the gas pipeline, whereas HiHi is typically set 1 or 2 PSI below the MOP level. In the Jan 2013-Jan 2016 time period, for distribution regulator stations, there were 272 MOP or HiHi alarms. PG&E's Overpressure Elimination Team conducts cause analysis on the alarms. Analyzing their reports, we find a cause breakdown shown below:

As shown, a substantial number of alarms (SP Too Close to MOP) are triggered due to the set-point of the regulator being set too close to the HiHi or MOP level. Likewise, a substantial fraction are triggered during maintenance procedures. For instance, in Station RZA, 2013, an overpressure event occurred when a filter was initially installed backwards during maintenance operations. While potentially serious, most maintenance-related overpressure events have the benefit of having PG&E personnel on-site to immediately correct

Cause	# Events	%
Equipment Failure	37	14%
Facility Design	3	1%
Maintenance-Related	29	11%
SP Too Close to MOP	183	67%
Calibration/Communication	15	6%
Unknown/Other	5	2%
Total	272	100%

Table 3.1: MOP/HiHi Events Jan 2013-Jan 2016, By Cause

the issue. False alarms are also occasionally an issue, with errors in sensor calibration or communication indicating an OP event where none actually exists.

Diagnosing specific equipment failures is more difficult, but of the 37 equipment-related failures, 5 had recorded sulfur build-up in the regulator and 8 showed debris (typically weld slag) contamination. Sulfur is naturally found in natural gas, and depending on the filtering and source of the gas, can build up in the diaphragm or pilot of the regulator. Likewise, debris such as welding slag can accumulate when construction or maintenance is performed upstream of the regulator.

Time-Scale	# Events	%
Days	5	25%
Hours	6	30%
Minutes	9	45%
Total	20	100%

Table 3.2: Time Scale of Equipment Failures

Looking closer at the time-scale and corresponding pressure patterns of equipment failures, we see that the time-scale varies. About half occur on the scale of minutes, very rapid changes in pressure. Another quarter are typically on the scale of hours. And the final quarter show pressure deviations on the scale of days. Apart from immediate spikes in pressure, equipment failures over longer time periods show a characteristic ramp in pressure. In particular, sulfur build-up created ramps in pressure over several days at Station Y and Z in 2014. Therefore, the anomaly detection algorithms are focused on detecting abnormal ramps in pressure, on the scale of days or hours.

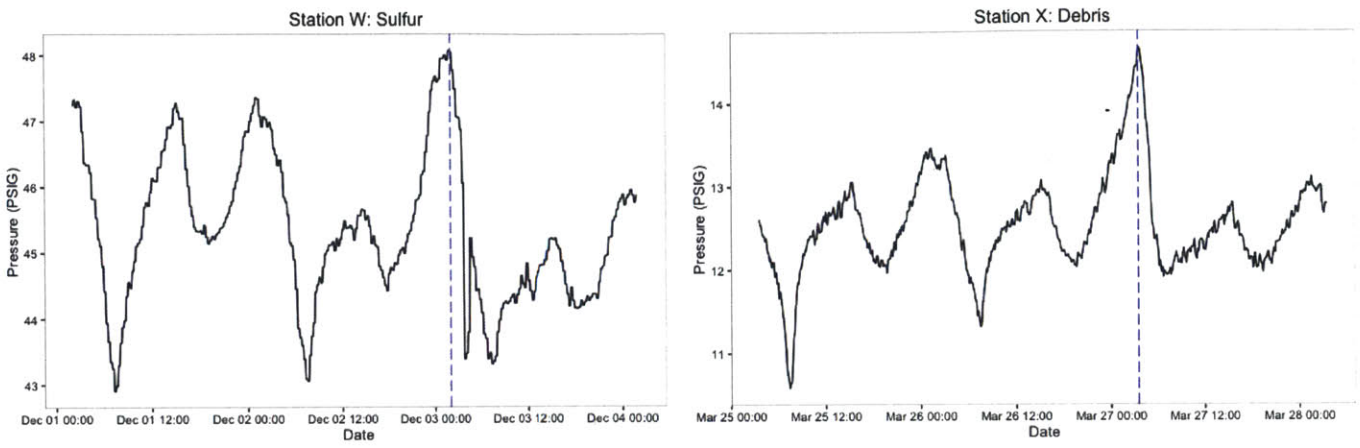


Figure 3-2: Hours-Scale Overpressure Events

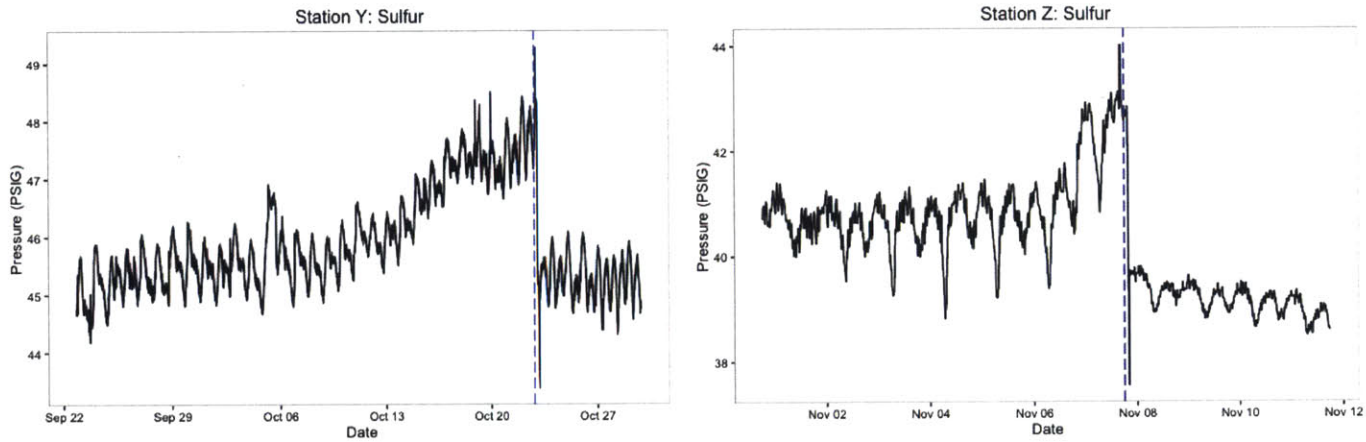


Figure 3-3: Days-Scale Overpressure Events

3.4 Feature Extraction

With the collected data, the next step is processing the time-series data to extract relevant features out of the time-series. For each station with have a time-series of downstream pressure readings at 1 hour intervals. The first step in feature extraction is separating out longer-term volatility with intra-day volatility. This is particularly important given the large amount of demand-driven volatility over the course of a day. Therefore, we separate out the daily average pressure from the intra-day fluctuations in pressure. Figure 3-4 shows the downstream pressure time-series for a station in Daly City, the black line being the hourly pressure, and the turquoise being the daily average. Note the sudden step in pressure in early 2015 when a maintenance crew adjusted the set-point of the regulator.

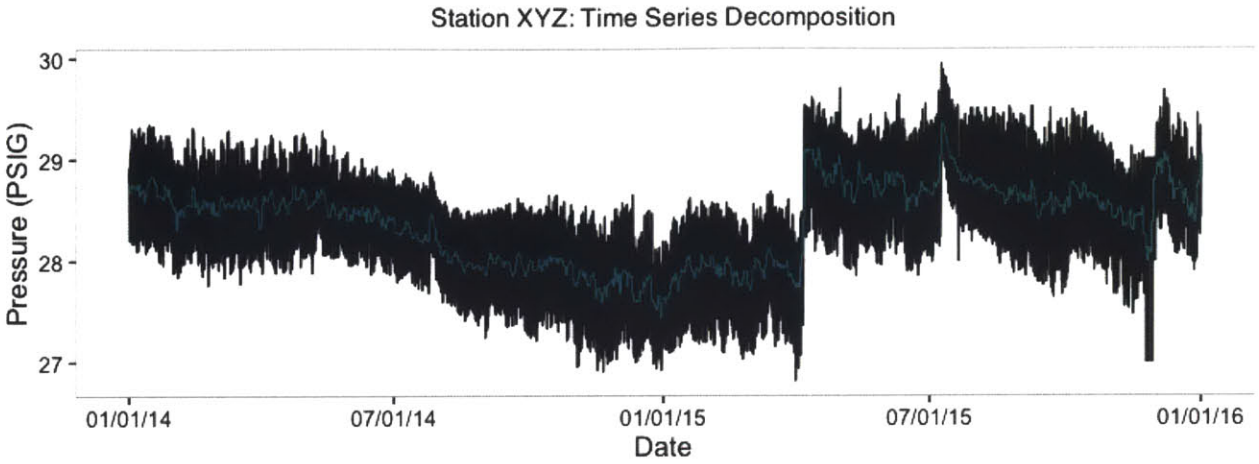


Figure 3-4: Hourly and Daily Average Downstream Pressure

Examining these features in the context of many stations over a year, we use both histograms and kernel density estimation to estimate and view the distributions of the extracted features. Whereas histograms assign observations to different "bins", kernel density estimation uses a weighting function called a kernel for estimating the probability density function given observations. The probability density at a given point x is estimated as:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \tag{3.1}$$

Where K is the kernel function. We use the common Gaussian kernel, so that the weight of each point's contribution dies off exponentially.

3.4.1 Daily Average Pressure Features

We first analyze the downstream pressure time-series on a longer scale, looking at the daily average pressure. Feature extraction and analysis of longer-term trends is relevant for detection of sulfur and/or debris build-up. Surprisingly, there is substantial variation of pressure patterns across stations and over long time scales. We define the downstream pressure time-series for a station as X $(x_1...x_t)$.

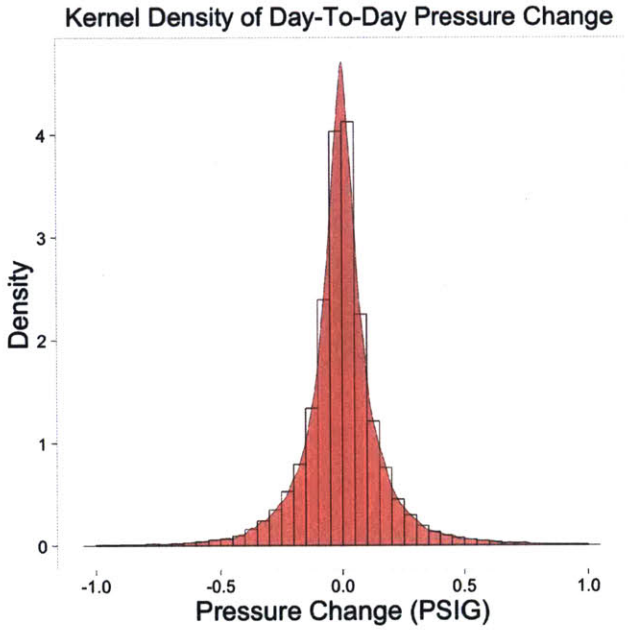
Daily Change in Average Pressure

We first examine the day-to-day change in daily average pressure, defined as:

$$\Delta x_t = x_t - x_{t-1} \tag{3.2}$$

This feature helps to indicate abrupt changes in pressure. This feature can also be viewed as modeling the pressure time-series as a random walk. We can look at the autocorrelation functions and partial autocorrelation functions to see if this is an appropriate approach.

Examining 90 distribution stations, over the year of 2014, we can estimate the following distribution and percentiles for Δx_i (Figure 3-5). The large majority of observations lie within ± 0.5 PSIG/day, meaning we can set bounds on the daily average pressure, based on the previous day's average pressure.



σ	90th	95th	99th
0.206	0.165	0.260	0.580

Figure 3-5: Histogram and Density Plot of Daily Change in Pressure

Moving Averages of Pressure

Another set of features involves comparing the time-series to longer moving averages. We can extract a window of length w from a time-series, creating a subsequence of observations

to transform into features. Use of these sliding windows can generate features that provide information about the local behavior around a particular observation. With a simple moving average, we define the feature to be:

$$\delta_t = x_t - \frac{1}{w} \sum_{j=t-w}^{t} x_j \tag{3.3}$$

Where w is the length of the moving average. Effectively, this feature measures the deviation from longer term trends. We examine the distribution of δ_t for window sizes of 3 days, 7 days and 14 days (Figure 3-6). The 99th percentile of this deviation varies from 0.405 PSIG for a 3-day moving average, to 0.934 PSIG for a 14-day moving average. Put another way, downstream pressures are typically within 1 PSIG of longer term moving averages.

