Industrial Data Analysis and Prediction Modeling

by

Ruben Araya

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering
University of Alberta

Abstract

Multiple studies have reported results of activities focused on development of different algorithms for prognosis of failures of apparatuses and machines. Failure prediction allows to schedule maintenance activities, increase productivity, and decrease inventory of spare parts. Construction of prediction models requires multiple procedures and thorough analysis of available data. For that, processes of Data Mining and Machine Learning can be applied. Data Mining provides necessary tools for storing, refining data, and analyzing data. Machine learning, on the other hand, includes a variety of approaches and techniques for finding patterns in data and building data models.

This study addresses data-driven analysis of two industrial problems. In the case of the first one, we analyze a degree of "weariness" of suspensions in haul trucks. Here, we determine usage of a suspension via generating and integrating features representing struts' pressures. The second problem concerns prediction of outages in power systems. Feature selection is performed, and different prediction models are built. Additionally, we look at graph-based data representation and visualization of data with Neo4j database.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1. Motivation

Prediction of events is an important topic of academic and industrial research. Anticipation provides multiple benefits for long-term strategies. If it is possible to forecast a certain event, we can adopt proactive measures, and improve future states of a system. Currently, this topic is being addressed by different industries for betterment of procedures, preparedness of new conditions, and improvement of final products. In this aspect, maintenance procedures draw a lot of attention due to a novel and extensive usage of sensors (Saiied and Moe, 2017) and their integration with IoT (Internet of Things) devices (Yeon and Jun, 2016). Sensors, taking different types of reading from machines/systems, generate a lot of data that allows us to determine current states of machines/systems and potentially, the prognosis of any failures that might arise. Nevertheless, several steps and some of them quite complex are required in order to obtain any type of prognosis. Firstly, data has to be captured, stored, cleaned, integrated with other sources, and then analyzed. Secondly, a prediction model needs to be developed with the ability to generate prognosis about machine/system faults. In most cases, this involves such processes as feature selection and the design of an architecture of a prediction model. Selection of the most suitable procedures and technologies depends entirely on the context and goals set up for each stage. Therefore, development of prediction models is rarely straightforward, given that there is not a well establish procedure that defines how to achieve the desired results.

1.2. Thesis Goal and Contributions

Given the countless ways to solve different event-prediction tasks, our study focuses on the analysis of two different industrial problems by applying suitable approaches based on the context. The first problem addresses predicting failures of a suspension in haul trucks. For that, new features have been created using the measures of the strut's pressure. Interesting results have been obtained. However, it has become evident that construction of a prediction model requires more suspension data. The second problem focuses on building a model capable of predicting an outage in a power system. Prediction models have been constructed using features

representing different aspects of a power system itself and features describing weather. In this case, suitable models have been developed based on the available data. Also, it has been evident that more information and more integration is required to construct models in order to improve their performance.

The contributions of this thesis are as follows:

1. We propose a reference-like approach for dealing with different data related analytics: for time series in the case of our first problem – haul trucks; and data integration issues for our second problem – outages in power systems.
2. We propose an application of different technologies for storage, integration, visualization and manipulation of data.
3. For the suspension's failure prediction task: we provide a way to measure a degree of usage of suspension, and visually detect a suspension's failure.
4. For the power system outage prediction task: we provide a methodology for construing a model able to predict an outage using the available data.

1.3. Thesis Outline

This thesis is divided into five chapters. Chapter 2 describes work focused on event prediction procedures applied in industrial settings. Chapter 3 provides background on data mining, machine learning and time series analysis. Details related to analysis of data related to the suspension's failures in haul trucks are presented in Chapter 4. Chapter 5 describes the analysis done on power system outage data and construction of prediction models able to predict specific types of outages. Finally, Chapter 6 presents the conclusion of the thesis.

CHAPTER 2

Related Work

As technology progresses, industry becomes more competitive. Now, companies not only put their focus on the final product/service but also, they try to add *intelligence* to their processes, optimizing them. With this addition, the costs are reduced, the productivity increases, and the final output becomes less cumbersome to produce, unveiling new venues to explore; the available resources may be used to generate new business opportunities. In the light of this necessity, the application of machine learning plays a main role, granting this new added value, *intelligence.* In the pursuit of optimization, there are numerous applications that can be mentioned; one of the most common ones has to do with the prediction of "events" (failures) in machines. In this regard, the main task is to find a suitable model capable of predicting (with high accuracy) the defined "event". If is not possible to predict the event with high accuracy, the combination of different models is needed in order to generate the desirable output. Interesting approaches have been done in this case, in which multi-model approach outperforms a single-model approach (Kim et al., 2018) reducing the probability of an error. Here, the study analyzes different factors of degradation in railway tracks and buildings, proposing a linear model to forecast degradation in normal conditions, in combination with an exponential model to predict degradation when a defect has been detected.

Other studies that try to address proactive maintenance combine different techniques depending on how far in time the prediction is. In (Raziyeh et al., 2016) the task predicted if a pipe was prone to failure considering 3 factors: pipe-intrinsic features (such as material, diameter, and age), operational features (such as corrosion, pressure, and external stresses) and environmental features (such as temperature, rainfall, and soil conditions). The research divided the prediction in two categories: mid-term prediction, and long-term prediction. For the mid-term (or annual) prediction, they used Evolutionary Polynomial Regression (EPR). For the long-term prediction, they first clustered the similar pipes using K-Means, and then applied EPR on each cluster to determine if the pipe was going to suffer damaged. The accuracy in the prediction is 83%.

The implementation of other machine learning tools is also utilized in industry. Alternatives such as Genetic Programming (GP) might grant good results when other conventional

technologies do not perform as expected. An interesting work done using GP predicts the performance degradation of a gas compressor (Safiyullah et al., 2018). The performance degradation prediction provided by the GP model is compared against the actual performance degradation of the centrifugal gas compressor. Both indicators are compared in order to determine if maintenance is necessary, having two levels or alerts: “alarm 1” when the difference between the indicators is 20% (at this stage, the spare part needed are requested) and “alarm 2” when the difference is 30% (here, the compressor must be shot down and taken to maintenance). The GP model estimates the degradation performance with 92% of accuracy.

Research related to outage prediction focuses its attention in weather to forecast a potential blackout. Complex models can predict an outage in the presence of an extreme weather conditions (such as storm, heavy rain and lightning) applying fuzzy rules (Asma and Vali, 2015). The model was built using the weather data of 3 years and tested using the weather data of 2014 in 3 different locations, having an average accuracy of 94.23%, specificity 94.2% and sensitivity of 89.53%. Other approaches look at a single weather event, and then assess the likelihood of an outage. An example of this method is described in (Rozhin et al., 2017) using a linear regression for predicting the outage. The accuracy of the study is 90%, nonetheless, the authors concluded that further studies were needed given that artificial data points were used to measure the performance of the model.

In time series, main challenge has to do with the amount of data and features that usually are involved. Due to this complexity, a general architecture for tackling time series tasks is defined (Soheila and Mohammad, 2015). In this study, different ideas and suggestions are proposed depending on the type of difficulty that the time series task presents, having two categories: *data challenges* and *algorithm challenges*. In addition, based on the nature of the data involved, a suitable set of algorithms are recommended for implementation.

CHAPTER 3

Background

3.1. Data Mining

Ubiquitous devices (e. g. sensors, portable computers, cellphones) and computers are capable of collecting a huge amount of data in a short period. Depending on the context and the nature of the data, there is the potential of providing additional information to generate new knowledge, discover patterns, or predict eventual anomalies. In order to achieve any desired goal, the data has to go through the following processes (Sumathi, 2006):

1. Collection: it corresponds to the process of gathering data. The main sources of data are:
    - Human (e. g. a person who enters data through a web portal).
    - Sensors (e. g. a sensor that measures heart rate).
    - Other systems (e. g. a retail software system that generates statistics regarding their customers).
2. Extraction: is the process of extracting unstructured data and transforming it into a more structured (or semi-structured) data. This task is essential for humans to understand the data, and for computers to easily process it.
3. Storage: depending on the nature of the data and the desired output, it can be stored in a well-defined structure (i. e. relational database) or in a looser structure (i. e. graph database). Some of the more common ways of storage are:
    - Relational (SQL) database such as: MySQL, SQL Server, Oracle.
    - Non-SQL database such as: MongoDB, Neo4j.
    - Flat files: also known as text files, in which the data is organized based on a pre-defined structure, such as JSON, XML or CSV.
4. Preprocessing: corresponds to the necessary activities needed to prepare the data for analysis. The main tasks are:
    - Data cleaning (e.g. outlier removal).
    - Data integration (i. e. gathering data form different sources and merging it in a single dataset.

- Data transformation (or data normalization).
- Data reduction (i. e. reduce the amount of records when, for example, the data is unbalanced).

5. Analysis: after the data goes through the previous steps, a new data subset is generated, which will be the input of our algorithm. Depending on the task (desired output), our algorithm could derive in the form of a model (generated using a machine learning technique) a statistic process or a new system.

All previous tasks are part of a broad field called Knowledge Discovery in Database (KDD). KDD is process of gathering and preparing data for the purpose of solving a problem, extracting knowledge, or recognizing a pattern. Due to the broadness of this concept, in the 1990's a novel concept was proposed that focuses on data analysis that is called Data Mining. Data Mining is a field of computer science that applies machine learning and statistics to discover patterns in the data (Sumathi, 2006).

3.1.1 Applications

Recent progress in databases and machine learning, has granted a wider spectrum of Data Mining Application. Some of the most common are:

Telecommunications: Telecommunication related data sets are of high quality, initially they were call records collected for billing purposes, but now they have been used for fraud detection and consumer marketing. For example, detection of patterns in the data leads to identification of users' preferences.

Climate: The data collected by satellites (Earth snapshots) and terrestrial observations (temperature, precipitation and pressure, just to name a few) allows the prediction of weather, ecological problems and the future health of the planet.

Banking: The transactions done by a customer within a period are usually used to define the customer's profit, behavior and preferences, allowing it to assign a suitable segment for credit scoring. Nevertheless, current applications are used for fraud detection in case the customer presents an unusual pattern.

Criminal Investigation: One of the features of crime analysis is the relationship between a crime and the criminal. In this regard, the analysis allows for the definition of a pattern based on the modus operandi of an offender. As the pattern gets better defined (based on other similar crimes), it is possible to define a set of suspects when a new crime is committed.

3.2. Machine Learning

The necessity of making a computer able to "learn" has been a purpose followed since computers were invented. This concept was introduced by Alan Turing in 1950 in his paper *"Computing Machinery and Intelligence"* in which he proposed the idea of making a computer *behave* like a human, rather than *think* like a human (Turing, 1950). If that is possible, the computer will acquire the ability to "learn" based on "experience". For instance, a computer that controls a house would be able to optimize the usage of every resource (power, water, gas and even data traffic) based on the behavior of its inhabitants. In other words, the house *learns the patterns and applies them based on experience* (Mitchell, 1997). Currently, it is not possible to make a computer behave like a human, but it is possible to make them "learn" through a well-defined input. Therefore, Machine Learning is the process of making a computer able to detect a pattern using input data (i.e. "learn"), applying performance measurements as feedback to improve (i.e. "experience").

3.2.1 Supervised and Unsupervised Learning

Most of the Machine Learning tasks fall into one of these two categories: Supervised and Unsupervised Learning. In Supervised Learning, there is a set of data points (or observations) and for each observation there is a response. In this case, we build a model that relates each response with an observation. Thus, given a new observation, the model is capable of providing (predicting) a response. In contrast, Unsupervised Learning there is a set of observations, but not an associated response. In this scenario, we try to find relationships within the observations (James, 2013).

Given the context of our work, we will focus our attention on two tasks: Classification (Supervised Learning) and Clustering (Unsupervised Learning).

Classification: In this task, there is a set of defined responses (or classes). Therefore, we can refer to this as a qualitative response. If the response has continuous values (quantitative response) we refer it as a *Regression* task. Some examples of classifiers are: Random Forests and Neural Networks.

Random Forest: This is a tree-based classifier. In this case, the set of observations is divided into a number of segments. In order to define the most suitable segment for a given observation, the algorithm looks at the most commonly occurring class (i.e. mode) of each segment. In Random Forest, every time that the segment is split, a random sampling of the observations is taken, considering only one of them as a potential class, decreasing the eventual correlation that might exist in each class (James, 2013). This approach is called "tree" due to the shape of the classification, in which each segment is evaluated as a suitable class.

