# Modeling failure of oil pipelines

Kimiya Zakikhani[1], Tarek Zayed[2], Bassem Abdrabou[3], Ahmed Senouci[4]

**Abstract:** As the safest means of transporting gas and hazardous materials, pipelines transport invaluable petroleum material. However, a considerable number of accidents have happened on these facilities, leading to economic losses and environment impacts. Several inspection techniques are used to provide safety for these pipelines. Despite their accuracy, these techniques are time-consuming and costly procedures. Some failure prediction and condition assessment models were recently developed to tackle these inefficiencies. However, most of these models only predict one failure source or they rely on subjective expert survey results. This research developed three objective models based on artificial neural network (ANN) and multi nominal logit (MNL) regression to predict failure source in oil pipelines. An ANN model was developed for prediction among mechanical, corrosion, and third-party failures with an average validity percentage (AVP) of 73.7%. Another ANN model was developed for prediction between corrosion or third party failures with an AVP of 72.8%. In addition, a MNL model was developed for prediction among mechanical, corrosion, or third party failures with an AVP of 73.7%. Pipeline operators and decision makers can use these models to identify pipeline failure sources. They can also be applied to facilitate in-line inspection prioritization to take appropriate maintenance measures.

Keywords: Oil pipelines; failure source prediction; regression; artificial neural network, multinomial logit.

---

[1] Department of Building, Civil and Environmental, Engineering, Concordia University, Montreal H3G 1M8, QC, Canada, Email: kimiya.zakikhani@gmail.com ,
[2] Department of Building and Real Estate (BRE), The Hong Kong Polytechnic University, Hong Kong, Email: tarek.zayed@poly.edu.hk
[3] Listuguj first nation government, 44 Dundee Road Listuguj, QC, Canada, Email: Abassem77@gmail.com
[4] Construction Management Department, University of Houston, Houston, United States, Email: asenouci@central.uh.edu

## Introduction

Pipelines, which are critical components in the oil and gas industry, transport invaluable worth of goods. Compared to rail and highway transportation, pipelines represent the safest transportation mean of petroleum products. However, statistics show that pipeline accidents can lead to catastrophic environmental impacts and severe economic losses (Dey et al. 2004). CONCAWE (Davis et al. 2010), which is a European organization that investigates environmental, health and safety issues for the oil industry, reported that pipeline failures are due to natural hazards, operational, corrosion, third party and mechanical sources. It was established in 1963 to carry out environmental research for the majority of European oil companies (Davis et al. 2010).

Over the past 20 years, new inspection techniques have been developed to detect pipeline anomalies and defects without stopping production such as ultrasonic testing (UT) and magnetic flux leakage (MFL). Because these techniques are highly expensive and time consuming, researchers have developed several failure prediction models and condition assessment procedures for oil pipelines. These models allow decision makers to prioritize inspection actions to prevent pipeline failure. However, most of the current models are either dependent on expert subjective opinions or evaluating only one failure type. Due to these limitations, historical data based objective models are needed to predict failure sources for oil pipelines.

This research aims to develop several failure prediction models for the identification of failure types of oil pipelines. The models can predict the failure type, which can be either mechanical, corrosion, or and third party. These failure types cover more than 80% of oil pipeline accident sources (Davis et al. 2010). Three failure prediction models are developed;

(i) An artificial neural network (ANN)-based model that predicts failure among mechanical, corrosion or third-party failures;

(ii) An ANN-based model that predicts failure between corrosion or third-party-failures;

(iii) A Multinomial logit (MNL) model that can predict failure caused by mechanical failure, corrosion, or third-party failures.

These models can help oil pipeline operators to make decisions on the required actions on pipelines to protect them against threats and to prioritize inspections.