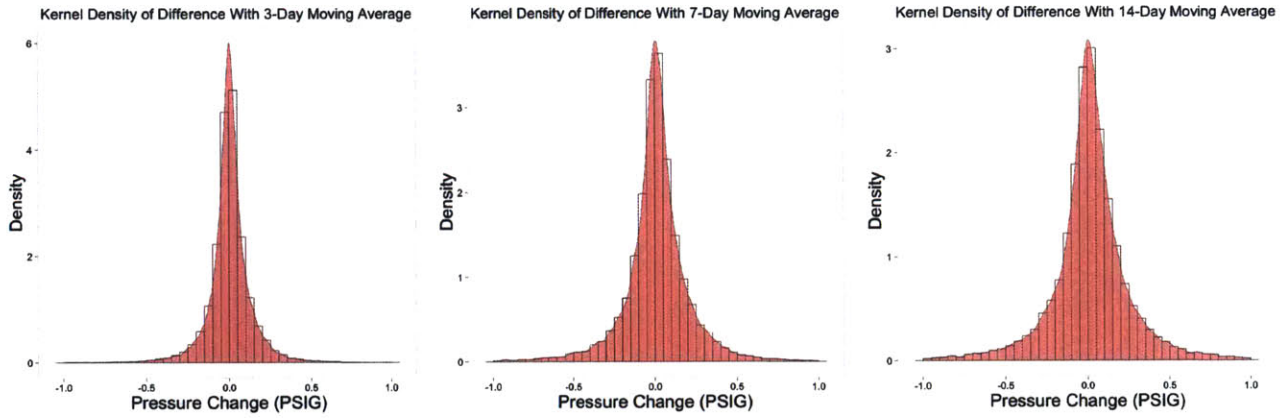


Figure 3-6: Histogram and Density Plot of Moving Average Difference

	σ	90th	95th	99th
3-day	0.159	0.138	0.206	0.405
7-day	0.267	0.222	0.330	0.650
14-day	0.360	0.284	0.432	0.934

Similarly, instead of a simple moving average, we can use an exponentially weighted moving average (EWMA), and examine the differences between a time-series observation and the EWMA, μ, using a smoothing parameter λ.

$$\mu_t = (1 - \lambda)\mu_{t-1} + \lambda x_t \tag{3.4}$$

$$\delta_t = x_t - \mu_t \tag{3.5}$$

Again, we examine the distribution of δ_t, for EWMAs with varying smoothing parameters (Figure 3-7). Similar with the simple moving averages, downstream pressure again tends to be within 1 PSIG of the longer-term moving averages. Therefore, we can use this feature to identify anomalous points as those significantly greater or less than 1 PSIG away from the longer trend.

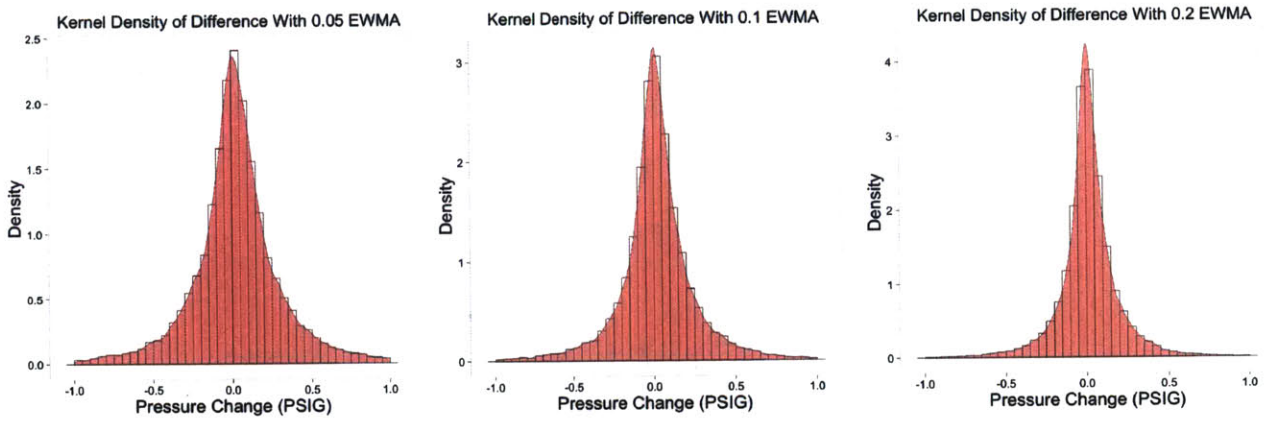


Figure 3-7: Histogram and Density Plot of EWMA Difference

λ	σ	90th	95th	99th
0.05	0.243	0.202	0.303	0.608
0.1	0.336	0.273	0.417	0.871
0.2	0.433	0.364	0.553	1.15

Local Regression of Pressure

Beyond taking moving averages, we can perform more complicated analysis on the subsequence of values extracted from a sliding window. In particular, for a window of size w, we can fit an OLS linear regression, and extract the coefficients and residuals as features. For windows of size 7, 14 and 28 days, we examine the distribution of slope coefficients and the error of prediction. Local regression therefore allows extraction of features related to the first order trends in a given window. Looking at the results, the 99th percentile of the slope coefficient varies from 0.233 PSIG/day for a 7-day window, to 0.073 PSIG/day for a 28-day window. Obviously, sustained ramps in pressure over longer time-periods are rarer than similar slope ramps over shorter time periods. We can use this feature for identifying abnormal ramps in pressure. For instance, a ramp of 0.5 PSIG/day over 7 days can be

considered an outlier, and would trigger an alarm.

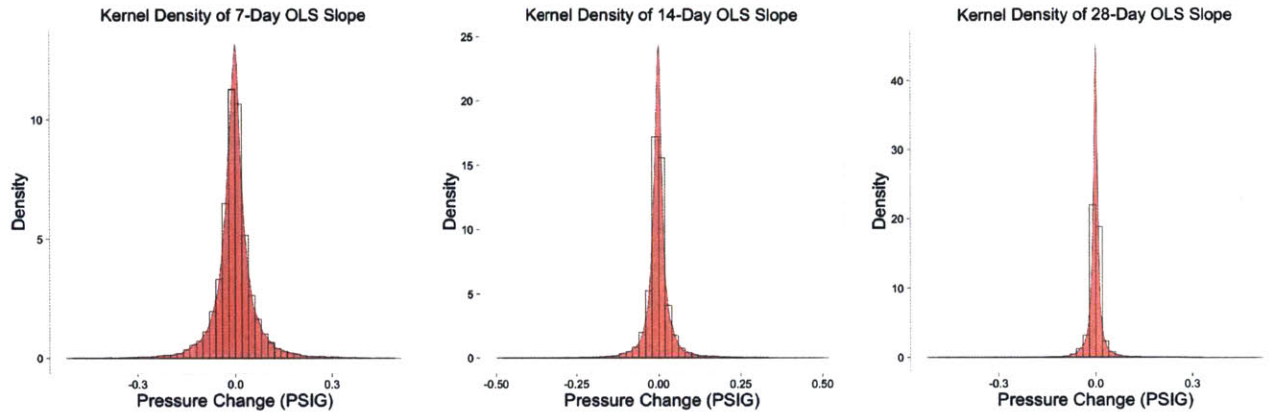


Figure 3-8: Histogram and Density Plot of Sliding Window Slope Coefficient

	σ	90th	95th	99th
7-day	0.081	0.060	0.097	0.233
14-day	0.049	0.032	0.053	0.142
28-day	0.029	0.016	0.029	0.073

Piecewise Linear Segmentation

For time-series representation it is often useful to represent a time-series with piecewise linear segments. This condenses a time-series into a more data-efficient format and aids in feature extraction and clustering. There are many methods of segmenting time-series, but we focus on two in this thesis, building off of techniques presented in [16].

First, we look at a binary, top-down approach. This works in an off-line mode, so it is unsuitable for real-time anomaly detection, although it can be used in a batch mode. However, it is still useful in generating features for classification and clustering of stations. The basic algorithm takes a time-series, and determines the best location for splitting the time-series into two linear segments. Each of those linear segments is then split by the algorithm. For defining the best location to split a time-series, we use a informational approach, minimizing the Bayesian Information Criterion (BIC). The BIC is defined as:

$$BIC = -2 \ln \hat{L} + k \ln(n) \tag{3.6}$$

Where $\hat{L}$ is the maximized likelihood of the model, k is the number of free parameters in

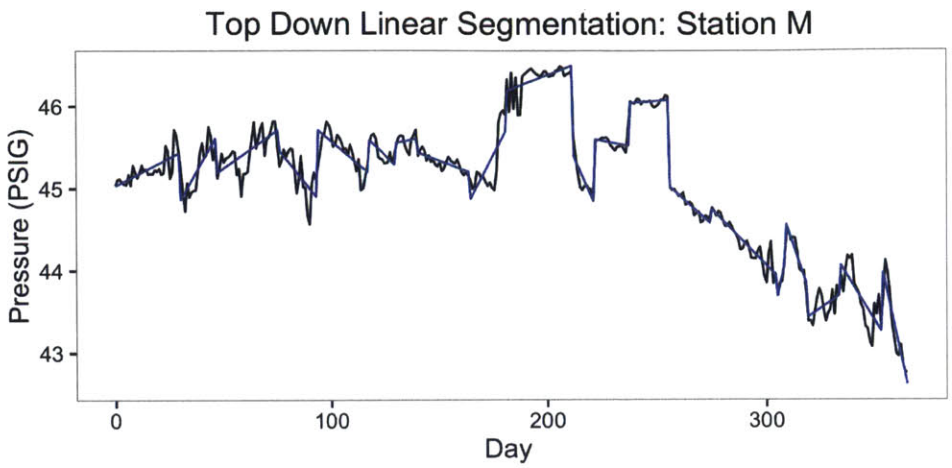


Figure 3-9: Time Series and its Piecewise Linear Approximation

the model, and n is the number of observations. In essence, the BIC penalizes more complex models (ie models with more free parameters). In splitting a time-series, we use the BIC to compare a model with one linear segment, vs a model with two linear segments split at the break-point. The pseudocode (implemented in R) for this top-down approach is shown below:

For online piecewise linear segmentation, a linear segment is grown from the beginning of the time-series until the some error bound is reached. A new segment is then started at the point where error bound was reached and the previous segment ended. The sequential, online nature of the segmentation algorithm allows it to be used both for feature extraction for clustering as well as real-time anomaly detection. There are many choices available for the nature of the error bound, we focus on using either the maximum absolute error or the mean squared error of the segment.

With both of these piecewise linear segmentation algorithms, we can now extract a set of features for each time-series.

Weekend-Weekday Differences

Another feature we extract is the difference between weekend and weekday pressure. Given that downstream pressure is heavily affected by flow through the regulator, large swings in pressure between weekday and weekend are indicative of significant industrial consumers downstream. These industrial consumers use significant quantities of natural gas during the

Algorithm 1: Top-Down Linear Segmentation

```
    procedure DetectChangePoint(x, y)          ▷ Find the Break-point for a Sequence
2:      Fit ← LinearModel(y ← x)
        BIC_1 ← BIC(Fit)
4:      for k in 2:length(x) do
            changetest ← x[k]
6:          Fit2 ← LinearModel(y ← x * (x < changetest) + x * (x >= changetest))
            BIC_2[k] ← BIC(Fit2)
8:      end for
        if min(BIC_2) < BIC_1 then
10:         return (changepoint = x[BIC_2 = min(BIC_2)])
        else return null
12:     end if
    end procedure
14: procedure FindChangePoints(x, y)▷ Recursively Find Changepoints In Time-Series
        changepoint ←DetectChangePoint(x,y)
16:     if changepoint = null then
            return (changepoint)
18:     else
            x1 ← x[x < changepoint]
20:         y1 ← y[x < changepoint]
            x2 ← x[x >= changepoint]
22:         y2 ← y[x >= changepoint]
            leftchange ←DetectChangepoint(x1,y1)
24:         rightchange ←DetectChangepoint(x2,y2)
            return (concatenate(leftchange, rightchange))
26:     end if
    end procedure
```

Algorithm 2: Online Linear Segmentation

```
   procedure SEGMENTTIMESERIES(x, y)                    ▷ Online Segmentation
 2:    segmentbegin ←1
       for i in 2:length(y) do
 4:       x1 ←x[segmentbegin:j]
          y1 ←y[segmentbegin:j]
 6:       Fit ←LinearModel(y1 ←x1)
          if Slope(Fit) > UCL then Alarm[j] ←1*
 8:       end if
          if max(residuals) > upper OR min(residuals) < lower then
10:          segmentbegin ←j
             changepoint[k] ←j
12:          k ←k + 1
          end if
14:    end for
       return changepoint
16: end procedure
```

Figure 3-10: Example of Stations with Different Weekend-Weekday Pressures

work-week, but upon shutting off for the weekend, there is a large flow decrease through the regulator and a subsequent increase in pressure. We define this feature for each station as:

$$Median[X_{weekend,i} - X_{weekday,i}] \tag{3.7}$$

Where $X_{weekend,i}$ is the average pressure for the station during the weekend for week i. Using the median for each station, rather than an average provides a more robust estimate. The feature aids in classification of stations by downstream demand; consumption data from Gas Planning was unavailable for the purposes of this study, so this feature serves as an effective proxy.