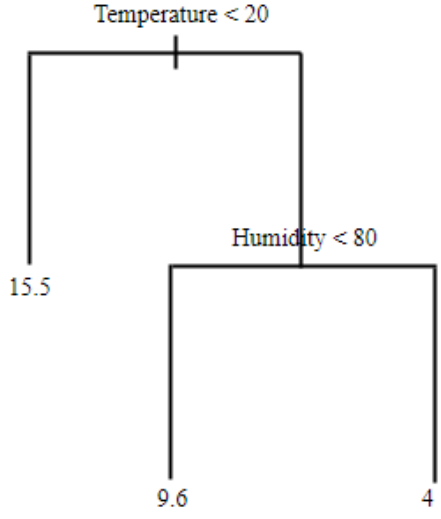

Figure 3.1 Decision tree

Neural Network: Based on biology, Neural Networks (NN) mimic the structure of the human brain's neural network. An artificial neuron (AN) is connected to others artificial neurons. Every time that an AN receives a signal as input, it will generate an output signal to the rest of connected ANs. The input signal might be intensified or diminish depending on the associated weight of the connection. The output signal of each AN is dictated by a non-linear function that computes all the incoming signals from the connected ANs. Hence, a NN is composed of several ANs that are connected though one or several layers (Andries, 2007). Usually, the anatomy of a

NN is dictated by an input layer, one or several hidden layers, and an output layer. Going back to the previous concept, the "input signal" would be the observation, and the "output signal" would be the classification (keeping in mind that the output signal might go back and forth if the NN is recurrent).

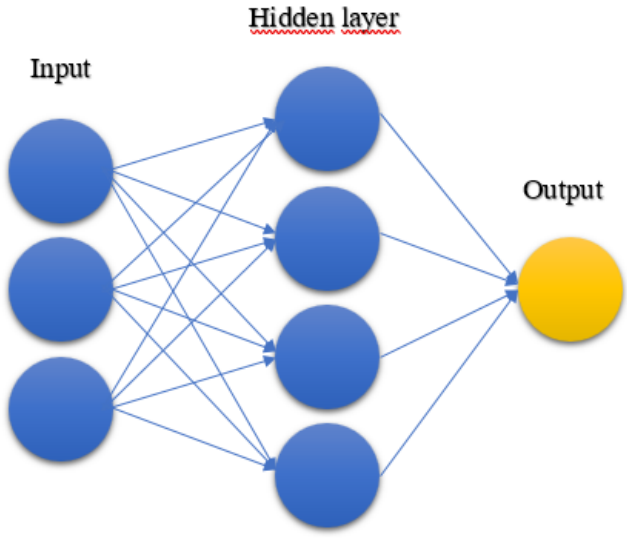

Figure 3.2 Artificial Neural Network

Clustering: The idea of clustering is to look for a relationship within the data points. For that, the observations are divided in subsets or *clusters*. Once the clusters are derived, it is possible to analyze them and conclude whether there is a relationship between the subsets or not. Of course, any conclusion has to be related to the nature of the data being clustered. One of the most popular techniques is K-Means.

K-Means: Proved to be one of the best methods for Clustering, K-Means utilize a pre-defined number of cluster *K* to group the observations. After that, each observation is randomly assigned to one of the *K* clusters. Now, the following steps are iterated until the *local optimum* is reached:

1) The *centroid* of each cluster is computed. The *centroid* corresponds to the feature means of all the observation of the *K* cluster.
2) Re-assign each observation to the nearest *centroid.* For that, K-Means calculates the *Euclidian distance.*

The *local optimum* is achieved when the *sum-of-squares* is minimum and no longer changes in the next iteration. But ¿how can we determine the initial number of clusters *K*? this can be done using experimentation. The idea is to run K-Means several times for a different number of clusters, until the objective value is minimum, and there is a clear (visual) definition of each cluster (James, 2013).

3.2.2 Applications

In real world problems, these are some of the most common applications of Machine Learning:

Speech recognition: The process of transforming spoken words into plain text (or bytes, using a lower granularity) is one of the most common applications of Machine Learning. The World's most famous assistant *Siri* is not only capable of recognizing a question with high accuracy (around 95%) but also capable of providing an answer (with lower accuracy).

Healthcare: Computer Assisted Diagnosis (CAD) allows the prediction of an eventual disease. Using the patient's medical data, the Machine Learning algorithm is capable of detecting patterns that are not visible to clinicians. In this way, if there is a high probability of occurrence of a disease, the clinician can act and mitigate any risk.

Online advertising: Also known as online marketing, applies Machine Learning to detect the pattern of thousands of customers and then define the preferences of a new customer. Thus, the new customer receives promotional marketing that might be of interest, according to his/her preferences.

3.3 Time Series Analysis

In most of the previous applications, the analysis assumes that the observations are independent of each other, and there is no order in the data points. Therefore, it is not a factor that has to be considered. For instance, if we have a classification task, in which we need to predict the salary of a person based on features such as: degrees, years of study, experience, age and field, each data point will be independent, given that the nature of the prediction is not related to the time in which the observation was recorded. A different challenge would be weather prediction, in which each data point becomes a part of a discrete time sequence. This type of task is called

Time Series Analysis. Under this type of analysis, there are two subcategories: Time Domain Approach and Frequency Domain Approach.

3.3.1 Time Domain Approach

This approach works under the presumption that current values are dependent on past values. For example, in astrology it is commonly used to determine how an astrological object changes over time. For that, it is imperative to consider past observations to determine the current or future variations of the object (Shumway, 2011).

3.3.2 Frequency Domain Approach

Contrarily, in this approach the variation of the data is subjected to a periodic oscillation that depends on external factors. In this context, most of the studies are associated with stock market prediction, social behavior and weather. In these cases, the current behavior is dictated by external (or seasonal) features (Shumway, 2011).

3.3.3 Applications

Time Series Analysis is not strictly related to prediction. Generally, the analysis tries to find a pattern of behavior or relationships between different components:

Brain's Response: Using Magnetic Resonance Imaging (MRI) the analysis tries to find how different regions of the brain react before, during and after an external stimulus.

Weather: One of the most common uses of Time Series Analysis is related to weather behavior. Using seasonal patterns, the analysis tries to study not only temperature, but also other weather phenomena (such as *El Niño*) and other features associated with location (e.g. snow, wind, humidity).

Stock Market: Over the past decade, stock market prediction is an objective that has been pursued because it might yield huge profits if it is done properly. Applying values from past trends in the prediction, several (and sometimes hundreds) of features (such as current price, open price, quarter, volume just to name a few) are applied. However, current models are not able to accurately predict the price in the short or medium term.

CHAPTER 4

Maintenance-oriented Prediction

4.1 Problem Statement

Haul *ultra-class* trucks are high-performance heavy-duty machines that require regular and scheduled maintenance due to the constant harsh environment that they are subjected to, mainly enormous payloads, uneven roads, and extreme weather. The planning of their maintenance plays a fundamental role in their utilization. Any downtime of even a single truck diminishes the productivity and has a direct impact on the company's profit. In the case of haul trucks, one of the most severe problems is related to failures of suspension. The implications of such failures involve:

- Taking care of trucks' payloads and bringing them to a shop, this happens when trucks are carrying payloads and their suspension fails;
- Re-arranging the current work schedule because of a smaller number of trucks able to move the planned payload;
- Inability to determine availability of trucks in a case of lack of spare parts, and a long waiting time for re-supplying the inventory.

Therefore, to minimize any possible loss in productivity, it is imperative to be able to predict maintenance timelines and needs for working trucks.

Objective: To perform a feasibility studies regarding need for maintenance tasks, and to determine the degree of wear of a suspension based on different data analysis methods. Further, to forecast a life-span of trucks' suspensions based on their utilization and predict their potential failures.

4.2 Data Description

The trucks are equipped with multiple sensors that take readings of their speed, payload, and pressure of all suspensions, i.e., left front (LF), right front (RF), left rear (LR) and right rear (RR).

4.2.1 Measurement Data

There are two formats of data used in this study: 1) time series data representing continuous measures strut pressure, and 2) information about notifications and failures. The data values are measured in short intervals that contain the following values:

- Time stamp.
- Payload.
- Ground speed.
- Strut pressures: LF (Left Front), RF (Right Front), LR (Left Rear) and RR (Right Rear).

The original format of this data is:

<Truck-ParamName-ReadTime-FloatValue>.

The data from six different trucks has been used. At different points of the studies, we have also analyzed data from another eleven trucks.

4.2.2 Notification/Failure Logs

Log data contains entries representing Maintenance Notifications & Failure Codes. This data comes in the format: <Unit-Notification-Creation(date)-Description-Symptoms-WorkOrder-WOCreation-FunctionalLoc-Part-DamageCode-Failure Code>. The data have been obtained for four different trucks.

4.2.3 Data Pre-processing

The format of the obtained measurement data has not been suitable for time series style analysis. A program has been developed to translate the original format of data sets into the format that resembles time series waveforms. The new format is the following:

<Time-pressures: LF, RF, LR, RR>

The program allows us to identify a desired time step between measurements. The results presented here are based on data with “measurement interval” of 5 mins. At the same time, the program has been written in a way that an issue of missing points has been addressed. At this stage, we have adopted an approach of filling the missing data with interpolated values. Most of

the suspension's pressure values (95%) are between the range of 15000 and 30000 Pa, thus, all the data points greater than 30000 Pa were treated as outliers and removed from the dataset

4.3 Initial Analysis of Data

4.3.1 Measurement Data Analysis

The first set of experiments were performed on formatted and clean data from Truck_A. The plots representing suspension pressure in LR, and RR cylinders are shown in Figure 4.1. Additionally, Figure 4.1 contains the plot of values of PITCH – one of three auxiliary values, besides RACK and BIAS[1]. In the study, we investigated these quantities only at the beginning – we have replaced them with different values, Section 4.3.2.

Figure 4.1. Plots of times series for Truck A: Feb 2014 – Feb 2016

The plots presented in Figure 4.1 have not provided "clear" indicators that could lead to identification of abnormal/failure conditions of struts. In this case, the failure is spotted using the

[1] Terms PITCH, RACK and BIAS represent different measures: PITCH is a Front and Rear Axle Wheel Strut Pressure; RACK is a Diagonal Wheel Strut Pressure; and BIAS is a Side to Side Strut Pressure.

notification/failures logs previously described. The measurements contain a lot of variations in magnitude of quantities. Over some period of time, the values have looked suspicions: “clouds” of largely spread data points before the time of “RR replacement” on RR and PITCH plots. All this has prompted us to investigate and introduce an additional processing of data.

4.3.2 Data Modifications

As the result of analysis of measurement data, we have proposed two modifications:

- Generation of “new” quantities called *differences*, and;
- Elimination of “jittering” of data via application of moving window average.

New quantities: a set of six new quantities, Figure 4.2, have been introduced in order to better capture differences between measures strut pressures at four axis: LF, RF, LR and RR. These quantities are:

- Diff left = (LF) – (LR)
- Diff right = (RF) – (RR)
- Diff front = (LF) – (RF)
- Diff rear = (LR) – (RR)
- Cross 1 = (LF) – (RR)
- Cross 2 = (RF) – (LR)

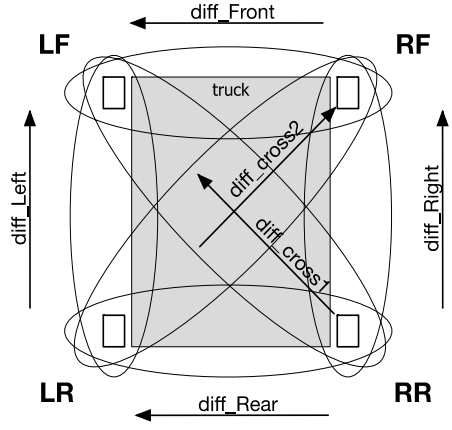

Figure 4.2 Suspension’s axis

4.4 Analysis

At first look, the dataset presents a high level of variation in values, or "jittering" behavior (Figure 4.3 (a)). To reduce those variations, a smoothing factor has been identified, taking the average of a specified number of readings. Figure 4.3 (b) represents a "smoothed" version of previous plot, taking the average of the last 200 readings. As can be observed, it is much easier to see a behavior of the measure quantity.