## Background

Over the past decade, significant efforts have been conducted on the prediction of failure in petroleum pipelines including the probability of failure, rate, consequence, source, age, and pressure. As one of the most studied parameters, probability of failure has been estimated through application of different methods such as Fuzzy technique, AHP, Fault tree, and simulation. Kabir et al. (2016), Bertuccio and Moraleda (2012), Yuhua and Datao (2005), Zhou et al. (2016), Li et al. (2016) estimated the probability of failure for petroleum pipelines using a Fuzzy technique. Li et al. (2016), Kabir et al. (2016) and Hasan (2016) determined the probability of failure by developing a hierarchical risk structure based on expert opinion using AHP method. Omidvar and Kivi (2016), Yuhua and Datao (2005) and Li et al. (2016) predicted the probability of failure to specific causes using fault trees. Moreover, Li et al. (2009), Witek (2016), Wen et al. (2014), and Dundulis et al. (2016) estimated the probability of failure using experimental equations or finite element analysis results of burst pressure due to corrosion failure or a combination of failure types.

Parvizsedghy and Zayed (2015a) predicted the financial consequences for gas pipelines using neuro-Fuzzy technique and a bow-tie model. Moreover, Restrepo et al. (2009) estimated financial consequence of oil pipelines using binary logistic and least square regression methods. Even though they relied on historical records, the methods were merely capable of predicting monetary consequences and did not consider environmental and safety costs.

Few studies were conducted on predicting the cause of failure. Senouci, El-Abbasy et al. (2014) established a Fuzzy-based model to estimate the source of failure in oil pipelines based on historical records. In another study, Senouci, Elabbasy et al. (2014) predicted failure source in these facilities using a back-propagation neural network and multiple regression. In both studies, the prediction of failure source has been treated as a predictive problem rather than classification, which may question the accuracy of the results. In other words, the nominal variables were treated as numerical values. Moreover, Bertolini and Bevilacqua (2006) developed a multi- regression model using a regression tree (C& RT) approach, that predicts failure causes using available records in CONCAWE database (Davis et al. 2010). However, these models lack a validation procedure, which may question their accuracy.

Several research studies were conducted to predict failure rate and pressure. Liao et al. (2012) and Caleyo et al. (2009) predicted corrosion rate in gas pipelines, respectively, using a back propagation neural network and Monte Carlo simulation. Papavinasam et al. (2010) predicted the internal corrosion-pitting rate of oil and gas pipelines using laboratory experiments on pitting corrosion rate. The research conducted by Liao et al. (2012) and Papavinasam et al. (2010) was limited to corrosion failure prediction. Moreover, the methods developed by Caleyo et al. (2009) and Papavinasam et al. (2010) relied on performing expensive in-line inspection results. In a similar study, Ma et al. (2013) developed a model to estimate magnitude of failure pressure in petroleum pipelines using finite element analysis.

Parvizsedghy and Zayed (2015b)developed two regression models to estimate the failure age of oil pipelines (Parvizsedghy and Zayed 2013; Parvizsedghy and Zayed 2015b). Even though their diagnostic measures in these studies were satisfactory, the models were based on a combination of different failure sources. This issue may become problematic since the mitigation measures for

each failure source is distinct from the others.

The literature review of the studies on the failure prediction of oil and gas pipelines identified few important research gaps. The first limitation is that few of these studies were limited in the number of failure parameters considered in the model. Ma et al. (2013), Parvizsedghy and Zayed (2015a) and Restrepo et al. (2009) only considered one or two failure consequences. Moreover, Hasan (2016), Omidvar and Kivi (2016), Bertuccio and Moraleda (2012) and Li et al. (2016) only considered one or two failure types. As another limitation of the reviewed models, few were subjective, which lead to a lack of generalization. Kabir et al. (2016), Yuhua and Datao (2005), ( Li et al. (2016) proposed models based on expert judgments, which are very sensitive to the responses provided in the opinion surveys. Finally, few of the developed models were based on a limited number of records, which limit their generation. The models developed by Zhou et al. (2016), Dundulis et al. (2016) and Ma et al. (2013) were either based on limited in-line inspections, finite element analysis, or historical records. Therefore, the application of such methods to other pipelines may be ineffective (Abdrabou 2012). Most of the developed models were subjective. In other words, they are not applicable to all pipelines. They are also case specific due to their reliance on subjective measures such as expert opinions. On the other hand, few models can predict merely one type of failure among five different failure sources that will be introduced later. Finally, few of these models were developed using a limited number of records obtained from in-line inspections, which leads to a limitation in their application to pipelines with different attributes (Abdrabou 2012).