3.4.2 Intra-day Pressure Features

Intra-day fluctuations in pressure are primarily driven by consumer demand downstream of the regulator. Understanding these intra-day fluctuations is critical for prediction and anomaly detection of minute or hour-scale pressure anomalies. Pressure tends to peak around 1 AM, when flow is at its lowest. Downstream pressure declines substantially to about 7 AM,

as consumers use heat and hot water. Regulators show substantial variation in the magnitude of pressure variation over the course of the day, as well as the typical daily pattern.

3.5 Clustering of Stations

Because of the large variability in pressure patterns for different stations, we seek to group stations with similar behavior together, so that anomaly detection algorithms can be calibrated on an individual group. Most time-series clustering methods involve comparing the overall shape of each time-series, often using Euclidean distance. Distance metrics such as dynamic time-warping can group time-series with similar time-distorted shapes. However, the overall shape of each pressure regulator time-series tends to be governed by set-point changes from maintenance; we are more interested in grouping time series with similar operating characteristics and behavior, rather than similar high-level shape. Of particular interest is grouping well-behaved, low volatility stations together, and grouping hard-to-predict, highly volatile stations in another cluster. We use k-means clustering for a given set of features. Several different sets of features were tested and used for clustering.

Method 1: Daily Pressure Change Distribution

For Clustering Method 1, we cluster using just the distribution of daily change in average pressure (Δx_i). This should effectively group similarly volatile stations together. To cluster stations with similar distributions, we discretize Δx_i, by assigning observations to buckets y_i:

$$y_i = \begin{cases} a & \text{if } \Delta x_i > 0.5 \\ b & \text{if } 0.25 < \Delta x_i \leq 0.5 \\ c & \text{if } 0.1 < \Delta x_i \leq 0.25 \\ d & \text{if } 0.05 < \Delta x_i \leq 0.1 \\ e & \text{if } 0 < \Delta x_i \leq 0.05 \\ f & \text{if } -0.05 < \Delta x_i \leq 0 \\ g & \text{if } -0.1 < \Delta x_i \leq -0.05 \\ h & \text{if } -0.25 < \Delta x_i \leq -0.1 \\ i & \text{if } -0.5 < \Delta x_i \leq -0.25 \\ j & \text{if } \Delta x_i \leq -0.5 \end{cases} \tag{3.8}$$

Then, for each station, we cluster on the count of each bucket for each station (ie # y_i = a, b, c...). The counts in each bucket are scaled for use in k-means, which ensures equal weighting of the features. K-means requires a specified number of clusters; to select the number of clusters to use, we first cluster using different number of clusters, and examine the within-group sum of squares. Based on the results, it appears partitioning into 5 clusters will account for a significant amount of variation between stations.

We can now examine the different features across clusters. In particular, we examine the distribution of change in daily average pressure, difference between daily average and EWMA, and slope coefficient for a sliding window. There are substantial variations between clusters; in particular, clusters 2 and 5 have much lower volatility than clusters 1 and 3. Both the day-to-day pressure changes (Figure 3-12) and deviation from moving averages (Figure 3-13) have a significantly wider range for clusters 1 and 3. The longer term, 28-day slope coefficient (Figure 3-14) is more similar across the clusters, but clusters 1 and 3 still show wider ranges of that feature, indicating than longer ramps in pressure are more common in these clusters.

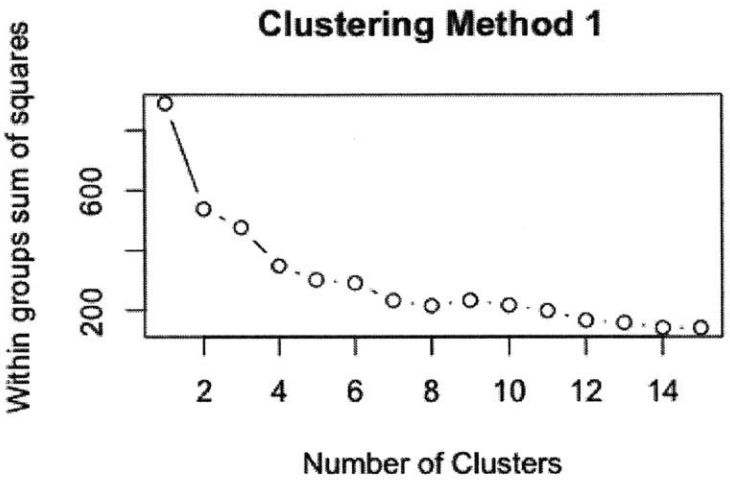


Figure 3-11: Within-group Sum of Squares Vs Number of Clusters

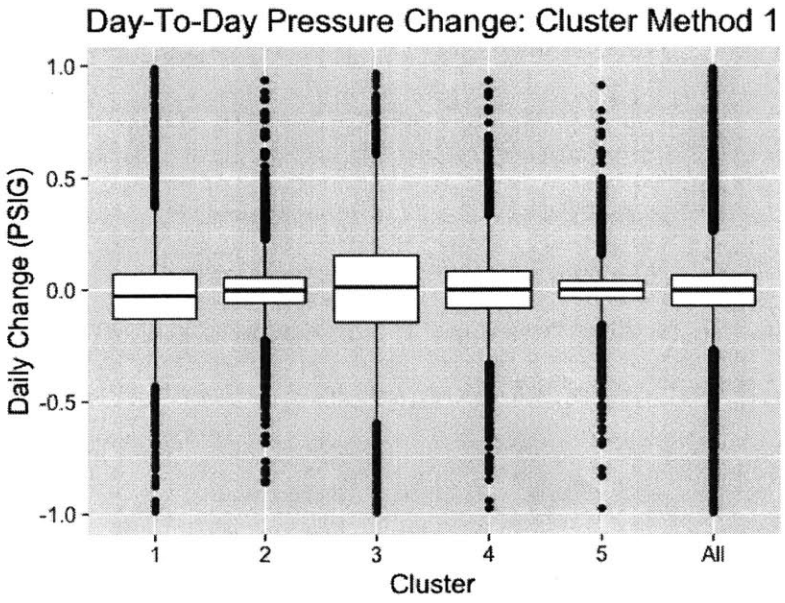


Figure 3-12: Box-plots of Daily Average Change Across Clusters: Method 1

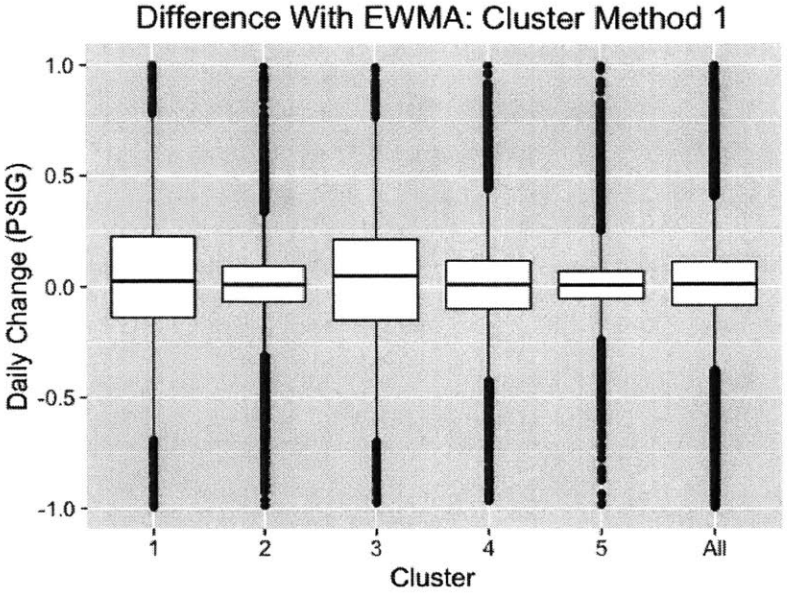


Figure 3-13: Box-plots of EWMA Difference Across Clusters: Method 1, $\lambda = 0.1$

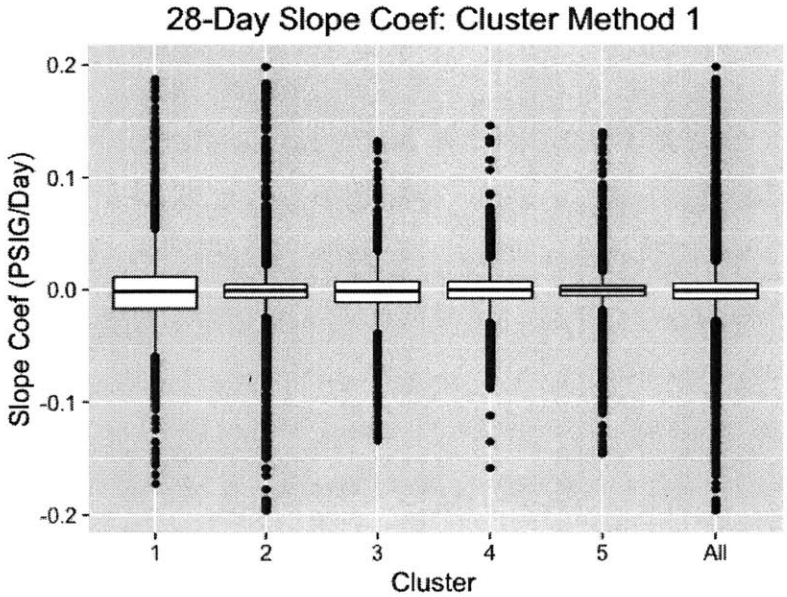


Figure 3-14: Box-plots of Local Regression Slope Coefficients Across Clusters: Method 1

Method 2: Piecewise Linear Features

While the previous method helps differentiate stations by volatility, we're interested in differentiating by certain pressure shapes. Stations can have similar distribution of changes in daily pressure, but have substantially different behavior. Here, we make use of the piecewise linear segmentation algorithm, creating features for clustering. For Clustering Method 2, we cluster using the following features for each station:

- Number of Segments with (Slope $<$ -0.05 and Length $>$ 14 days)
- Number of (Jumps between Segments $>$ 0.5)
- Standard Deviation of Slope of Segments, Weighted by Length of Segment
- Standard Deviation of Residuals of Piecewise Linear Model and Time-Series
- Median Difference between Weekend and Weekday Pressure

Again, these features are scaled before k-means clustering, so that each feature is equally weighted. Looking at the Within-group sum of squares for varying numbers of clusters, we again select 5 as the number of clusters. Here, clusters 1 and 3 again show larger ranges in day-to-day pressure changes (Figure 3-15) as well as deviations from longer term moving averages (Figure 3-16). Cluster 1 shows a significantly wider range in the longer-term, 28-day slope coefficient (Figure 3-17), indicating that substantial ramps in pressure occur more frequently in cluster 1 stations.