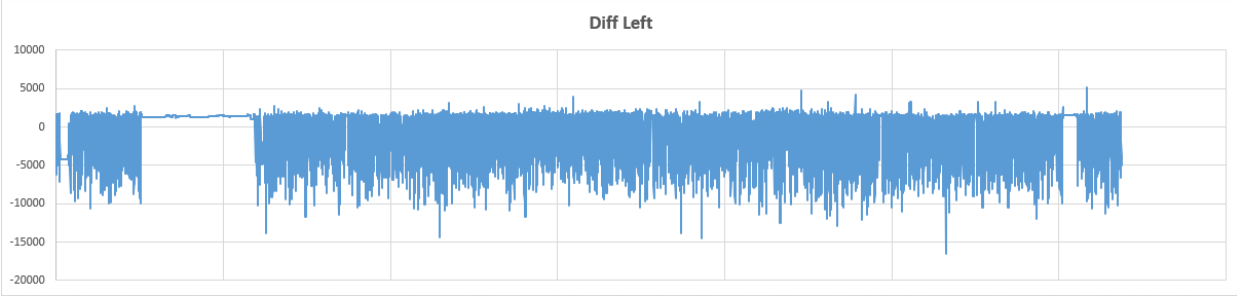

Figure 4.3 (a) Sample of feature "diff left"

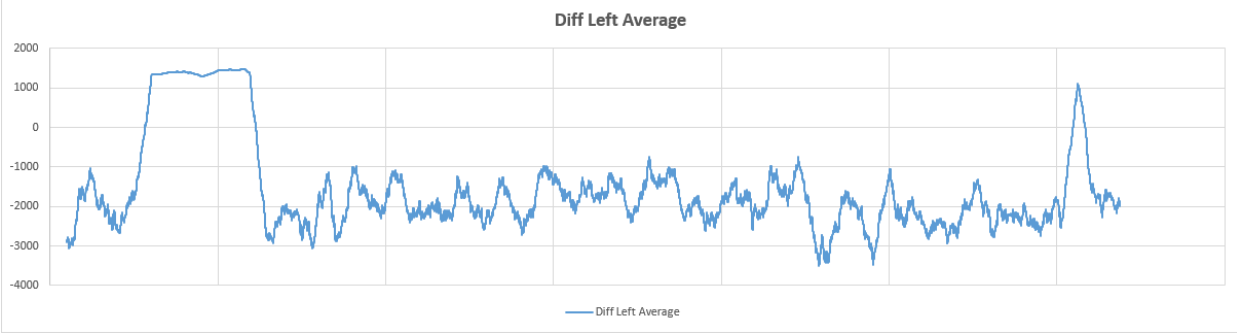

Figure 4.3 (b) Averaged "diff left"

All studies have been conducted using such a smoothing factor. In order to illustrate the difference with the smoothing factor, Figure 4.4 (a) shows all the suspension involved in the study, and Figure 4.4 (b) shows the same features within the same time range applying average with moving window.

Now, to predict the "life span" of a suspension, it is necessary to measure *the accumulative usage of a suspension over time*, result that will lead to determining a failure of a suspension. The accumulative usage will be dictated by the equation:

$$Usage = \int_{t_0}^{T} (suspension\_pressure)\,dt \tag{4.1}$$

Applying (4.1) it is possible to analyze the behavior of the suspensions given the usage. The plot below (Figure 4.5) shows a sample of "integrated" pressure values (using the six features defined), with a moving window of a week (i. e. summation of all the data points of the last 7 days), step size of a day (i. e. the window moves one day in every summation).

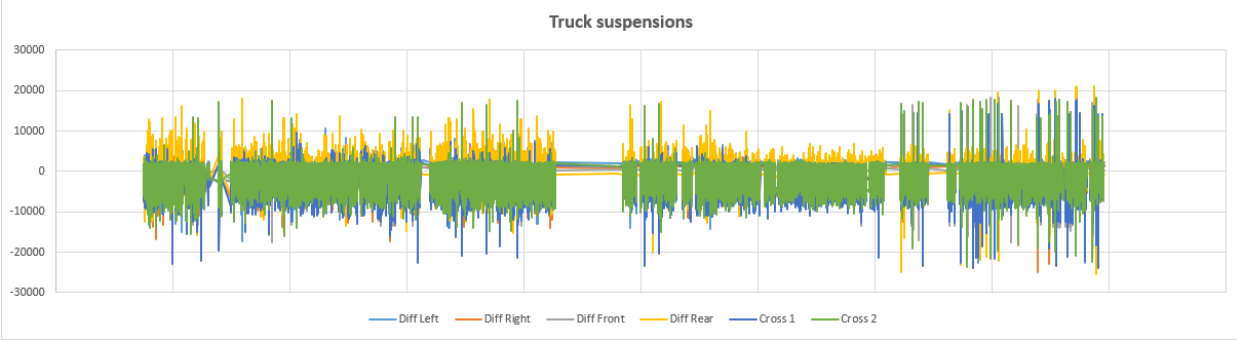

Figure 4.4 (a) Sample of all features

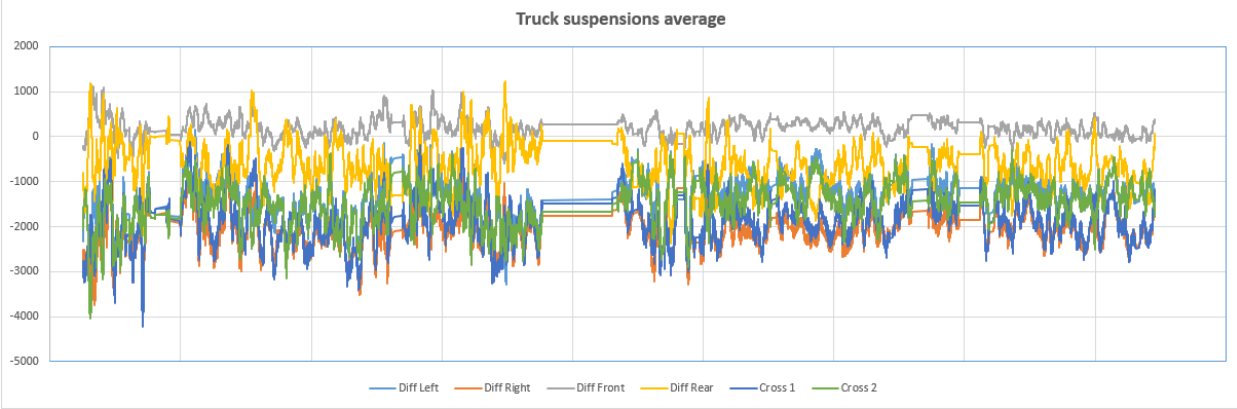

Figure 4.4 (b) Averaged all features

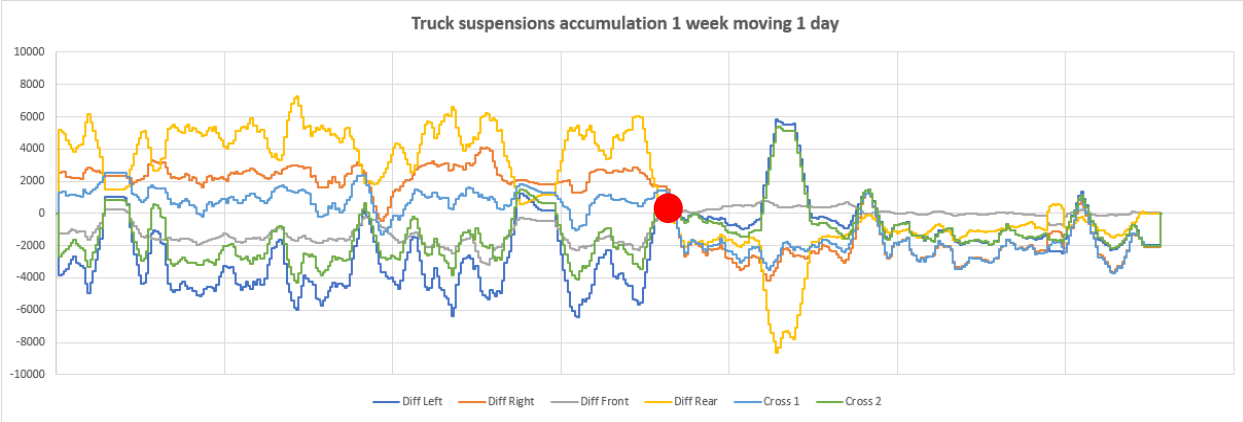

Figure 4.5 Behavior of the suspension's differences using a moving window of a week, with step size of a day

Similar results can be derived using only the values of each axis (LF, LR, RF, RR). The following graph shows the behavior.

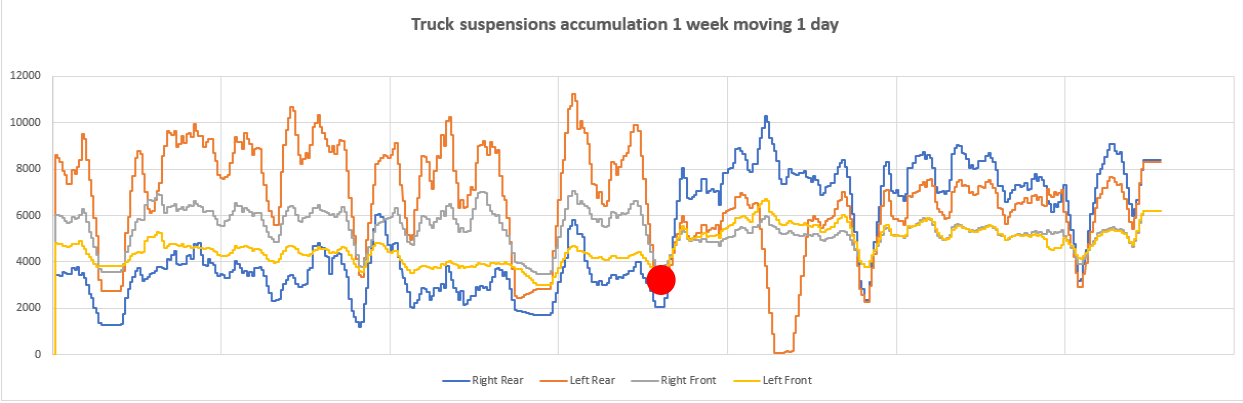

Figure 4.6 Behavior of the suspensions using a moving window of a week, with step size of a day

The red dot on both plots indicates the time of replacement of RR suspension. After the replacement, it is possible to observe that the signal is steady, having a smaller difference among the struts. This approach does not consider *accumulative* usage. In other words, given a day, the summation of the last 7 days of "usage" (pressure) is computed and plotted. For the following days, the same logic is applied, without keeping the values of previous days.

4.4.1 Integration of moving windows

More interesting results can be obtained when the summation of the pressure is kept over time; the accumulation of usage can be seen clearly when the same event happens (i.e. replacing a suspension that is failing). This behavior is depicted below.

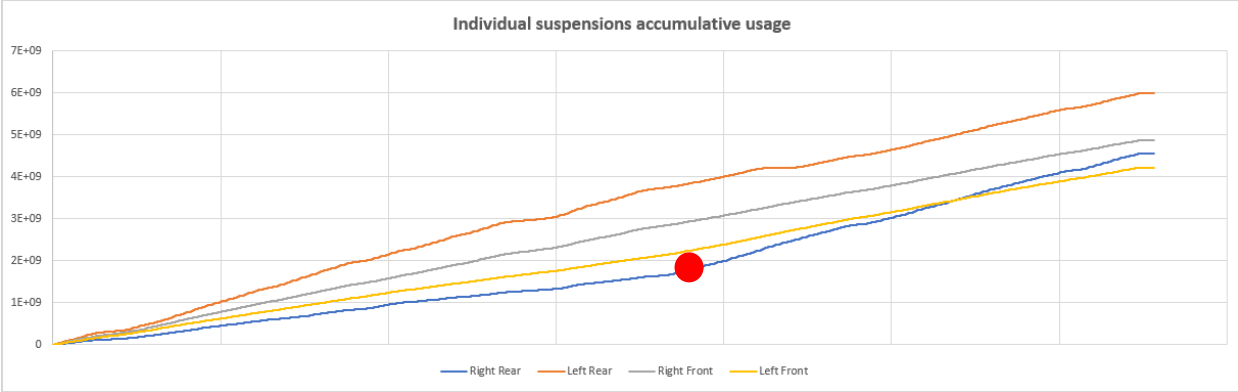

Figure 4.7 Behavior of the suspensions using accumulative usage

Likewise, the red dot shows the moment in which the defective suspension RR was replaced. The blue line represents the RR suspension. Before the change, the pressure in RR was low due to its malfunction, therefore the slop is smaller compared to the rest of the suspensions. After the replacement, the new RR suspension picks up the payload, recovering its normal level of usage. When the truck is moving with a payload, most of the weight 'is taken' by the rear suspensions. If the suspensions work correctly, the suspensions' wave-forms look steady and symmetric, showing a minor difference among the suspensions' curves of the integrated pressures LF and RF (front) have similar slopes, indicating similar "load" on each strut. The curve of LR is different, at the beginning its slope is the highest, taking part of the load from the defective RR struct. After the replacement, the gap between the rear suspensions is smaller, showing balance in the payload. These results have led to further analysis on "integrated" values of pressures.

4.4.2 Balanced signals

Using a similar approach previously described, here the level of usage is calculated based on how "balanced" the struts are. Figure 4.8 presents the main idea: focusing on suspension LR, compute the level of usage considering the differences between the suspension (i. e. "diff left", "diff rear" and "cross 2").