Comparing Cluster Results

In addition to looking at how features vary across clusters for the different clustering methods, we can measure which stations are grouped together. A statistic known as the Rand index is a measurement of the similarity between two clustering methods. The Rand index compares pairs of stations, measuring which ones are in the same cluster, versus different clusters. It can have a value between 0 (no similarity between clusters)and 1 (complete similarity between clusters). An adjusted Rand index adjusts for the probability that two stations

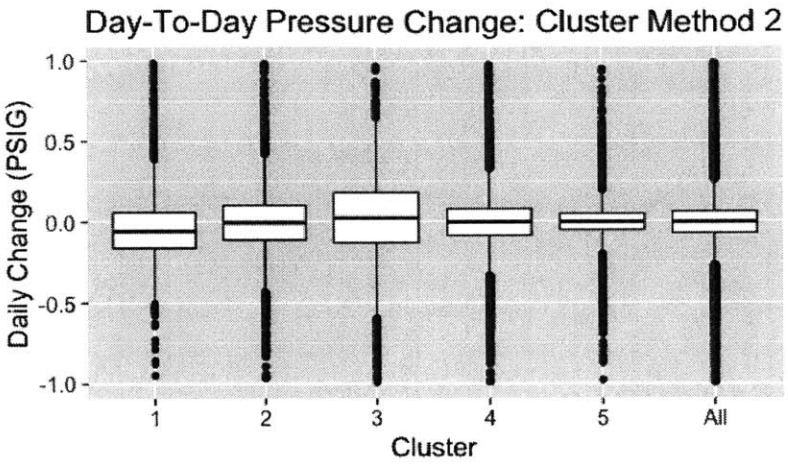


Figure 3-15: Box-plots of Daily Average Change Across Clusters: Method 2

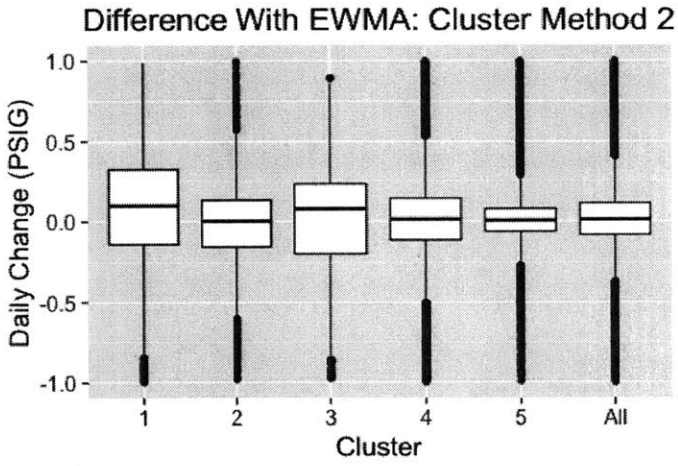


Figure 3-16: Box-plots of EWMA Difference Across Clusters: Method 2, $\lambda = 0.1$

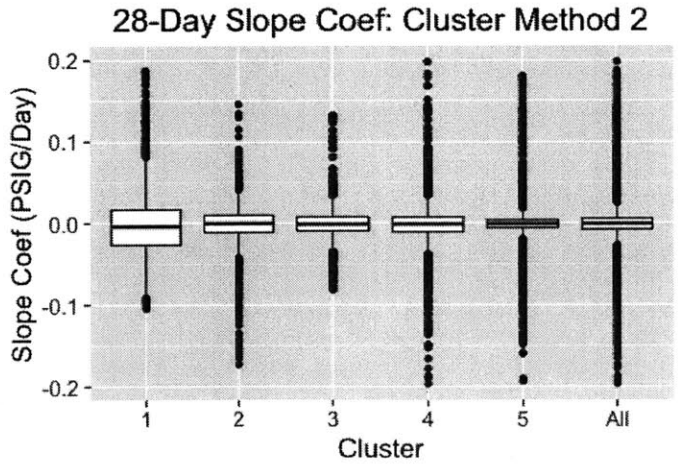


Figure 3-17: Box-plots of Local Regression Slope Coefficients Across Clusters: Method 2

may be in the same cluster by chance. We measure the adjusted Rand index for Method 1 and Method 2, as 0.277.

3.6 Detection Techniques

Using the features extracted and the clusters of stations, we use several methods for identifying anomalies in the downstream pressure time-series.

3.6.1 CUSUM-EWMA

CUSUM techniques are widely used for detection of changes in the mean of a process. In the case of regulator stations, however, the process mean is not a fixed number, so we estimate a local mean, in this case using an EWMA. Our CUSUM statistic is therefore given by:

$$S_t = max[0, S_{t-1} + (x_t - (\mu_t + K))] \tag{3.9}$$

With μ_t being an EWMA with smoothing parameter λ and K as a slack amount (also known as a reference value). Because we're concerned primarily with overpressure events, we focus on just the upper CUSUM limit. We can then set an upper control limit (UCL) on S_t. Breaching this alarm limit (Alarm = 1 if $S_t >$ UCL) results in resetting the CUSUM statistic S_t back to zero and alerting gas control. Figure 3-18 shows an example CUSUM chart, with alarms triggered at the red dashed lines. The parameters for this method are the smoothing parameter λ, the UCL on the CUSUM, and K, the amount of slack.

3.6.2 Local Regression

This method uses a sliding window to extract a subsequence at each time-series point. An OLS regression is fit to the subsequence, and the slope coefficient β is extracted as a feature. The slope coefficient is then used as the measurement in a Shewhart chart, with a given UCL. The alarm rule is then Alarm = 1 if $\beta >$ UCL. The parameters for this method are the upper control limit and size of the sliding window. An example of this detection method is shown in Figure 3-19, using a 7-day window.

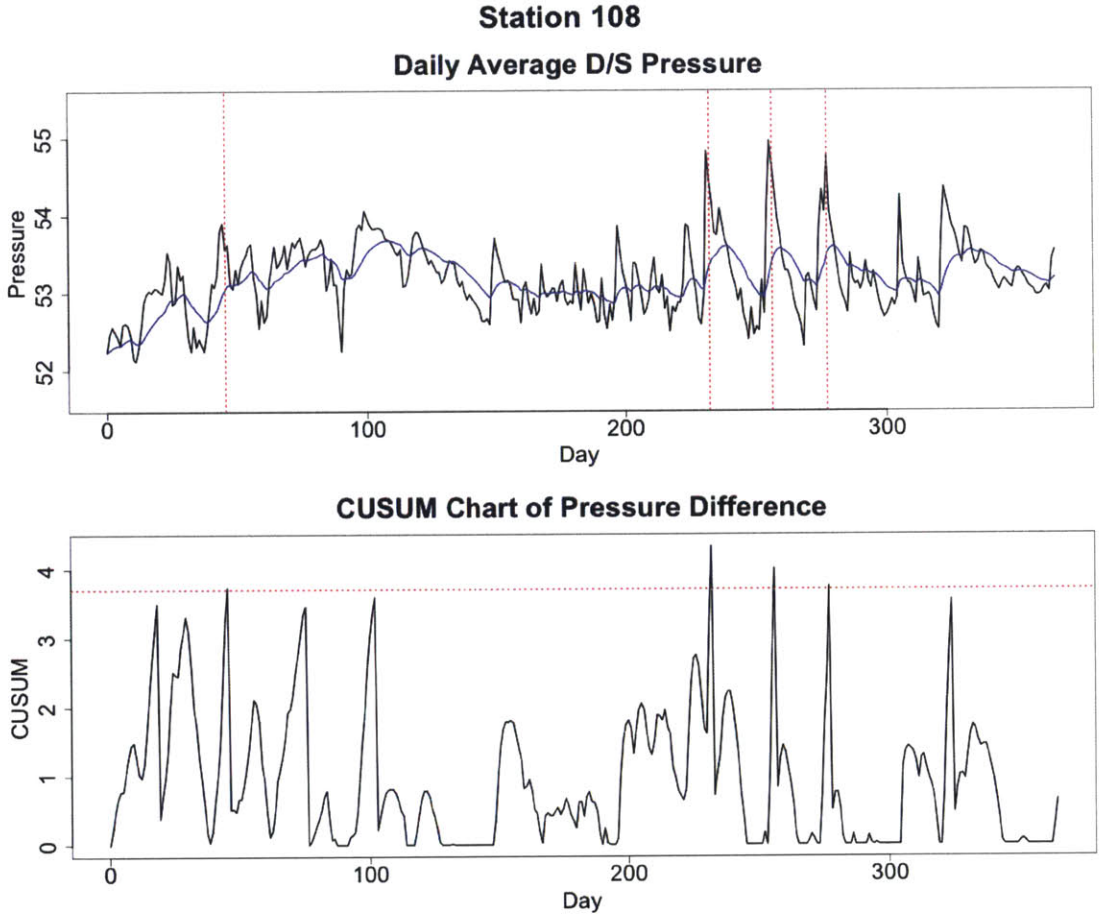


Figure 3-18: CUSUM Detection Technique for Station 108

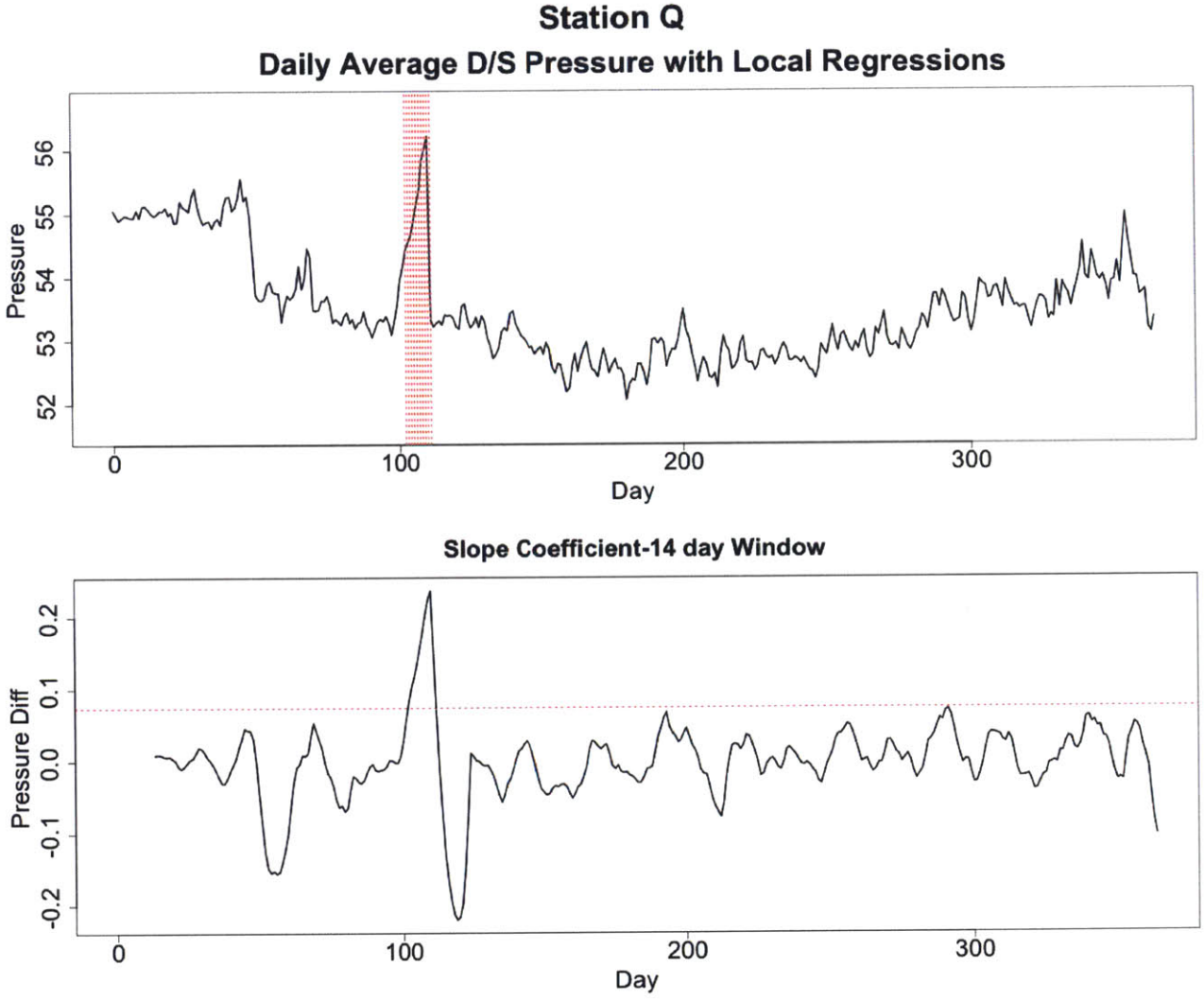


Figure 3-19: Local Regression Detection Technique for Station Q

3.6.3 Adaptive Window

The third detection method uses the sliding window piecewise linear segmentation algorithm (Algorithm 2) previously used for clustering. Given the last change-point, we extract a window and apply a linear regression to it. We can then monitor how the slope corresponds to the probability distribution of slope coefficients for windows of the same length (line 7 of Algorithm 2). This method requires keeping track of the last change-point in the time-series in order to build the current window.

The parameters for this are the error bound for the piecewise linear algorithm (which controls when a new segment begins), and the upper alarm bounds for slope coefficients. Analyzing 2014 data, we examine the quantiles of slope coefficients for varying segment lengths. We fit an exponential to the results, allowing us to parametrize the upper alarm bounds with the scalars, α and β which determine the UCL for a given window size W. Fitting an exponential to the 75th, 90th, and 95th percentile curves, we obtain values of β in the range of -0.3 to -0.2.

$$UCL(W) = \alpha e^{\beta W} \tag{3.10}$$

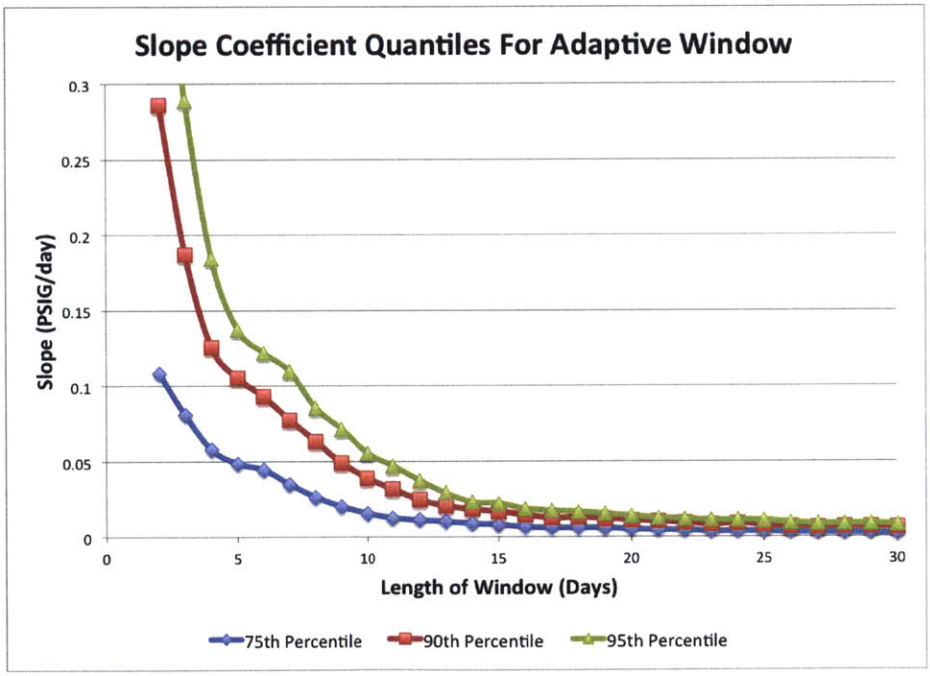


Figure 3-20: Slope Coefficient Quantiles By Segment Length

3.6.4 Conclusion

In summary, we have presented the steps in building the anomaly detection algorithms for natural gas regulator stations. First, data is collected, and then pre-processed for feature extraction. The distributions of these features give us insight into what constitutes an outlier or anomalous point. These features are then used for clustering, to group stations by characteristic behavior and better tailor the detection algorithms. Finally, the detection algorithms, themselves, use modifications to traditional statistical process control techniques to detect anomalies in these downstream pressure time-series. With the detection algorithms created, it's now critical to assess their performance for varying parameters and anomalies, to better understand the trade-off between detection of actual anomalies versus false alarms.

Chapter 4

Testing and Results

4.1 Overview of Testing and Results

After formulating our anomaly detection algorithms, we test their performance. We test on simulated data-sets, which were created by inserted synthetic anomalies into normal downstream pressure time-series data. We can then test the fraction of synthetic anomalies detected, versus the number of false alarms. We test the detection methods for multiple values of parameters and for different synthetic anomalies. We also test with and without clustering to better understand how detection methods can be tailored to a specific cluster of stations. Lastly, we analyze specific stations which are responsible for false alarms or display other anomalous pressure patterns.

4.2 Testing and Evaluation of Detection Methods

4.2.1 Simulation of Overpressure Events

Due to the low number of historical failure events on the time-scales of the detection methods, it was necessary to simulate overpressure events to estimate accuracy, sensitivity and specificity of the techniques developed. A ramp in pressure is added to each time-series at a given location. We tested using ramps of varying height and length. For ramp height, we first examined the difference between the Maximum Overpressure (MOP) level and the

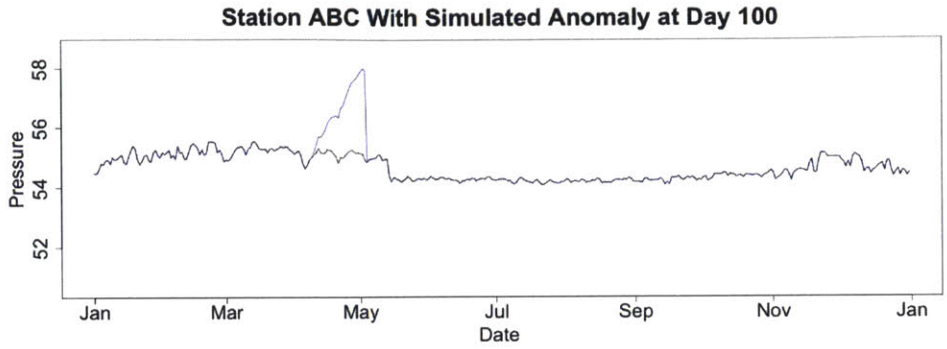


Figure 4-1: Simulated Spike in Time-Series

average pressure for each station. Across 92 stations in 2014, the minimum difference was 2.96 PSIG, and the median 5.35 PSIG. We therefore test ramps of 3 PSIG and 5 PSIG height. For ramp length, given the analysis of previous overpressure events, it's clear that ramps in pressure can occur on varying time-scales; we therefore examine ramps with length 6 and 24 days.

4.2.2 Detection Evaluation

We illustrate the performance of the anomaly detection techniques using receiver operating characteristic (ROC) curves. The vertical axis shows the true positive rate, also known the detection sensitivity. In our evaluation, the true positive rate is measured as the number of simulated ramps detected versus the total number of simulated ramps in the sample set. While a single ramp can trigger multiple sequential alarms, we only count whether the ramp has triggered at least one alarm. The horizontal axis shows the false positive rate. We define a false positive as an alarm triggered in a time period not containing a simulated ramp, and the false positive rate as (Number of False Alarms)/(Number of Days Not Containing a Simulated Ramp). Along each curve, the alarm bounds (the UCLs) are varied, so we can see how the detector performs for different alarm limits. The resulting curves show the trade-off between high detection rates and false positives.

For evaluation, we look at 89 stations with 2014 and 2015 data. The clusters and algorithm parameters are calibrated using 2014 data. We test both in-sample on 2014 time-series, as well as out of sample on the 2015 time-series.

CUSUM-EWMA Performance

We first examine the performance of CUSUM-EWMA method. Testing against the 89 stations with 2014 data, we examine the performance of the detection algorithm for different values of the EWMA smoothing parameter, λ (Figure 4-2). Interestingly, the effect of λ varies depending on the length of ramp being detected. Using the largest $\lambda = 0.2$, which weights recent observations more heavily in the EWMA, detects rapid ramps the best, but its performance is the worst for the longer 24-day ramp. The reverse isn't true for the smallest λ; it's possible the types of false positives being generated are responsible for this. Many of the false positives are due to set-point changes, which result in a rapid step-function in the pressure time-series; a smaller λ (slower varying EWMA) might have better detection of ramps, but may produce more false alarms after a set-point change since the EWMA doesn't adjust as quickly afterwards.

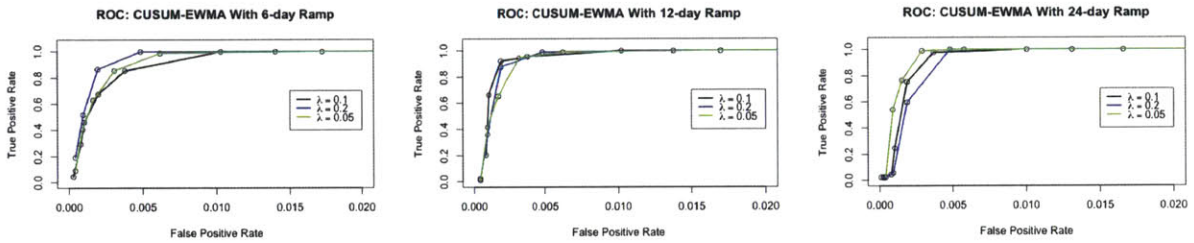


Figure 4-2: Performance of CUSUM-EWMA on 2014 Dataset, Varying λ, K = 0

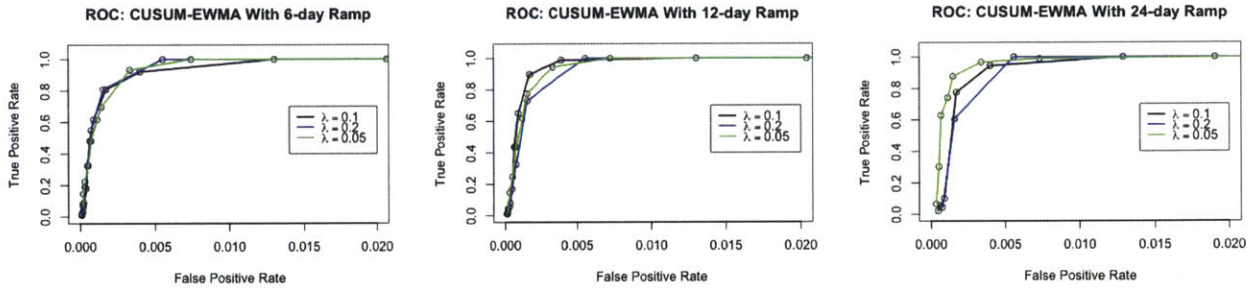


Figure 4-3: Performance of CUSUM-EWMA on 2015 Dataset, Varying λ, K = 0

We also examine the effect of varying the slack parameter, K, of the CUSUM method (Figure 4-3). Again, there appear to be some non-linear effects in how K influences the detection efficiency. Positive values of K perform well in detecting 6-day and 24-day ramps, but perform worse with the 12-day ramp.

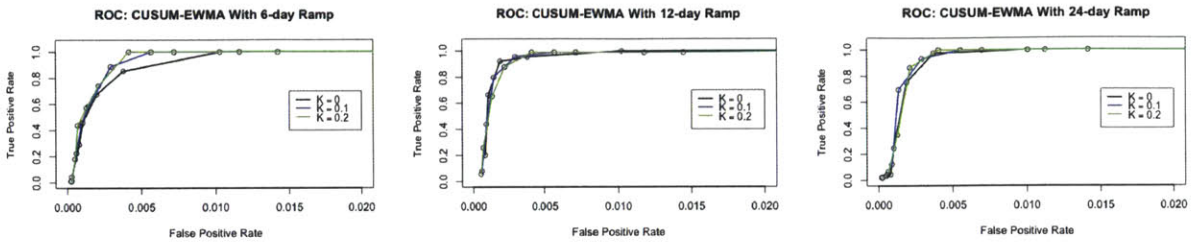


Figure 4-4: Performance of CUSUM-EWMA on 2014 Dataset, Varying K, $\lambda = 0.1$

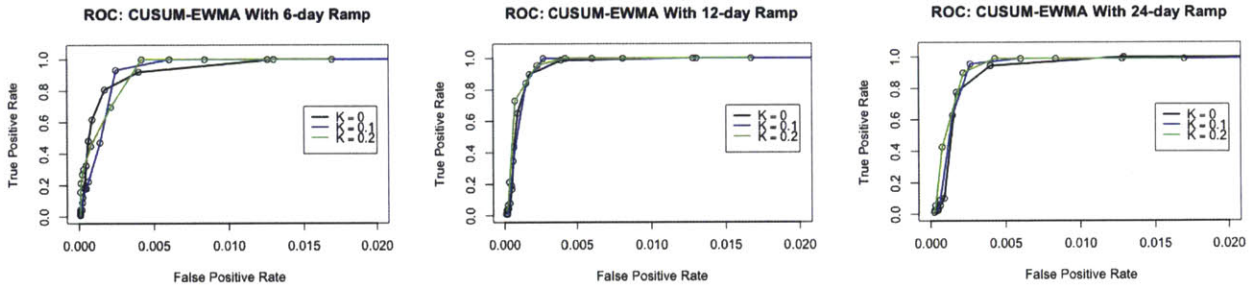


Figure 4-5: Performance of CUSUM-EWMA on 2015 Dataset, Varying K, $\lambda = 0.1$

Overall, the CUSUM-EWMA detection method allows nearly 100% detection of the simulated ramps at less than 0.5% false positive rate. However, given there are about 250 distribution regulator stations with RTUs, this still would result in about 1 false alarm per day. Given the number of personnel and capabilities of PG&E's Gas Control Center, this is probably an acceptable rate. However, further installation of SCADA systems, would result in correspondingly more false alarms.