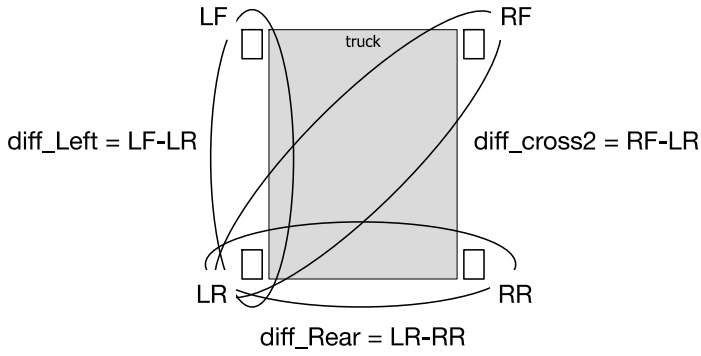

Figure 4.8 Involvement of strut LR's pressure in the suspension's differences

Based on Figure 4.8, we can say that the amount of "stress" put in LR is given by:

$$\text{Diff Rear} - \text{Diff Left} - \text{Cross 2} \tag{4.2}$$

Which leads to:

$$3LR - RR - LF - RF \tag{4.3}$$

The fact that "diff left" and "cross 2" are constantly negative (pressure in the rear is higher than in the front) requires a correction. Thus, the constant differences between the front and the rear is given by:

$$\text{Delta}_L\text{: Diff Rear} - \text{Diff Left} \tag{4.4}$$

$$\text{Delta}_{C2}\text{: Diff Rear} - \text{Cross 2} \tag{4.5}$$

Adding those constants to (4.3) the perfectly balanced suspensions is given by:

$$3LR - RR - LF - RF - \text{delta}_L - \text{delta}_{C2} = 0 \tag{4.6}$$

The equation represents the perfect balance between LR pressure (times 3, given that we have to consider the same factor in the other suspensions) and the others strut's pressures, with two constant values (deltas) between the rear and the front. If the values are plugged in the equation, it would be possible to observe unusual loads (positive and negative) in reference to a normal condition. These results can be further processed to determine if the suspension has been over-used due to a malfunction in one of the other suspensions. If we plug the values to (4.6) the following results are obtained:

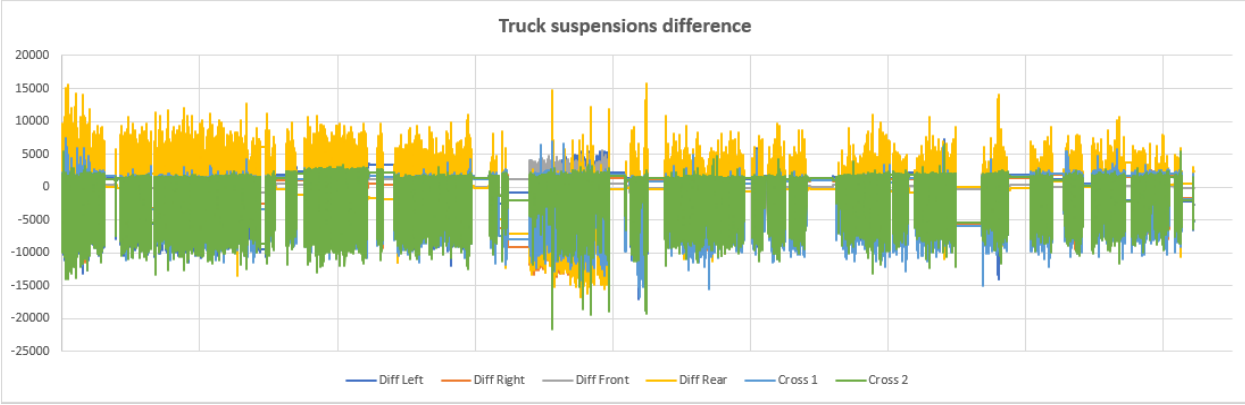

Figure 4.9 (a) Strut's differences in a normal behavior

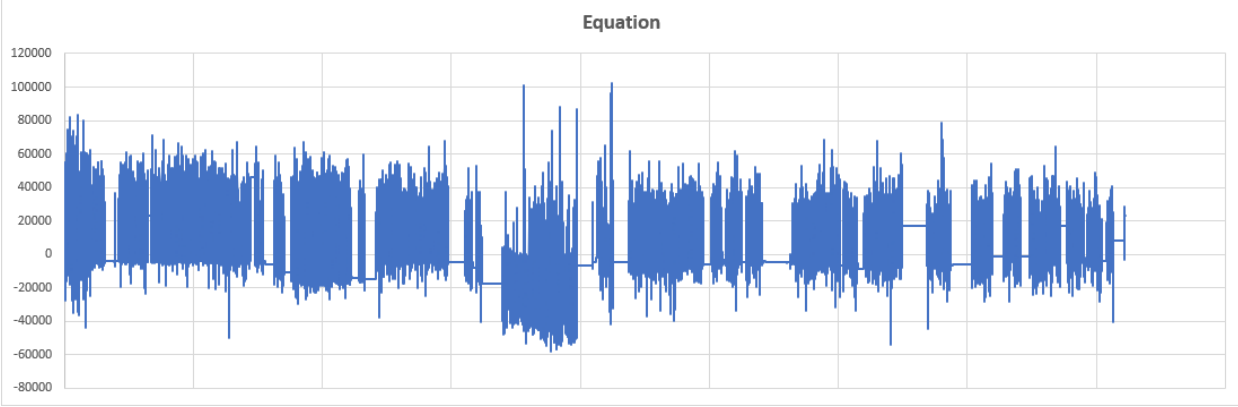

Figure 4.9 (b) LR behavior applying the equation from (4.6)

Figure 4.9 (a) depicts the original values, while Figure 4.9 (b) shows the results of the equation applied in LR. The values above zero indicate a usage higher than normal, and below zero indicates lower than normal. If we further process the results from Figure 4.9 (b) it is possible to define a threshold-based approach combined with summation to determine the level of the strut's usage. Based on the observations, three thresholds were defined the determine levels of "stress" that the suspension has been exposed to. In other words, each time a value exceeds a threshold, the "signal" is replaced by one of three possible numbers:

- If the threshold of 2000 is crossed – a signal of value 1 is used;
- For crossing the threshold of 5000 – a signal of value 2 is formed;
- When crossing the threshold of 10000 – a signal of value 3 is created.

Figure 4.10 (a) illustrates the results of the application of the thresholds, based on the wave presented in Figure 4.9 (b). Finally, Figure 4.10 (b) shows the summation of the pulses. This represents a way to determine the level of usage of a given suspension.

Figure 4.10 (a) Pulses generated based on the thresholds

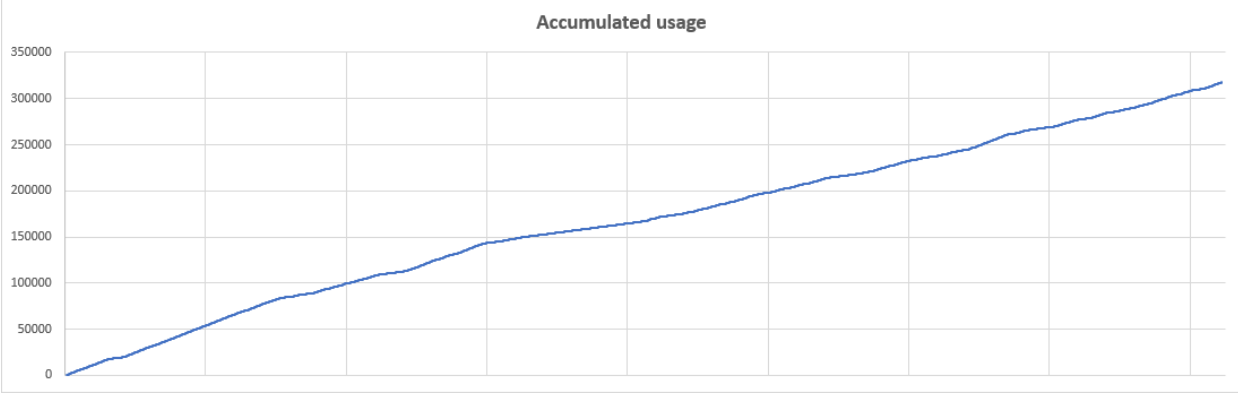

Figure 4.10 (b) Accumulated values generated based on the pulses.

4.5 Discussion

The generation of new features (“diff right”, “diff left”, “diff rear”, “diff front”, “cross 1” and “cross 2”) with moving averages enabled us to clearly see the behavior of the suspensions when the truck is carrying a payload. Most of the weight (pressure) is taken by the rear suspensions, making them more prone to failure. In a normal state, the gaps between the “differences” is minor (i.e. the amplitude of each wave form is close to zero). When one of the suspensions has a defect, the gap between each feature is visible, making the amplitude of each wave greater.

The first approach (integration of moving windows) allows to accumulate the *usage* of each suspension. In this case, a failure can be detected, but over time; the change of the angle of the slope is visible long after the suspension is replaced. The second approach also grants with the accumulation over time but using the balance among the struts.

With these findings, we can conclude that: first, the level of usage of a given suspension can be measured; second, detect when a failure is occurring. Combining these two events, it would be possible to predict a malfunction of a suspension. In other words, based on the lifespan of a brand-new suspension (provided by the manufacturer) it is plausible to estimate when the suspension will complete its lifespan by accumulating its usage, and then look at its behavior after it reaches this threshold. Any abnormal behavior is readily detected looking at the gaps between each suspension's wave form, which can be done using threshold-based algorithm or a machine learning model. However, to reach this stage, further studies have to be performed, given that all the studies were done over "old" struts, thus it is necessary to define the values of these thresholds (for usage and gaps between the struts) using a brand-new suspension (or in an ideal scenario, a brand-new truck).

CHAPTER 5

Event Prediction in Power Systems

5.1 Problem Statement

Outages are events that are difficult to handle in a timely manner. When they happen, they are usually reported by one or several users that were affected by a given outage, or by someone who saw that something wrong happened to a power line, a switch or a transformer. Once the affected area is approximated, a group of technicians is sent to the location to inspect the place and try to identify a problem. When the problem is detected, reparations begin. It is not possible to determine how long they take place, especially, considering such factors as: difficulty of repairs, availability of needed supplies and man power. In this case, the most affected stakeholders are users who have several problems created due to the absence of electricity.

The necessity to anticipate an outage is imperative to avoid any major problems (and hazardous situations) caused by an interruption of electricity.

Objectives: To integrate and analyze outage related data. To build outage prediction models. To identify, automatically, types of outages.

5.2 Data description

In power systems, a large number of quantities are being measured with different frequency. Under this scenario, we can divide measured variables into two groups:

- Internal: these are variables that represent different entities/measures of the system; some examples of these are: power consumption, number of customers connected to a transformer, age of a given device, number of repairs done on a device, etc.
- External: these variables represent quantities not related to the system, but that might affect its operations; most common of those are weather features: temperature, humidity and wind speed.

As it will be shown, these variables are combined in different ways to construct models predicting an eventual outage.

All measured quantities are stored in multiple databases that are distributed among multiple locations in electrical utility sites. The process of preparation of data involved a number of activities that required interaction with utility personnel in order to understand meanings of individual attributes/features of collected data. One of the important activities has been data integration. Fig 5.1 represents a snapshot of tables from two databases: Outage Management System (OMS Database) and Maintenance database (Maximo):

- OMS: a database system that contains information about the outages reported by customers, entered by operators, and provided by personnel involved in taking care of an outage.
- Maximo: a database system that contains information about the maintenance done in a given equipment (e. g. poles, transformers, switches, etc.).

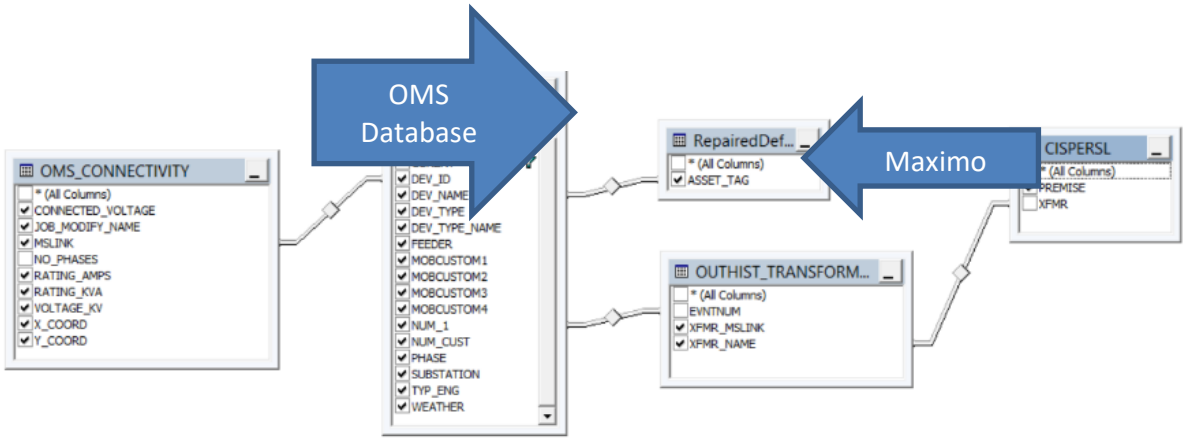

Figure 5.1. Tables and integration "links" between them

5.3 Visualization and interactive analysis of data

Provided the complexity and the number of features describing an outage, it is essential to understand all the quantities involved in representing an outage, and how they are related to each other. In this context, graph databases provide an interesting view at the tuple <node-relationship-property> to define a schema:

- Node: represents an entity within the model.
- Relationship: represents a link (or association) between different entities defined within the model.

- Properties: represents the entity's features.

At the beginning we have modeled the available data using a special schema (Section 5.3.1) and then visualized the model accordingly.

5.3.1 Schema

In database systems, a schema represents the structure of the data described in a formal language. The main entities of the schema are represented by tables (nodes) that are connected to other entities through constraints (relationships). After analyzing the data available to us, the main entities involved in an outage are the following:

- Outage: has general information of the outage (such as id, date, location, number of affected customers).
- Event: has information regarding the cause of the outage.
- Cause: domain node that defines all possible causes associated to an outage.
- Calls: represents all the calls from the customer that reported the outage.
- Root transformer: contains the location and the id of the root transformer that was related to the outage.
- Transformers: contains the id of all the transformers involved in the outage.
- Customers: contains the id and the location of the affected customers.
- Location: defines the X and Y coordinate of the location of the customers.
- Breaker: contains the path of all the devices affected by the outage, starting from the root transformer and ending in the customer's transformer.

The graph model is depicted in Figure 5.2:

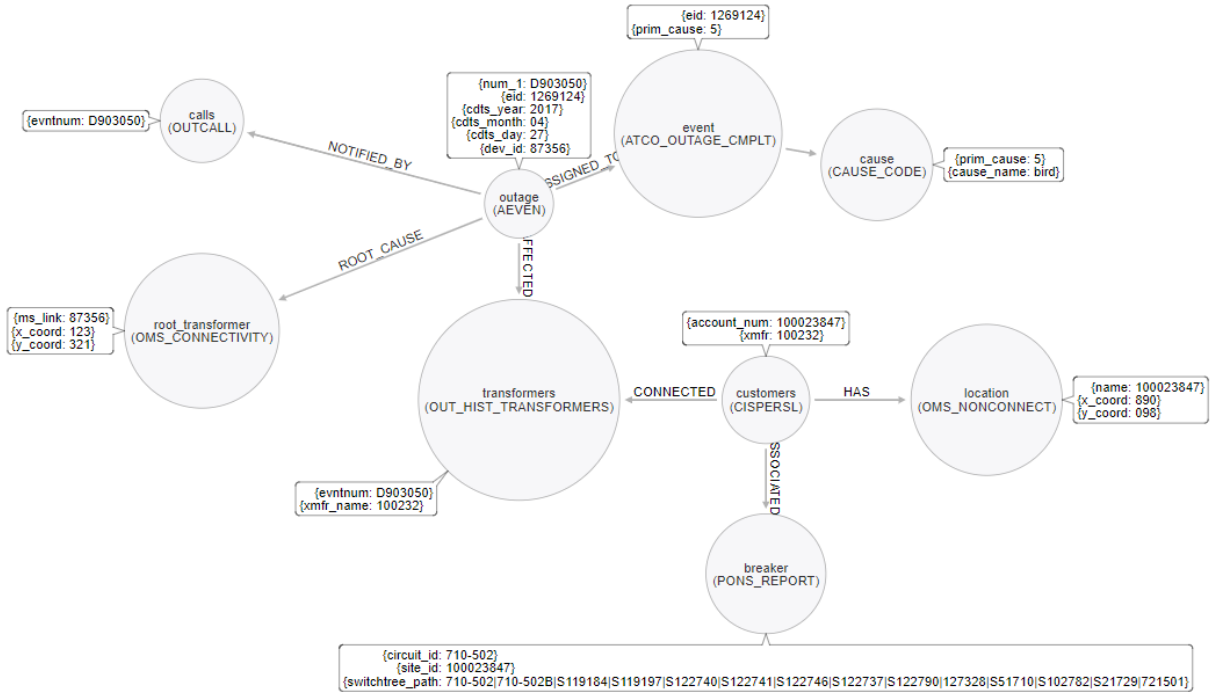

Figure 5.2 Schema with the entities, relationships and properties involved in an outage.

Once the conceptual model has been defined, we have loaded it into a graph database called Neo4j (Neo4j, 2018).

5.3.2 Database population

For manipulating and interacting with the database, we use *Cypher* which is a simple query language for graph databases. The process of creating the model is divided in three parts:

- Create nodes: using the data source, we create all nodes of the model. Cypher command below shows how to create a node "Outage" with properties "eid" (outage id), "num_1" (event id) and "cdts" (date in which the outage happened) when loading the data from a *CSV* (Comma Separated Values) file.

```
LOAD CSV WITH HEADERS FROM
'file:///AEVEN_01.csv' as line
WITH line
WHERE line.CURENT = "T"
```

CREATE (o:Outage {eid:line.EID, num_1:line.NUM_1,
cdts:substring(line.CDTS,0,8), dev_id:line.DEV_ID})

- Create constraints: to maintain integrity of the data, it is necessary to create the constraints according to the structure of the data; a Cypher command below defines a unique constraint for the node "Outage", property "num_1".

  CREATE CONSTRAINT ON (o:Outage) ASSERT o.num_1 IS UNIQUE

- Create relationships: applying conceptual schema defined, we create all relationships of the model; a Cypher command shown below defines a relationship between two nodes "Outage" and "Transformers".

  MATCH (o:Outage),(t:Transformers)
  WHERE o.num_1 = t.evntnum
  CREATE (o)-[:AFFECTED]->(t)

Once the graph database is loaded, we can query the model and extract the data for our analysis.

5.3.3 Querying the graph

Using defined relationships, we are able to obtain values of properties from the nodes that we are interested in. Depending on what we need to obtain, and the relationship defined among the nodes, we prepare proper graph queries. For example, in order to retrieve the customer locations from all the outages we follow the model defined in Figure 5.2, using the following query:

MATCH (o:Outage)-->(t:Transformers)-->(c:Customers)-->(l:Location)

RETURN o.num_1 as outage, t.xfmr_name as transformer, c.premise as customer,
l.x_coord as x, l.y_coord as y

Locations of customers involved in all outages can be retrieved based on the relationship defined in the model: "match", and "return", and the properties "num_1" (outage), "xfmr_name" (transformer), "premise" (customer), "x_coord" (coordinate X) and "y_coord" (coordinate Y). Table 5.1 shows a sample of the result returned by the query.

Outage	**Transformer**	**Customer**	**X**	**Y**
D964096	937575	1169101272	28345735	603335040

D964096	663111	7941289	28418864	603283266
D964097	P2575452	3109389	15324631	577123292
D964096	663089	7941287	28212441	603272844

Table 5.1 Result returned by the query (first four records)

An alternative way to visualize the data is via displaying relevant nodes. For that, we can select all the nodes using a command:

MATCH (n:Outage) *RETURN* n

Once we display nodes representing outages (Figure 5.3) we can expand nodes and see their relationships with other nodes (Figure 5.4). A simple graph shown in Figure 5.4 illustrates a set of relationship between an outage node (red), a transformer node (pink), a customer node (yellow) and a location node (blue). The number in each node represents an id of that node.

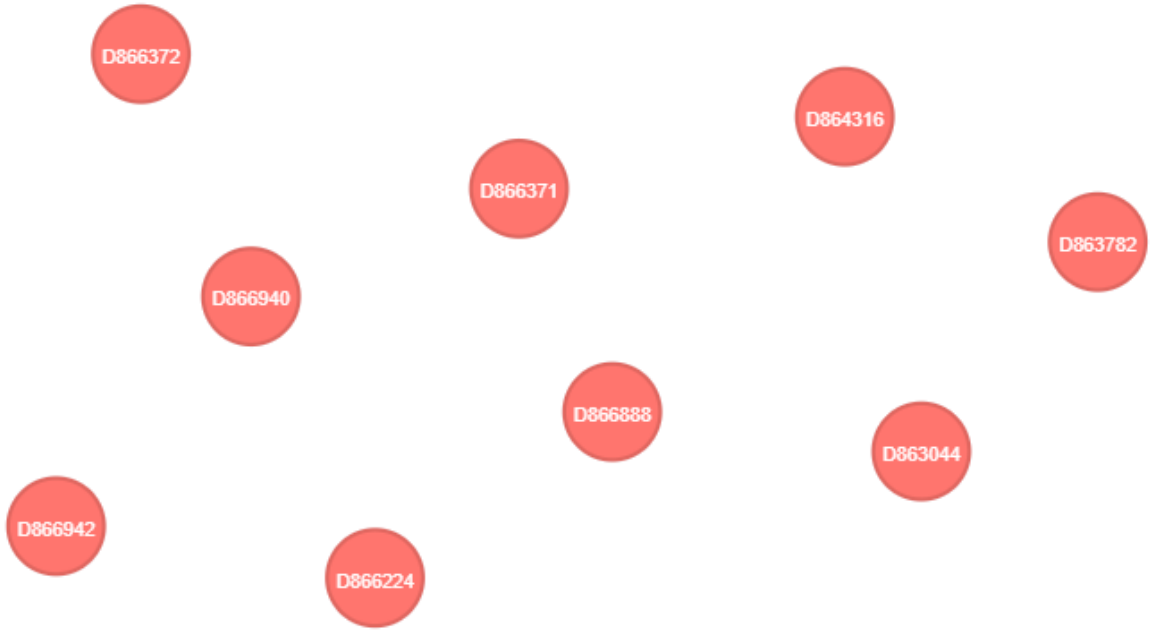

Figure 5.3 Displayed outage nodes

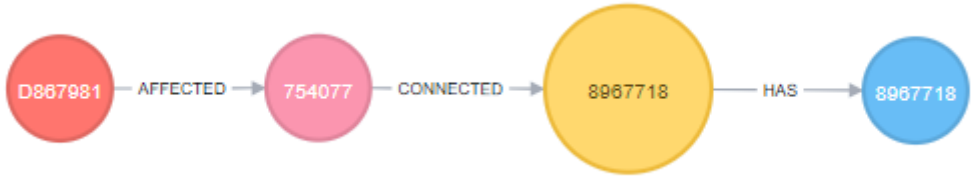

Figure 5.4 Alternative visualization provided by the graph database

Previous example shows that the selected outage has one "affected" transformer, which is "connected" to one customer, and that the customer "has" a location. Further visualization provides more complex situations as depicted in Figure 5.5, in which multiple users are connected to a single transformer affected by an outage.

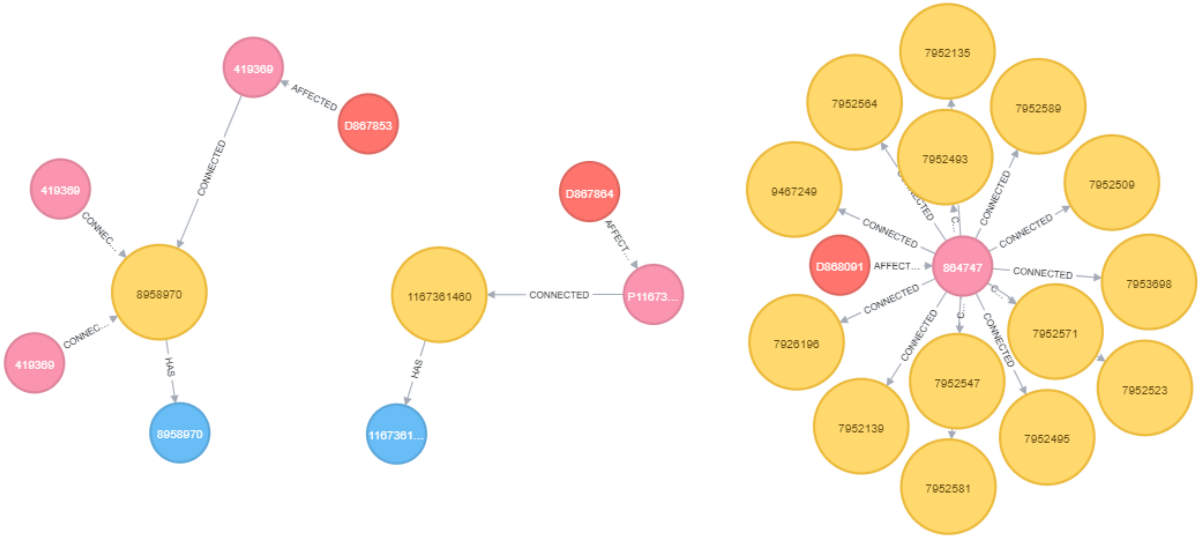

Figure 5.5 Three outages were selected and "expanded" to see their relationships; one of them (on the right) with several customers affected

The graph visualization provides an interactive way of analyzing multiple different relationships that leads to visual exploration of data according to the users' needs.