Local Regression Performance

We now look at performance of the Sliding Window, Local Regression method. Here, the key parameter is the length of sliding window, W. As shown in Figure 4-4, the window size effectively determines the type of anomaly that can be most effectively detected. The 7-day window is the most effective on the 6-day ramp, the 14-day window is most effective on the 12-day ramp, and the 28-day window is most effective on the 24-day ramp. In addition, performance drops off significantly for window sizes different than the corresponding anomaly length.

Overall, the local regression detection method has generally worse performance than the

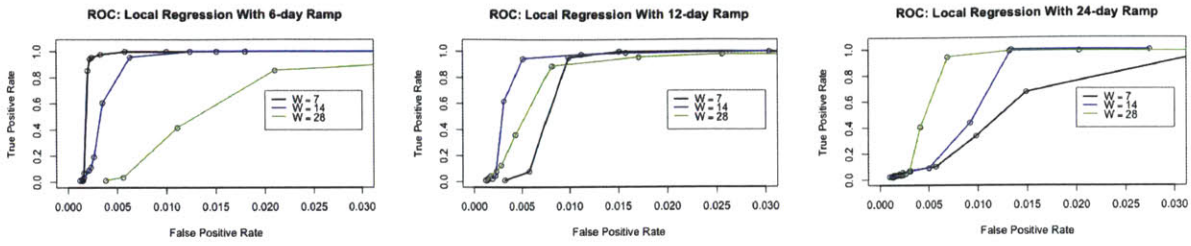


Figure 4-6: Performance of Local Regression Method on 2014 Dataset, Varying W

CUSUM-EWMA, unless the window size precisely matches the length of the anomalous ramp.

Adaptive Window Performance

We also tested the adaptive sliding window detection method on the simulated anomaly data-sets. For constructing the ROC curves, we varied the α parameter of the UCL function ($\text{UCL(W)} = \alpha e^{\beta W}$), with a fixed β for each curve. Overall, the performance of the algorithm is similar to the CUSUM-EWMA method, achieving nearly 100% of simulated anomalies detected, with less than 0.005 false alarms per station-day. Analyzing the effect of varying β, we see that a steeper decreasing exponential (β = -0.3) performs better; this effectively increases the UCL for smaller windows, which in turn likely reduces false alarms due to short-term volatility. We did not investigate varying the error bound for starting a new piecewise linear segment; a new segment is created when the absolute maximum residual of the segment is greater than 0.2 PSIG.

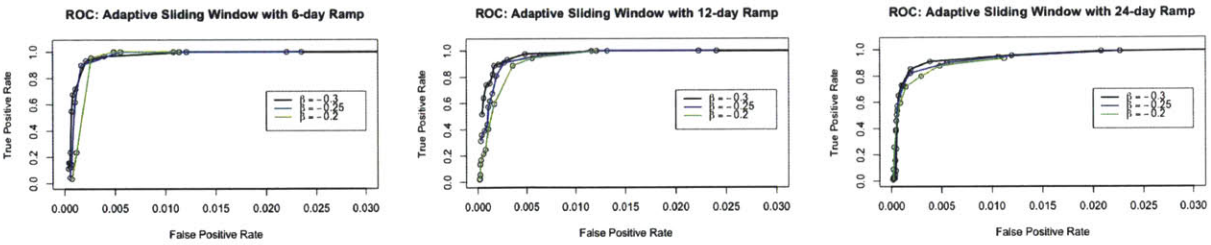


Figure 4-7: Performance of Adaptive Window Method on 2014 Dataset, Varying β

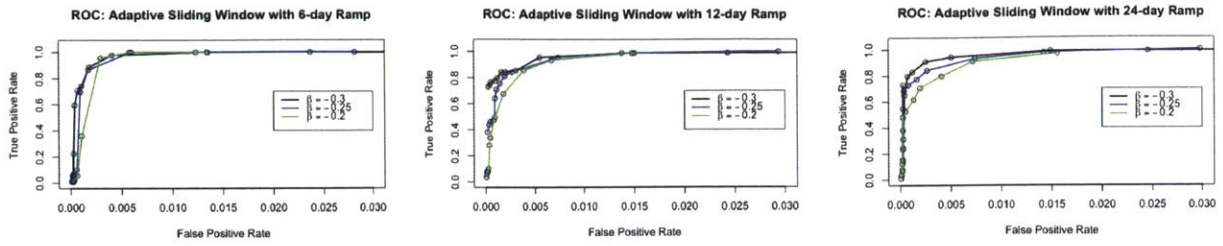


Figure 4-8: Performance of Adaptive Window Method on 2015 Dataset, Varying β

Performance By Cluster

We can also look at how the detection methods perform against the individual clusters we created previously. Looking at our first clustering method, using just the distribution of changes in daily average pressure, we see cluster 1, which had more volatility than the other clusters, has significantly worse performance in the CUSUM, Local Regression and Adaptive Window methods. Clusters 2, 4, and 5, on the other hand, perform well, and our previous analysis showed those 3 clusters as having lower volatility.

Looking at our second clustering method, using features extracted from the piecewise linear segmentation, we see cluster 1, which also had substantial volatility, causing poor performance in the three detection algorithms. Interestingly, cluster 5 has excellent performance under all three algorithms, and is the largest cluster, as well, making up 50 of the 89 stations. This highlights the ability for time-series clustering to identify stations very amenable to anomaly detection methods, while partitioning off the poor-performing stations. For implementing these algorithms in PG&E Gas Operations, the stations that perform well under these algorithms can be setup for monitoring first, while stations in clusters that perform poorly can further investigated to understand the underlying causes of their volatility.

4.3 Identification and Classification of Stations

The previous clustering of stations also provides the opportunities for asset management. Ideally, these station clusters would be related to station characteristics, including downstream demand, equipment type and recent maintenance information. Unfortunately, the bulk of this data was not available for the purposes of this study. We examine some recur-

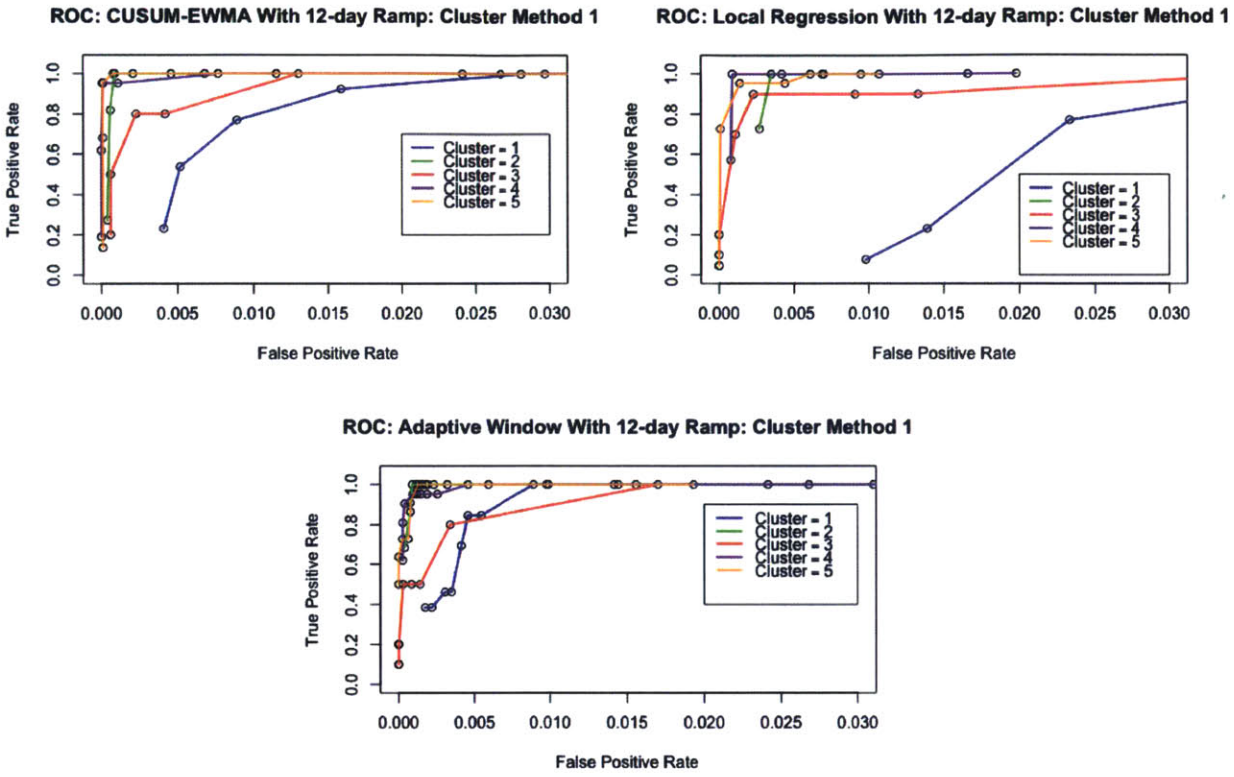


Figure 4-9: Performance of Detection Methods on 2014 Dataset, By Cluster For Cluster Method 1

ring patterns of interest for asset management that are worthy of further study. From an anomaly detection standpoint, we examine these stations to gain a better understanding of the particular pressure patterns causing false alarms, and why the detection efficiency varied across clusters. We analyze the false alarms (red dashed lines) from stations generated by the CUSUM-EWMA method, with UCL = 3.7, K = 0, and $\lambda = 0.1$.

Saw-tooth Pattern

One particular pattern seen in several stations is a distinct saw-tooth, with daily average pressure rapidly increasing over the course of a day or two, and then ramping slowly downward over the span of weeks or months. These patterns have been identified both through visual inspection and the analysis of the piecewise linear segments. In particular, one specific area has several regulators with this pattern, most of them in a single hydraulically independent system. Other examples exist for stations in other divisions. In particular,

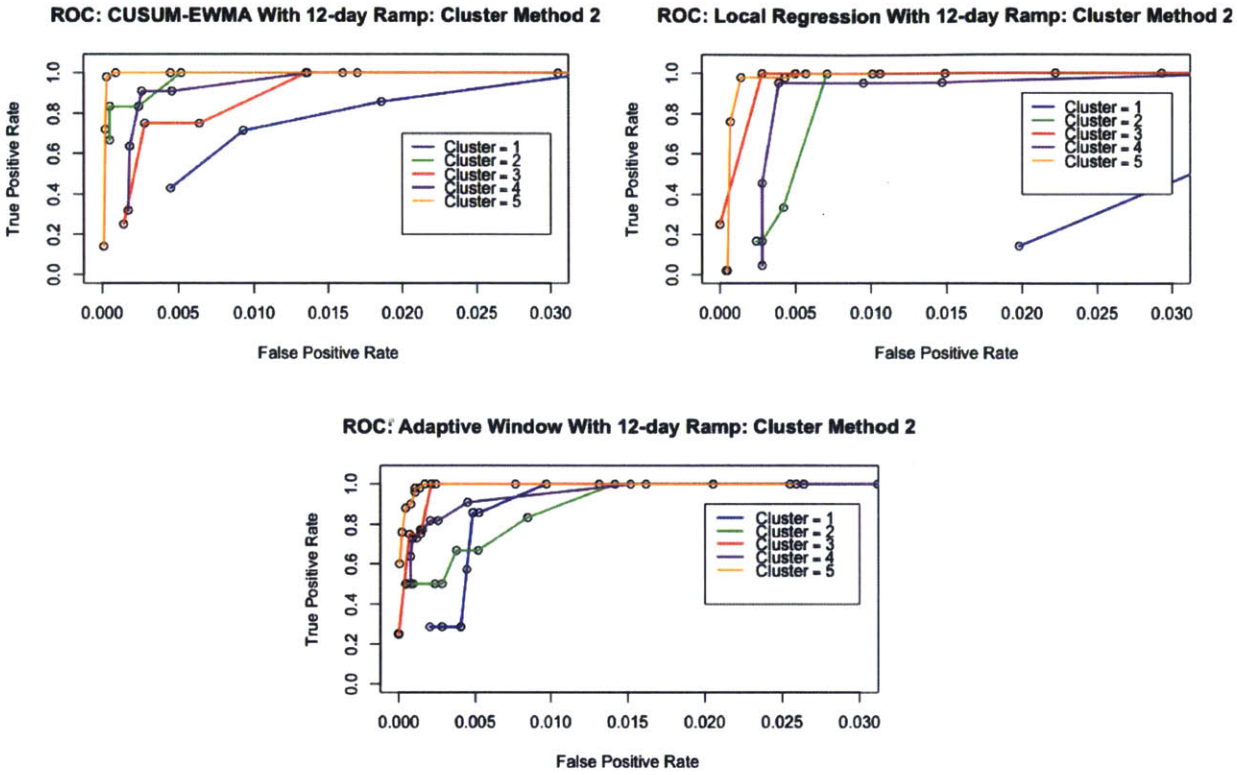


Figure 4-10: Performance of Detection Methods on 2014 Dataset, By Cluster For Cluster Method 2

the three stations shown, are all in cluster 1 in both clustering methods. These sawtooths therefore appear to be partly responsible for the poor false alarm rate of cluster 1 under both clustering methods. It remains unclear as to the overall cause of this pattern and whether it constitutes a potential safety risk or sign of equipment degradation.