5.4 Unsupervised techniques in analysis of outage

5.4.1 Clustering process

The unsupervised learning technique allows us to determine if there are isolated groups – clusters – within the data. The technique we use here is *K-Means*. If clusters are determined, they can be used to "classify" a new data point based on its closeness to one of the clusters. As a feasibility study, we focus on a small portion of the dataset, more precisely, outages reported in January of 2017. Only non-continuous features have been selected for this purpose:

- Customer_no: number of customers affected by the outage.
- Voltage: connected voltage.
- Dev_type_name: type of the device affected.
- Weather: description of the current state of weather; for example: "normal", "storm", "hurricane", etc.
- Mobcustom2: range of temperature at the time an outage happened; for example: "-19 to -30", "-6 to -18", etc.
- Mobcustom3: wind's strength; for example: "light", "strong", "moderate", etc.
- Mobcustom4: weather phenomena involved; for example: "normal snow", "light rain", "heavy rain", "none", etc.
- Expr1: corresponds to the phase at which an outage occurred.

All the nominal values, such as "dev_type_name", "weather", "mobcustom2", "mobcustom3", "mobcustom4", have been transformed to numerical values that are required by the K-Means algorithm. Using the described features, we have analyzed data using three different combinations of features. In each case, the optimum *k* number of clusters is three.

Case 1: all features are used. Fig. 5.6 shows the result of the clustering. In this scenario, there is no conclusive result; there is no a clear separation between the clusters. In addition, the diameter of cluster nº1 is almost twice the size of the cluster nº3.

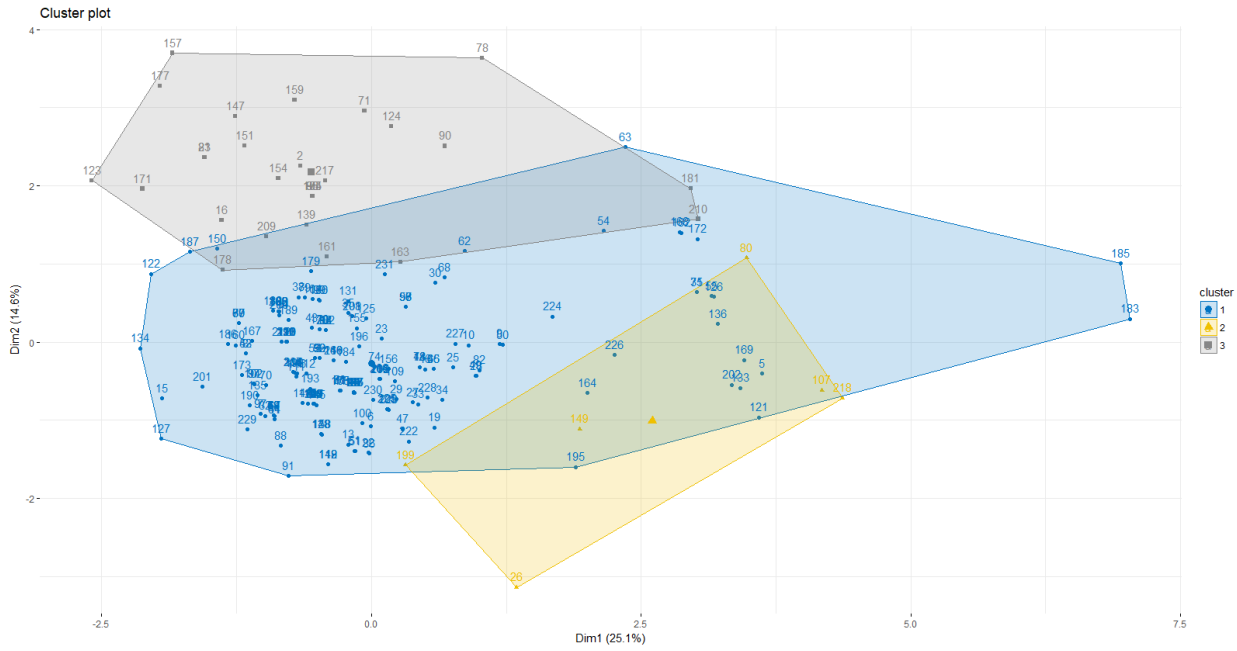

Figure 5.6 Clustering of all features

Case 2: two features "customer_no" and "expr1" are discarded. Fig 5.7 shows the result of the clustering; there is a minor improvement. In this experiment, there seems to be less overlap between the clusters, but again it is possible to observe a cluster (nº2) with a long diameter.

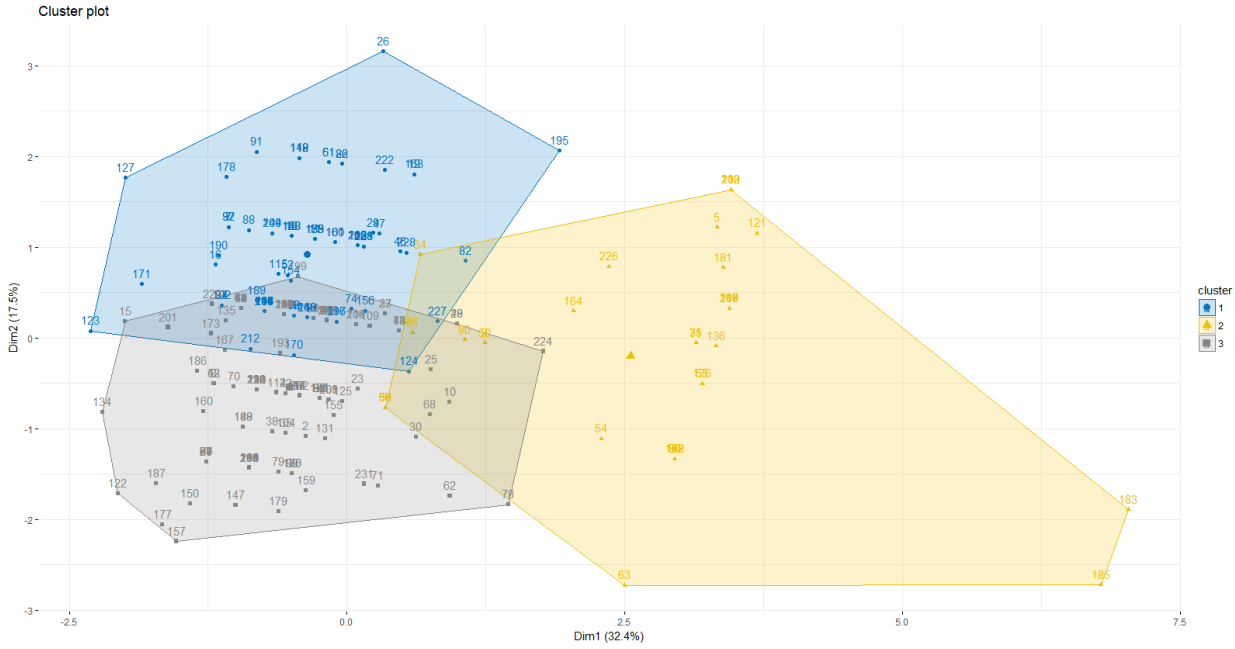

Figure 5.7 Clustering of features: voltage, dev_type_name, mobcustom2, mobcustom3 and mobcustom4

Case 3: Another feature – "voltage" – is discarded. In this case, two of the clusters are practically overlapped, therefore removing this feature has not improve the results, Fig 5.8.

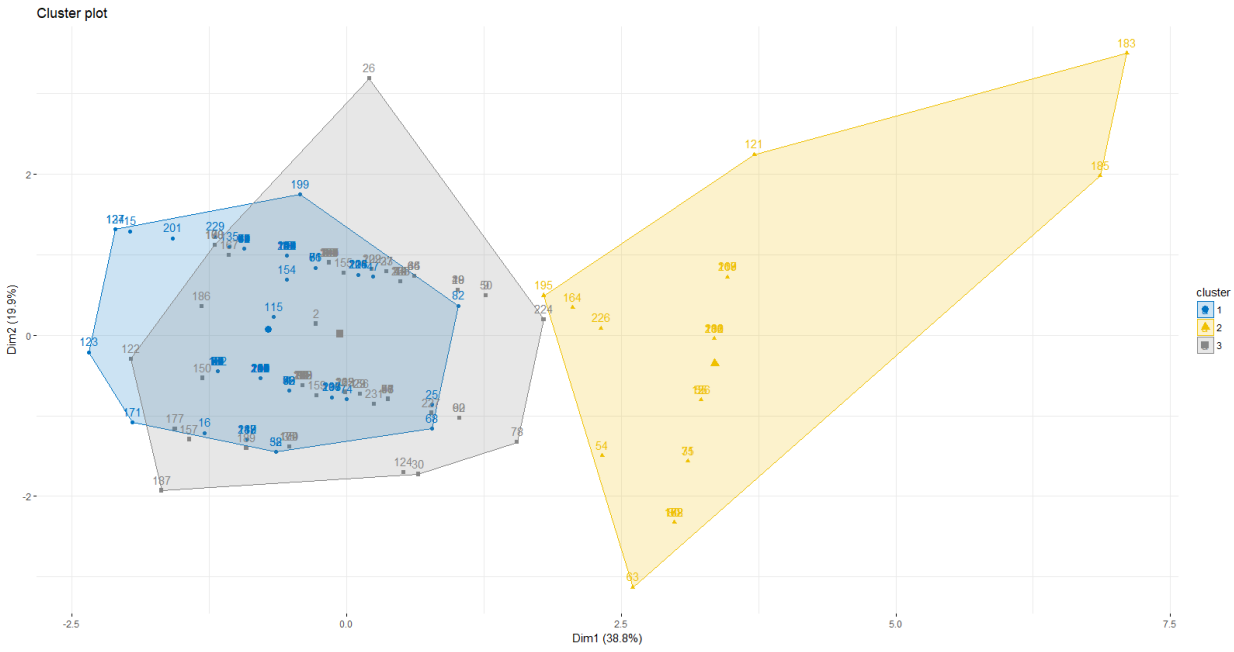

Figure 5.8 Clustering of features: dev_type_name, mobcustom2, mobcustom3 and mobcustom4

Visually, case 1 and case 2 presents similar results given the shape of the clusters.

5.4.2 Statistical analysis

In order to compare the results of clustering, two metrics are used to determine how good the clusters are in each case. These performance metrics are: *total sum of squares*, and *between sum of squares*:

- **Total sum of squares ("Total_ss"):** it represents a ratio the sum of squared distances of each data point to the global mean (Tusell, 2016).
- **Between sum of squares ("Between_ss"):** it represents a ration the sum of square distance of each *centroid* to the global mean (Tusell, 2016).

According to these definitions, if centroids of clusters are close to the global mean, it indicates that the clusters are not distinguishable between each other. To determine which results are the best among all cases, we apply the formula:

Between_ss / Total_ss

Thus, the closer to one this ratio is, the better the clustering result is. Now, the values of the ratio for each case are:

- Case 1: 0.85
- Case 2: 0.37
- Case 3: 0.43

Based on the ratio, Case 3 represents a better result than Case 2, having a better measure of distance between centroids; however, this might be misleading, considering the overlap between the clusters obtained for Case 3, Figure 5.7 and Figure 5.8. Case 1 represents the best result.

In summary, we can state that unsupervised-based analysis of data does not provide a very good result. The obtained clusters have provided us with a basic understanding of relations between individual data points and features.

5.5 Supervised techniques in analysis of outage: prediction models

A process of building a model requires a number of steps that can be quite different every time a model is constructed. There is a lot of flexibility in methodology that defines which approach to follow for a given dataset/classification task. Of course, there are well-known actions that must be done to ensure reliable results, such as: preprocessing, data cleaning, normalization, and feature selection.

The data we used for model construction include:

- OMS (Outage Management System).
- Weather data: features related to the environmental conditions; some of these features are: temperature, humidity, solar radiation, wind speed, pressure, just to name a few

In the following sections, we describe different approaches we have applied to construct models suitable for prediction outages using the available data.

5.5.1 Prediction based on OMS and weather (temperature and humidity) data

This first approach shares some similarities with the unsupervised analysis described in the previous section. Most of the features used for building a prediction model are the same.

5.5.1.1 Data description

The data set has information over a time range from 2012 until 2015. It contains information about all outages reported during that period. In comparison to the unsupervised learning we have performed, such features as "voltage", "dev_type_name", "expr1" are kept, while the feature "customer_no" has been removed due to its dependence on extensive post-mortem analysis of outages.

The features: "weather", "mobcustom2", "mobcustom3", "mobcustom4" have been replaced by new features: "temp" (temperature) and "hum" (humidity). These new values have obtained using the R Studio library "rclimateca" (https://cran.r-project.org/web/packages/rclimateca/index.html). This library has allowed us to query the weather data from The Environment Canada climate repository (Historical Climate Data, 2018), using a time and the "X" and "Y" coordinates of an outage.

The following new features have been also added as part of the data:

- Substation: power system that provides electricity to a service area.
- Feeder: power line that transfers electricity from the substation to the transformers.

Table 5.2 depicts a sample of the data used. The cause codes in the table represents the causes of outages as identified by personnel servicing a given outage (see Appendix A):

Voltage_kv	Dev_type	Substation	Expr1	Feeder	Temp	Hum	Cause_code
25	313	759S	ABC	759-502	23.7	31	9.6
25	313	NULL	ABC	763-504	-11.3	77	9.6
25	314	759S	A	759-502	7.7	91	4
25	300	777S	ABC	777-502	14.2	58	7
25	304	777S	B	777-502	32.4	18	4
999	313	728S	B	728-503	12.1	56	3
25	313	759S	ABC	759-502	23.7	31	9.6
25	314	700S	A	700-504	17.1	52	5
25	300	720S	ABC	720-502	19	20	3
25	300	720S	ABC	720-504	19.7	68	4

Table 5.2 A sample records of the dataset

5.5.1.2 Modeling

For the classification task, all the categorical values, such as: “substation”, “feeder” and “expr1”, have been transformed into numeric values. As a preliminary study, the data has been split 60% for training, and 40% for testing, taking a random sample for generating both subsets. The target of the classification is the value “cause_code” which represents the cause of an outage. The prediction model is Random Forests, which has been identified as the one that provides the best results.

5.5.1.3 Results

The process of constructing a prediction model has been done using R Studio (R Studio, 2018). A confusion matrix is shown in Table 5.3. The performance measures of the model are included in Table 5.4. We also provide the values of measure averaged over all types of outages:

- Average: sensitivity 0.501; specificity: 0.957; precision: 0.746;
- The global accuracy 74%.

	Reference							
Prediction	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9.6**
2	33	0	0	5	0	0	0	0
3	0	588	1	40	5	0	0	2
4	9	394	2626	280	343	335	0	26
5	797	111	105	2724	290	2	13	276
6	0	6	6	27	1538	0	0	24
7	0	0	0	0	0	34	0	0
8	0	0	0	0	0	0	0	0
9.6	0	22	24	56	3	1	0	1572

Table 5.3 Confusion matrix for each cause code

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9.6**
Sensitivity	0.0393	0.52453	0.9508	0.8697	0.7058	0.0914	0	0.8274
Specificity	0.9995	0.99571	0.8549	0.8265	0.9938	1	1	0.9898
Precision	0.8684	0.92744	0.6543	0.6308	0.9606	1	0	0.9368
Balanced Accuracy	0.5194	0.76012	0.9028	0.8481	0.8498	0.5457	0.5	0.9086

Table 5.4 Statistics by class

Worst results are obtained applying 10-fold cross validation:

- Average: sensitivity 0.206; specificity: 0.894; precision: 0.479;
- The global accuracy 63%.

	Reference							
Prediction	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9.6**
2	36	0	0	0	0	0	0	0
3	0	310	96	76	3	0	0	126
4	1250	2136	5115	3624	1779	717	963	3794
5	789	1966	2548	14743	3969	113	742	7385
6	30	1065	5005	5005	2632	312	804	3765
7	0	0	0	0	0	38	0	0
8	0	0	0	0	0	0	0	0
9.6	6	281	934	934	85	3	0	1651

Table 5.5 Statistics by class

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9.6**
Sensitivity	0.017	0.0538	0.538	0.604	0.310	0.032	0	0.098
Specificity	1	0.995	0.766	0.615	0.811	1	1	0.967
Precision	1	0.508	0.26	0.457	0.19	1	0	0.423
Balanced Accuracy	0.508	0.524	0.675	0.609	0.561	0.516	0.500	0.533

Table 5.6 Statistics by class

As it can be seen, the worst results are obtained for the cause codes “loss of supply” (2) and “adverse environment” (7). In “loss of supply” and “adverse environment” the low sensitivity is due to the number of missed-classifications. “Human element” (8) is a special case because it does not have a significant number of instances, thus the indicators are not meaningful (nevertheless, it is considered as part of the overall accuracy of the classifier).

The specificity obtained for each class is high due to a large number of “true-negatives”; in this scenario we compare one class against the rest. The precision is lower for such classes as “lightning” and “defective equipment” which might be attributed to the fact that these cases are more difficult to predict, and they represent a substantial portion of the observations. The balanced accuracy in “loss of supply” and “adverse environment” is also misleading because it considers the accuracy of the rest of the classes.

5.5.2 Prediction based on OMS and weather

In the second approach, we consider a full set of features related to weather (not only temperature and humidity). The number of features used is large due to the sampling of weather conditions considering not only the time when an outage occurred, but also the weather conditions before and after the outage took place.

5.5.2.1 Data description

The dataset contains information about outages that occurred from 2015 until 2018. As has been mentioned before, the number of attributes used for weather is larger than in the previous case. Now, we have 18 features (see Appendix B). The frequency of the sampling is a single measurement of each feature done on an hourly basis. We consider the weather conditions at the outage location starting 12 hours before, during, and 3 hours after an outage took place. Therefore, we have 288 features for the weather alone. Table 5.7 shows the sample of each feature, representing the weather 12 hours before the outage.

M2t_.12	D2t_.12	Wbt_.12	Rh_.12	Spk_.12	Wp_.12	Wd_.12	Cc_.12	Php_.12
-21	-25	-21	70	97.8	13	350	100	0
-19.8	-22.09	-19.8	99	94.1	10.9	312	99	0.00254
-18.9	-20.9	-19.1	84	93.1	18.4	30	100	0.05
-32.8	-35.8	-32.9	74	94.4	7.6	190	85	0
-22.2	-24.93	-22.3	98	95.6	4.2	252	100	0
-21.7	-24.24	-21.7	99	95.6	12.2	306	99	0
-23.1	-25	-23.3	84	96.2	7.6	290	95	0
-17.1	-19.2	-17.4	84	95.8	5.4	50	50	0.05
-11.5	-12.41	-11.5	99	90.1	9	262	100	0.022
-15.5	-17.29	-15.6	97	94.6	24.7	330	99	0.00254

Table 5.7 (a) Sample of the weather dataset; the first nine features

Dnr_.12	Dsr_.12	Dhr_.12	Wc_.12	At_.12	Hi_.12	Sf_.12	Press_.12	Wg_.12
0	0	0	-30	-30	-21	0	102.54	17.6
0	0	0	-27.3	-27.3	-19.8	0.3	104	54.1
0	0	0	-28.6	-28.6	-18.9	1.4	102.5	28.1
0	0	0	-40.9	-40.9	-32.8	0	102.9	13.3
0	0	0	-22.2	-22.2	-22.2	0	104.2	52.3
0	0	0	-30.1	-30.1	-21.7	0	104	72.1
0	0	0	-29.6	-29.6	-23.1	0	104.9	14.8
0	0	0	-21.3	-21.3	-17.1	0	105.4	7.6
15	78	75	-16.6	-16.6	-11.5	0.4	102.7	69.9

0	0	0	-25.7	-25.7	-15.5	0.1	102.9	96.1

Table 5.7 (b) Sample of the weather dataset; the second set of features

In addition to the weather data, we also have features from OMS that are related to the system and location only. Thus, from OMS we use the following six features:

- Service_Area: location in which the outage took place.
- Dev_id: unique identifier of the root device involved in the outage.
- Mslink: main transformer affected the outage.
- Scale: type of topology (rural or urban).
- Feeder: power line that transfers electricity from the substation to the transformers.

Table 5.8 illustrates a sample of the OMS data.

Service_area	Dev_id	Mslink	Scale	Feeder	Mw.load.corrected
Fort McMurray	7905203	7905203	Urban	820-501	0
Beaverlodge	8886006	8886006	Rural	815-503	0
Hanna	2894950	2894950	Rural	803-505	0
St. Paul	7903124	7903124	Rural	707-501	0
Falher	9064586	9064586	Urban	784-505	0
St. Paul	1.14E+09	1.14E+09	Rural	707-502	0
Vegreville	1.1E+09	1.1E+09	Rural	711-501	0
Hanna	2894890	2894890	Rural	801-502	0
Swan Hills	8882036	8882036	Rural	743-503	0
Bonnyville	7899418	7899418	Rural	859-503	0

Table 5.8 Sample of the first 10 records of the OMS dataset

The total number of features is 295, including the target, i.e., a cause code to be predicted. To reduce the number of attributes, feature selection process has been performed. It is described in next section.

5.5.2.2 Feature selection

In this task, the target class “cause_code” is transformed in two categorical values: “scheduled” and “unscheduled” outage. For that, all the “cause_code” type “schedule outage” are classified as “schedule” and the rest of the “cause_code” are classified as “unscheduled”. Using the subset evaluator “FirstBest” that provides the best subset of features, based on the selected target class,

from Weka (Weka, 2018), we obtain a subset of features that better defines our target class. The output of the process is described below:

Selected attributes:
2,3,6,25,28,32,45,138,143,146,152,154,155,164,165,169,170,171,182,184,185,186,187,188,189,191,277: 27

DEV_ID, MSLINK, MW.Load.Corrected, D2T_.11, D2T_.8, D2T_.4, WBT_.7, PHP_.10, PHP_.5, PHP_.2, DNR_.12, DNR_.10, DNR_.9, DNR_0, DNR_1, DSR_.11, DSR_.10, DSR_.9, DSR_2, DHR_.12, DHR_.11, DHR_.10, DHR_.9, DHR_.8, DHR_.7, DHR_.5, Press_1

The notation "_." (underscore point) plus "number" indicates that the value represents a quantity measured "number" of hours before an outage. In this case, if the hour is zero, the value corresponds to the time when an outage happened. The new subset has 28 features with the target class.

5.5.2.3 Modeling

In the approach of taking OMS and weather data, our target (output) variable represents only two distinct values determined based on the "cause_code". All the causes considered to be not related to weather have been removed, i. e., "unknown/other", "tree contacts", "defective equipment" and "adverse environment". The remaining "cause_code" has been labelled as follows:

- Scheduled outage → Non-weather outage
- Lightning → Weather outage
- Human element → Non-weather outage
- Foreign interference → Non-weather outage
- Adverse weather → Weather outage
- Non-outage → Non-weather outage

Therefore, the classifier's task is to predict outages that are related to weather, or not related to weather, rather than predicting all the possible causes with different classes. Two different sets of experiments have been conducted: one as a feasibility study where data has been split 70%

for training and 30% for testing, taking a random sample for generating both subsets; and one as a 10-fold cross validation. The prediction model is Random Forests, which provided the best results using the features selected as it was explained in the previous section.

5.5.2.4 Results

The confusion matrix is shown in Table 5.9. The performance measures obtained for the prediction model are:

- Sensitivity: 0.813
- Specificity: 0.750
- Precision: 0.778
- Balanced accuracy: 0.782
- The global accuracy of this approach is 78%.

	Reference	
Prediction	**Non-weather**	**Weather**
Non-weather	2450	699
Weather	564	2101

Table 5.9 Confusion matrix with the classification of each class

Similar results are obtained with 10-fold cross validation:

- Sensitivity: 0.813
- Specificity: 0.759
- Precision: 0.785
- Balanced accuracy: 0.786
- The global accuracy of this approach is 78%.

	Reference	
Prediction	**Non-weather**	**Weather**
Non-weather	8090	2203
Weather	1858	6925

Table 5.10 Confusion matrix with the classification of each class for 10-fold cross validation (cumulative matrix combining each experiment results)

Better results are obtained mainly because the classifier is predicting two well-defined classes, rather than trying to predict several “similar classes”. Here, all the similar instances are in the

same group, improving the sensitivity and precision. The balanced accuracy is almost identical to the global accuracy given that the dataset is balanced between the classes.

5.5.3 Prediction based on weather only

The approach that uses only weather data is presented here. The motivation of such a scenario has been motivated by the fact that some OMS data is difficult to provide in real time. The procedure described previously with a difference of removing all features related to OMS is applied. Now, the number of features is 25 (including the target class). The purpose of this experiment is to test if the removal of features related to the status of a system at the time of outage affects the accuracy of the prediction.

5.5.3.1 Modeling

The procedure as described in the previous section has been applied. As before we have performed a single experiment where the data is split 70% for training and 30% for testing, taking a random sample for generating both subsets, and a series of experiments contributing to 10-fold cross validation. The prediction model is Random Forests, which provides the best results.

5.5.3.2 Results

Table 5.11 contains the results of prediction obtained for Random Forest model.

- Sensitivity: 0.800
- Specificity: 0.720
- Precision: 0.763
- Balanced accuracy: 0.760
- The global accuracy of this approach is 76%.

	Reference	
Prediction	**Non-weather**	**Weather**
Non-weather	2463	765
Weather	605	2030

Table 5.11 Confusion matrix with the classification of each class

Similar results are obtained applying 10-fold cross validation:

- Sensitivity: 0.812
- Specificity: 0.725
- Precision: 0.763
- Balanced accuracy: 0.768
- The global accuracy of this approach is 77%.

	Reference	
Prediction	**Non-weather**	**Weather**
Non-weather	8082	2506
Weather	1866	6622

Table 5.12 Confusion matrix with the classification of each class with 10-fold cross validation

In this approach, we can readily conclude that the elimination of information related to the system affects the overall performance of the classifier, slightly diminishing the accuracy of it.