Set-point Changes

Increases in the set-point of the regulator can trigger false alarms in the detection algorithms. In particular, Station 50 produced a large number of false alarms due to a temporary, significant decline and then increase in the set-point of the regulator. Ideally, the algorithms should be linked to a table of all maintenance events (including set-point changes); unfortunately, much of this data was not available for the purposes of this study. Automated accounting of set-point changes would facilitate reducing false alarms.

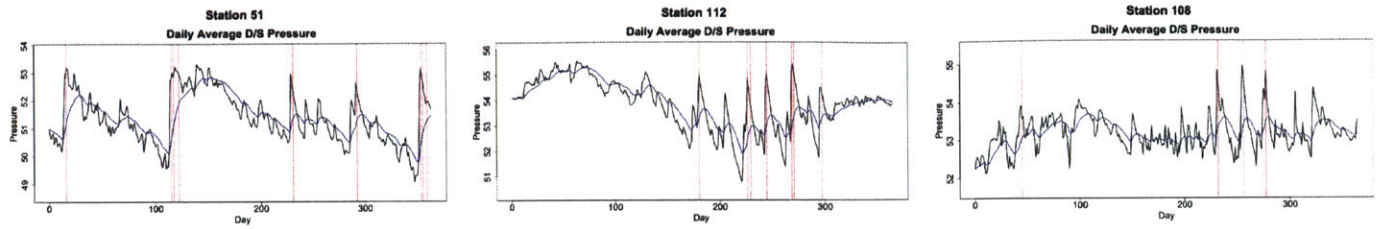


Figure 4-11: Sawtooth Pattern

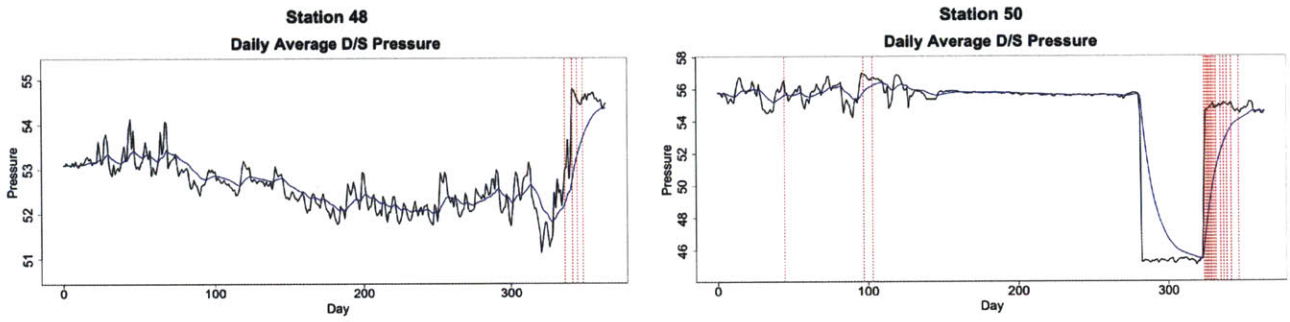


Figure 4-12: Set-point Changes

Winter Droop

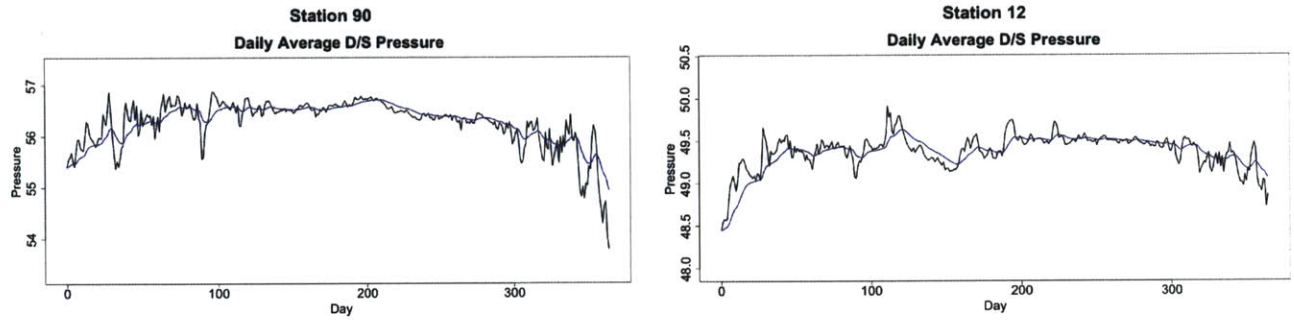


Figure 4-13: Winter Droop

Several stations have substantial drops in pressure during winter months. This is consistent with the physics of the regulators that experience pressure drops when flow approaches the maximum flow capacity of the regulator. These stations have been presented to the Gas Planning department for further analysis of downstream demand. Identification of stations with winter droop can help inform investment decisions for stations to upgrade.

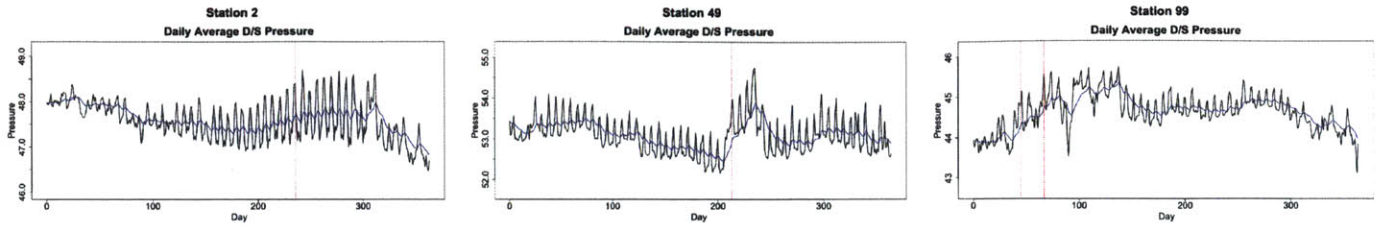


Figure 4-14: Weekend-Weekday Swings

Weekend-Weekday Swings

As mentioned in Section 3.2, volatility between weekend and weekday is significant for several stations, particularly those in the Fresno area. While this volatility doesn't contribute to a significant number of false alarms, in conjunction with other behavior (such as winter droop in Station 99), it can create sharp ramps in pressure that would trigger false alarms. Moreover, the substantial swings in pressure do indicate that these stations are liable to have different characteristic behavior. This illustrates the importance of measuring demand downstream of the regulator, as a way of better predicting regulator pressure patterns. Increased implementation of smart meters and use of smart meter data for Gas Operations would help resolve this.

Other Anomalous Pressure Patterns

Additional stations show highly variable and anomalous pressure patterns. Better understanding of the underlying causes and dynamics of these stations is critical for improving modeling of the pressure behavior of the regulators.

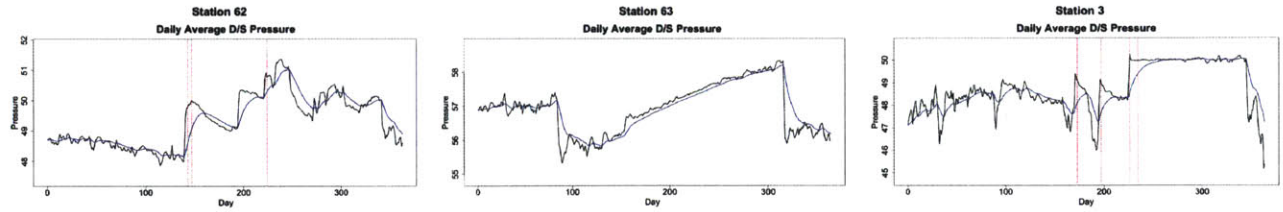


Figure 4-15: Other Anomalous Pressure Patterns

Chapter 5

Recommendations and Conclusions

5.1 Recommendations for Data Handling and Sensor Installation

Overall, given the data challenges faced with this project, we present several recommendations for better data handling.

Set-point Data

Accurate set-point data is critical for predicting pressure patterns in a regulator station. As shown, many false positives are due to set-point changes, so having efficient procedures for distinguishing these from actual anomalies is important. While set-point data currently exists in the SAP system in PG&E, improvements can be made to increase the effectiveness of this data. For example, every recorded set-point change features two numbers; some maintenance divisions record them as the "as-found" and "as-left" pressures, while others record them as the set pressures of the monitor and regulator. Better clarification and standardization around how set-point values are entered into SAP will improve data quality. Furthermore, set-point changes related to winter weather and temporary maintenance are not recorded in SAP. A combined database that integrates all set-point data together would be most useful for predictive analytics. A table showing the station, date, equipment modified (such as monitor vs regulator), set-point of the equipment and the reason for the set-point change

would allow for much easier identification of potential false alarms, improving the sensitivity of the detection algorithms.

Downstream Demand Data

Likewise, improvements can be made to downstream demand data and its use in predictive analytics. This data is currently fragmented in a multitude of different files. Centralizing data in relational databases such as Oracle or MySQL would provide easier access to this information. As shown, downstream demand is important in predicting and estimating pressure changes, both on longer time-scales and over the course of the day. Initially, a centralized database could show the percent residential versus percent industrial consumers in a given area. Further improvements could include a listing of the largest industrial gas consumers in an area and their typical consumption patterns. If downstream demand meter data is unavailable, better temperature data can help bridge the gap. Temperature is highly correlated to flow rate downstream, which in turn governs the downstream pressure. Combining PG&E temperature station data with other third-party temperature data could provide a better view into the temperature changes in a given area, and better predict gas consumption.

Equipment and Maintenance Data

One of the primary reasons for clustering the different station time-series was to group stations with similar operating characteristics and behavior. This was partly due to the fact that equipment and maintenance data differentiating regulator stations was not easily available for the purposes of the study. Data regarding equipment model, manufacturer, age, capacity and so forth would help in categorizing stations. It could also be critical in explaining some of the anomalous pressure patterns that were detected. Ideally, we would like to relate the clusters of time-series to physical characteristics of the regulator stations. For instance, all stations with regulator model X installed might show specific pressure fluctuations. Better understanding of the dynamics of regulator stations and the specific equipment installed will aid in differentiating false alarms from true anomalies.

Likewise, maintenance data can give insight into which stations suffer from sulfur prob-

lems, or other observations about the performance of the equipment. A standardized, centralized system for recording equipment and maintenance information is critical for leveraging analytics in predictive maintenance.

5.2 Ideal Model for Anomaly Detection

With accurate set-point, flow data and equipment data substantial improvements can be made for anomaly detection and asset management. The overall approach is to use additional data to create a predictive model for the downstream pressure of the time-series. We used an EWMA as the process mean in the CUSUM method, primarily because we were unable to assess the target mean of the regulator; cumulative deviations from the EWMA then triggered alarms. Rather than an EWMA, we could use additional data sources to create a prediction for the downstream pressure, and measure deviations from the prediction, using traditional SPC techniques.

Generally, the downstream pressure of a regulator is going to be a function of the set-point of the regulator, the flow through the regulator vs the capacity, and the upstream pressure. Accurate, standardized record-keeping by maintenance crews can provide the set-point data. Capacity of the regulator is governed by the equipment type and sizing. Upstream pressure can be measured with a sensor upstream of the regulator. Flow through the regulator can be measured with a flow sensor, currently used in a handful of regulators. However, flow sensor are fairly expensive; useful proxies for flow can be created by analyzing gas consumption downstream of the regulator. This consumption, in turn, is driven by ambient temperature and the individual consumption of large consumers such as industrial plants. For instance, we can compare the downstream pressure of a regulator to the "Heating Degree Day" defined as max(0, 62 - Daily Average Temp); as shown in Figure 5-1, several significant spikes or drops in pressure correspond to substantial swings in temperature.