5.5.4 Prediction of outage and associated weather phenomena

This work is slightly different from the one presented above. This time a prediction system is composed of a model to predict an outage, and three new models to determine which weather phenomena has been involved in the outage. For the development of these new models, a subcategory of "cause_code", called "supplementary_cause_code", has been used. We have distinguished with values:

- Icing/No Icing
- Extreme wind/No Wind
- Snow/No snow

Considering that these features are related to weather, for the feature selection process only weather features were used, having as a target class "supplementary_cause_code". As in the case of previous approach, the algorithm "BestFirst" from Weka has been used. For the outage prediction model, the same data features have been used with a single new feature "supplementary_cause_code"

The first stage of the prediction model, "outage model", predicts if an outage takes place. If "yes", the second stage uses three models: "icing model", "snow model", and "wind model" to

determine which phenomena is involved. Thus, one, two or all these factors might be involved. If the first stage predicts that there is “no outage”, no further prediction is done. Figure 5.9 depicts the integrated model.

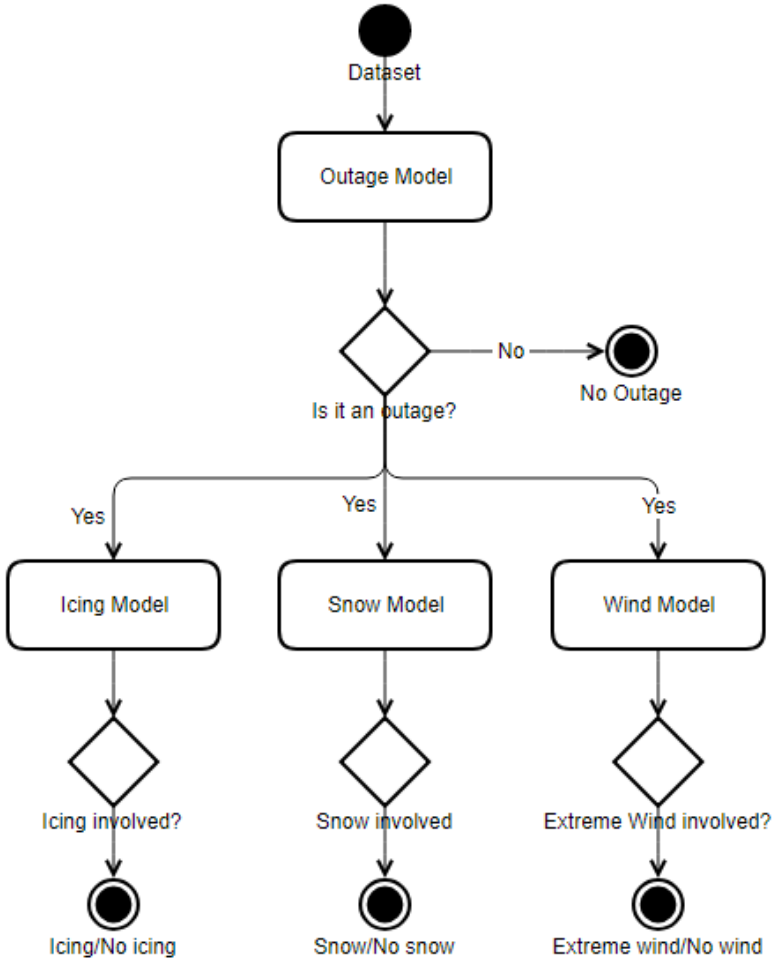

Figure 5.9 Flow chart of the integrated model

5.5.4.1 Modeling

For generation of these models, the data was split 70% for training and 30% for testing, taking a random sample for generating both subsets. The prediction model is Random Forests, which provides the best results.

5.5.4.2 Results

Table 5.13 contains the results of outage prediction obtained for Random Forest model.

- Sensitivity: 0.940
- Specificity: 0.740
- Precision: 0.80
- Balanced accuracy: 0.840
- The global accuracy of this approach is 85%.

	Reference	
Prediction	**Non-outage**	**Outage**
Non-weather	2844	707
Weather	161	2060

Table 5.13 Confusion matrix with the classification of each class

Tables 5.14-16 contain the results of the prediction results for each weather condition.

	Reference	
Prediction	**No-Snow**	**Snow**
No-Snow	2303	76
Snow	29	301

Table 5.14 Confusion matrix with the classification of Snow

	Reference	
Prediction	**Non-Ice**	**Ice**
Non-Ice	2202	57
Ice	92	405

Table 5.15 Confusion matrix with the classification of Ice

	Reference	
Prediction	**Non-Wind**	**Wind**
Non-Wind	2432	149
Wind	29	129

Table 5.16 Confusion matrix with the classification of each class

5.6 Discussion

Preliminary results show that an outage prediction model can provide better results if there is a combination of information coming from weather and the system itself. The clustering of the first dataset has not provided useful information; the overlap between the clusters does not

permit to further analyze a pattern. In contrast, the supervised approach granted good results combining internal (OMS) and external (weather) data.

The attempt to predict each “cause_code” (first approach) produced good results in terms of accuracy, but was inconsistent from the point of view of precision and sensitivity. Mainly, because of the similarity among the classes, and the unbalance distribution of the data. Grouping the classes (second approach) solved both issues, giving as a result with marginally higher accuracy, but a more consistent sensitivity and precision. The third approach proved that the deletion of internal information decreases the overall output of the model (using the same dataset and the same modeling procedure). The last approach attempts to also predict the weather. The high accuracy in the outage prediction is due to the addition of the new feature, however, it corresponds to a “post-mortem” feature. The low accuracy in the weather classification (“extreme wind”) responds to the unbalance positives and negatives samples of the dataset

This study has a limitation in terms of internal data availability. Currently, OMS does not provide any information related to the equipment involved in power system outages (for example, how old a device is? How prone to failure a power line is?) These sorts of features should be considered when trying to build a more accurate prediction model. Furthermore, the location plays an important role as well. Ideally, in a real-world system the model should predict an outage based on a selected location (or service area) indicating the probability of an outage for that service area. Location plays a fundamental role, considering that there are other external factors that have not been used, such as topology of the network, geography and vegetation, that might have influence on an outage. The integration of more data is needed to improve current approach and develop a model that provides a user with a probabilistic output for a given service area.

CHAPTER 6

Conclusion, Contribution and Future Work

6.1 Conclusion

In this thesis, we have presented a process of applying main concepts of Data Mining and Machine Learning to real world problems. In particular, two different problems from industry have been addressed. First, in the case of analysis of haul truck suspension data we have analyzed the behavior of trucks' suspensions by generating new features and visually representing waveforms of usage of suspensions. After that, a measure of suspensions' fatigue has been created using a concept of moving windows and compound suspension values. Second, in the case of prediction of outages in power systems we have addressed the problem predicting outage events. At the beginning we have used an unsupervised learning, i.e., K-Means clustering without good results. Further, we have applied supervised learning techniques to build prediction models – Random Forests. We have obtained prediction accuracy of 74%. Despite such relatively good results, our results have been impacted by limitations in available data impeding construction of a fully developed model. This sort of problem is very common in industry that struggles to create a reliable and comprehensive infrastructure that supports sensors and IoT devices. In order to take advantage of Data Mining and Machine Learning techniques, it is imperative to have available data of good quality.

6.2 Contributions

Processes of data analysis performed on industrial data lead to a number of contributions. Overall, our study has resulted in a set of findings regarding methodology of processing industrial data.

We can state that a number of different approaches can be adopted for data analysis depending on availability of data and desired outputs. Main factors influencing the choices are related to:

- Nature of data, would that be time series, aggregated or integrated data;
- Quality of data, i.e., consistency of data collecting processes, existence of missing values, and even availability of data for on-line modeling;

- Understanding of data, visual-based analysis of data inter-relationships and explanation of phenomena to be analyzed and predicted;
- And usefulness of data, such issues as: correspondence of features' names, importance of features for predicting purposes, availability of data for real-time prediction purposes.

Besides using ML/DM techniques for constructing prediction models, we apply novel technologies leading to better, deeper understanding of data. The most significant of them is a graph database. It has provided us with the ability to "see" data in a different way. That view focuses on relationships between data pieces representing different entities and aspects of phenomena that are important for analyzed problems.

6.3 Future Works

Further studies will address the following tasks:

- **Suspension's failure in haul-trucks:** More data collected on brand new trucks is required to better determine trends and characteristics of actual life span of a given suspension. The quantity "accumulated usage" can be determine applying the described method (Section 4.4). Similarly, more data and experiments need to be done in order to determine such quantities as threshold values and likelihood of suspension to fail. An important element that should be determined is the range of slopes of plots representing accumulated usage (Figures 4.7 and 4.10(b)) that represent a normal and abnormal utilization of suspensions. Additionally, another threshold-based analysis can be developed to determine when anomalies/failures occur in struts.
- **Outages in power systems:** An integration of multiple sources of data representing multiple aspects of power system (or *internal* data) is needed to improve accuracy of developed models. Currently, this type of data is available in several geographically distributed databases. This data integration must consider also "on-line" availability of information in order for model to predict in real-time. Moreover, more investigation is required to determine the impact of the localization on models' performance. Different prediction model architectures should be developed and tested in order to provide not only to predict outages but also probabilities of their occurrences.

References

Andries P. *Computational Intelligence;* Wiley: South Africa 2007.

Asma, R., Vali, D., Fatemeh, A. Applying Hierarchical Fuzzy Systems to Predict Unplanned Feeder Outages in The Yazd. *Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS).* 2015.

Cran package "rclimateca" (https://cran.r-project.org/web/packages/rclimateca/index.html) (accessed Aug 24, 2018).

Historical Climate Data (http://climate.weather.gc.ca/) (accessed Sep 08, 2018)

James, G., Witten D., Hastie, T., Tibshirani, R. *An Introduction to Statistical Learning;* Springer: New York 2013.

Kim, V., Bart, D., Robert, B. A Multiple-Model Reliability Prediction Approach for Condition-Based Maintenance. *IEEE.* 2018.

Mitchell, T. *Machine Learning;* McGraw-Hill: 1997.

Neo4j (https://neo4j.com/product/) (accessed Sep 08, 2018)

Raziyeh, F., Konstantinos, K., Kourosh, B., David, B. Pipe Failure Prediction in Water Distribution Systems Considering Static and Dynamic Factors. *International Conference on Water Distribution Systems Analysis.* 2016.

Rozhin, E., Amin, K. Machine Learning Based Power Grid Outage Prediction in Response to Extreme Events. *IEEE*. 2017.

R Studio (https://www.rstudio.com/) (accessed Sep 08, 2018).

Safiyullah, F., Sulaiman, S., Naz, M., Jasmani, M., Ghazali, S. Prediction on Performance Degradation and Maintenance of Centrifugal Gas Compressor Using Genetic Programming. *Energy.* 2018.

Saiied, M., Moe, A. Development of Fibre-optic Sensors for Australian Mining Industry. *Conference on Lasers and Electro-Optics Pacific Rim.* 2017.

Shumway, R., Stoffer, D. *Time Series Analysis and Its Applications,* 3rd ed.; Springer 2-3: New York 2011, 1-3.

Soheila M., Mohammad R. An Analytical Review for Event Prediction System on Time Series. *International Conference on Pattern Recognition and Image Analysis*. 2015.

Sumathi, S., Sivanandam, S.N. *Introduction to Data Mining and its Applications*, vol. 26; Springer: Poland 2006, 8-9.

Turing, A. *Computing Machinery and Intelligence;* Mind: 1950, 433.

Tusell, F. Interpreting result of k-means clustering in R, 2016. Stack Exchange Web site. https://stats.stackexchange.com/q/48528 (accessed Aug 11, 2018).

Weka (https://www.cs.waikato.ac.nz/ml/weka/) (accessed Sep 08, 2018).

Yeon, S., Jun, H. Issues and Implementation Strategies of the IoT (Internet of Things) Industry. *International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. 2016.

Appendix A: Cause code description

Outage's causes defined in OMS system.

CAUSE_CODE	CAUSE_DESC
0	Unknown/Other
1	Scheduled Outage
2	Loss of Supply
3	Tree Contacts
4	Lightning
5	Defective Equipment
7	Adverse Environment
8	Human Element
9.6	Foreign Interference
6	Adverse Weather
10	Non Outage

Appendix B: Weather code description

Weather's features used in the prediction task.

WEATHER_FEATURE_CODE	WEATHER_FEATURE_DESC
M2T	Temperature
D2T	Temperature Dew Point
WBT	Surface Wet Bulb Temperature
RH	Relative Humidity
SPK	Pressure Altimeter
WP	Wind Speed
WD	Wind Direction
CC	Cloud Coverage
PHP	Precipitation 24 hour
DNR	Direct Normal Irradiance
DSR	Downward Solar Radiation
DHR	Diffuse Horizontal Radiation
WC	Temperature Wind Chill
AT	Temperature Feels Like
HI	Temperature Heat Index
SF	Snow 24 hour
Press	Pressure Mean Sea Level
WG	Wind Gust