For further illustration, consider a simple predictive model (using simulated data) for downstream pressure X(t), based on the set-point, S(t) of the regulator and flow F(t). The flow results in variations in downstream pressure, Z(t), given by Figure 5-2; downstream pressure rises during low flow, and falls around maximum flow. We define the resulting

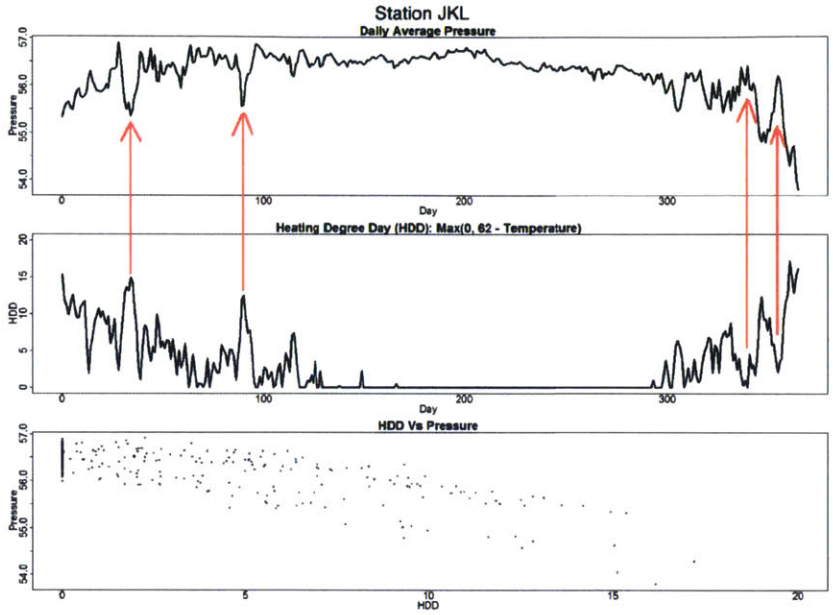


Figure 5-1: Downstream Pressure vs HDD

downstream pressure as: $X(t) = S(t) + Z(t)$. Figure 5-3 shows example time-series of set-points and flows, along with the predicted downstream pressure. In this example, the "actual" downstream pressure is generated by adding random noise and simulated anomalies into the predicted downstream pressure. Then, by analyzing the residuals of the (actual - prediction), we can identify potential anomalies (several were inserted into this simulated data-set). The overall idea is to remove as much noise and volatility from the downstream pressure time-series by finding explanatory variables (such as flow) for the noise. Simply looking at the downstream pressure alone wouldn't have indicated the anomalies due to the significant volatility; subtracting out this volatility using other variables and analyzing the residuals allows for this targeted anomaly detection.

5.3 Applications in Other Areas

The techniques used here for anomaly detection in natural gas regulators are applicable in other industrial areas, as well. Pressure regulators are used in a substantial number of other fields involving flow networks. Compressed air lines with regulators are used for many pneumatic tools and industrial processes. Likewise, steam pipes with regulators are used

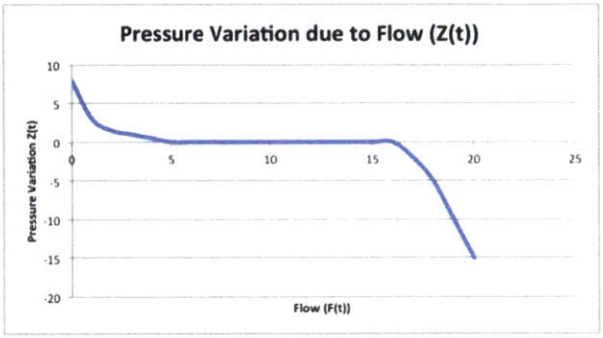


Figure 5-2: Example of Flow Versus Downstream Pressure Variations

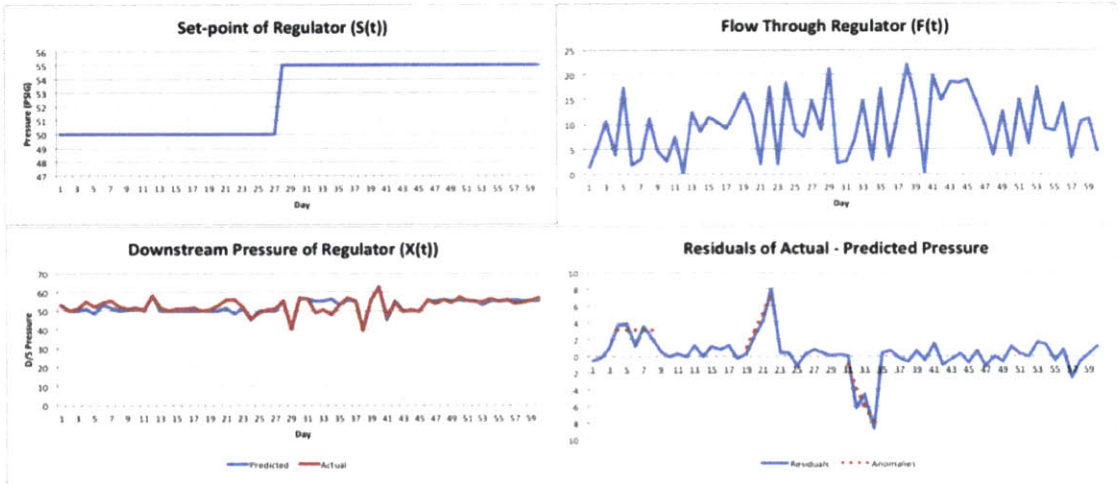


Figure 5-3: Illustration of Ideal Model

in power plants, refineries, paper mills and other manufacturing plants. Liquid pipes with regulators distribute water, oil and other chemicals. Natural gas regulators are also used in industrial settings such as at refineries and power plants. Early detection of potential overpressure events is critical in all of these areas. For PG&E, these time-series anomaly detection methods can be applied to transmission gas pipelines and compressor stations, and also be of use in detecting mechanical faults in electrical generators and turbines.

5.4 Conclusion

We have shown that a variety of anomaly detection techniques can be used to identify potential overpressure events in natural gas regulator stations. Feature extraction of time-series remains a critical step, allowing identification of relevant phenomena, and allowing clustering of stations with similar operating characteristics.

Further improvements to regulator station anomaly detection will require a better understanding of regulator behavior and improved collection of maintenance, equipment and consumer demand data. Better integration of these data sources is critical to creating better predictive models and identifying deviations from predictions.

Bibliography

[1] Charu C. Aggarwal. *Data Mining.* Springer International Publishing, 2015.

[2] Daniele Apiletti, Elena Baralis, Giulia Bruno, and Tania Cerquitelli. Real-Time Analysis of Physiological Data to Support Medical Applications. *IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE*, 13(3):313–321, MAY 2009.

[3] Sabyasachi Basu and Martin Meckesheimer. Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems*, 11(2):137–154, 2006.

[4] E. Peter Carden and James M.W. Brownjohn. {ARMA} modelled time-series classification for structural health monitoring of civil infrastructure. *Mechanical Systems and Signal Processing*, 22(2):295 – 314, 2008.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly Detection: A Survey. *ACM COMPUTING SURVEYS*, 41(3), 2009.

[6] Mooi Choo Chuah and Fen Fu. ECG anomaly detection via time series analysis. In Thulasiraman, P and He, X and Xu, TL and Denko, MK and Thulasiram, RK and Yang, LT, editor, *Frontiers of High Performance Computing and Networking - ISPA 2007 Workshops*, volume 4743 of *LECTURE NOTES IN COMPUTER SCIENCE*, pages 123–135, 2007. 5th International Symposium on Parallel and Distributed Processing and Applications/ISPA 2007 International Workshops, Niagara Falls, CANADA, AUG 29-31, 2007.

[7] RA Davis, TCM Lee, and GA Rodriguez-Yam. Structural break estimation for nonstationary time series models. *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, 101(473):223–239, MAR 2006.

[8] George Georgoulas, Petros Karvelis, Theodoros Loutas, and Chrysostomos D. Stylios. Rolling element bearings diagnostics using the symbolic aggregate approximation. *Mechanical Systems and Signal Processing*, 6061:229 – 242, 2015.

[9] M. Gupta, Jing Gao, C.C. Aggarwal, and Jiawei Han. Outlier detection for temporal data: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 26(9):2250–2267, Sept 2014.

[10] Z. Hameed, Y. S. Hong, Y. M. Cho, S. H. Ahn, and C. K. Song. Condition monitoring and fault detection of wind turbines and related algorithms: A review. *RENEWABLE & SUSTAINABLE ENERGY REVIEWS*, 13(1):1–39, JAN 2009.

[11] D Han and FG Tsung. A reference-free Cuscore chart for dynamic mean change detection and a unified framework for charting performance comparison. *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, 101(473):368–386, MAR 2006.

[12] DM Hawkins, PH Qiu, and CW Kang. The changepoint model for statistical process control. *JOURNAL OF QUALITY TECHNOLOGY*, 35(4):355–366, OCT 2003.

[13] VictoriaJ. Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[14] Andrew K.S. Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7):1483 – 1510, 2006.

[15] Wei Jiang, Lianjie Shu, and Daniel W. Apley. Adaptive CUSUM procedures with EWMA-based shift estimators. *IIE TRANSACTIONS*, 40(10):992–1003, 2008.

[16] E Keogh, S Chu, D Hart, and M Pazzani. An Online algorithm for segmenting time series. In Cercone, N and Lin, TY and Wi, XD, editor, *2001 IEEE INTERNATIONAL CONFERENCE ON DATA MINING, PROCEEDINGS*, pages 289–296. IEEE Comp Soc, TCPAMI; IEEE Comp Soc, TFVI; Insightful Corp; Microsoft Res; NARAX Inc; Springer Verlag, New York; StatSoft Inc, 2001. IEEE International Conference on Data Mining, SAN JOSE, CA, NOV 29-DEC 02, 2001.

[17] Xin Li and Zhi-Hong Deng. Mining frequent patterns from network flows for monitoring network. *EXPERT SYSTEMS WITH APPLICATIONS*, 37(12):8850–8860, DEC 2010.

[18] T. Warren Liao. Clustering of time series dataa survey. *Pattern Recognition*, 38(11):1857 – 1874, 2005.

[19] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing SAX: a novel symbolic representation of time series. *DATA MINING AND KNOWLEDGE DISCOVERY*, 15(2):107–144, OCT 2007.

[20] Jessica Lin and Yuan Li. Finding Structural Similarity in Time Series Data Using Bag-of-Patterns Representation. In Winslett, M, editor, *SCIENTIFIC AND STATISTICAL DATABASE MANAGEMENT, PROCEEDINGS*, volume 5566 of *Lecture Notes in Computer Science*, pages 461–477. LATG; Diamond Data Syst; Sun Microsyst; NOVACES; Univ New Orleans, 2009. 21st International Conference on Scientific and Statistical Database Management, New Orleans, LA, JUN 02-04, 2009.

[21] Douglas Montgomery. *Statistical Quality Control 7th Edition.*

[22] S Nandi, HA Toliyat, and XD Li. Condition monitoring and fault diagnosis of electrical motors - A review. *IEEE TRANSACTIONS ON ENERGY CONVERSION*, 20(4):719–729, DEC 2005.

[23] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215 – 249, 2014.

[24] Venkatesh Rajagopalan and Asok Ray. Symbolic time series analysis via wavelet-based partitioning. *SIGNAL PROCESSING*, 86(11):3309–3320, NOV 2006.

[25] Jaxk Reeves, Jien Chen, Xiaolan L. Wang, Robert Lund, and Qiqi Lu. A review and comparison of changepoint detection techniques for climate data. *JOURNAL OF APPLIED METEOROLOGY AND CLIMATOLOGY*, 46(6):900–915, JUN 2007.

[26] SN Rodionov. A sequential algorithm for testing climate regime shifts. *GEOPHYSICAL RESEARCH LETTERS*, 31(9), MAY 6 2004.

[27] GC Runger and MC Testik. Control charts for monitoring fault signatures: Cuscore versus GLR. *QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL*, 19(4):387–396, JUL-AUG 2003. Conference of the European-Network-of-Business-and-Industrial-Statisticians, RIMINI, ITALY, SEP, 2002.

[28] Galit Shmueli and Howard Burkom. Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. *TECHNOMETRICS*, 52(1):39–51, FEB 2010.

[29] LJ Shu, DW Apley, and F Tsung. Autocorrelated process monitoring using triggered Cuscore charts. *QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL*, 18(5):411–421, SEP-OCT 2002.

[30] Haining Wang, Danlu Zhang, and Kang G. Shin. Change-point monitoring for the detection of dos attacks. *IEEE Trans. Dependable Secur. Comput.*, 1(4):193–208, October 2004.

[31] Xiaozhe Wang, Kate Smith, and Rob Hyndman. Characteristic-based clustering for time series data. *DATA MINING AND KNOWLEDGE DISCOVERY*, 13(3):335–364, NOV 2006.