***Failure sources in oil pipelines***

According to CONCOWE database, failures in oil pipelines fall in one of five categories, namely, mechanical, corrosion, third party, operational, and natural hazards. Figure 1 presents the

contribution of various failure sources in oil pipelines according to CONCAWE database (Davis et al. 2010). This figure illustrates that 88% of total accidents were due to either mechanical, corrosion, or third party failures (Abdrabou 2012).

1) *Mechanical Failure**:*** Mechanical failure is due to poor construction or the use of low-quality materials (Dey et al. 2004). This failure can be further classified into two main categories (i.e., dents and gouges) that appear as deformations in the pipe wall. Dents are radial deformations, while a gouge follows along a surface deformation. These defects usually occur due to construction inefficiency. Depending on the severity, this failure can lead to immediate or delayed failure (Panetta et al. 2001). A mechanical failure can also be a result of design error, use of inappropriate materials, and/or faulty construction (Davis et al. 2010).

2) *Corrosion Failure:* In pipelines, corrosion is a slow process, which leads to a loss of metal in the wall leading to a failure. According to the US Department of Transportation (DOT), corrosion is classified as the second-most frequent failure source in pipelines. This type of failure can further be classified into three main categories, namely, external, internal, or stress cracking corrosion (SCC).

- External corrosion**:** External corrosion is due to atmospheric or subsurface sources. External corrosion can be deferred using cathodic protection and pipeline coating (Muhlbauer 2004).
- Internal corrosion: This corrosion type, which targets pipeline inner surface, is usually a function of the transported material. Product corrodibility and corrosion intervention are the two main factors that affect internal corrosion (Muhlbauer 2004).
- Stress crack corrosion: This type of corrosion leads to cracking of material due to the combined actions of corrosion and tensile stress. This type of corrosion has two modes,

namely, Classical SCC and low-PH SCC. Longitudinal cracks would link to long-shallow flaws in both modes. The formation of this type of flaws is due to having a potent environment, steel pipe susceptibility to SCC and sufficient magnitude of tensile stress (Beavers and Thompson 2006).!

3) *Third-Party Activity:* Third-party failure is a result of damages caused by factors that are not associated with a pipeline. The US Department of Transportation (DOT) pipeline statistics determine third-party activities as the most frequent failure source in oil pipelines. However, third party damage is the least-considered factor in pipeline hazard assessment (Muhlbauer 2004). Different factors can affect the occurrence of third party damage including coating, cover depth, public education program, etc.

4) *Operational failure:* Operational failure results from operational upsets such as malfunction, or inadequacy of one or more safeguarding systems or operators' error (Bersani et al. 2010). This type of failure is rare even though it can cause catastrophic consequences. The application of safety devices in addition to monitoring pipeline pressure may defer operational failures (Muhlbauer 2004). Moreover, the probability of operational failures may be minimized using operational procedures, supervisory control and data acquisition (SCADA) communication, drug testing, etc.

5) *Natural hazards:* This failure type corresponds to natural disasters, which rarely take place. Examples of this failure include flooding, land movement, volcanic activity, and earthquakes. Geotechnical and hydro-technical studies are performed prior to pipeline construction to prevent this failure type.

***Logistic and multi nominal logistic regression***

Logistic regression uses binomial probability theory that predicts the probability of having a

response equal to either 0 or 1. The maximum likelihood method (logistic function) includes several computational iterations to maximize the probability of observed data (Burns and Burns 2008). For a binary output (Y=0 or Y=1) and multiple independent variables ($x_1$ to $x_n$), a logistic regression calculates the probability of y equal to 1. A linear regression is formulated using the following equation (Agresti 2007; Abdrabou 2012):

$$y = w_1x_1 + w_2x_2 + \cdots + w_nx_n \quad (1)$$

Where *y* = dependent variable, $w_1$ to $w_n$ = estimators, and $x_1$ to $x_n$ = independent variables. On the other hand, the logistic regression is formulated using the following equation:

$$p(y = 1) = \frac{1}{1+e^{-z}} \quad (2)$$

Where "*e*" = natural logarithm number and "*z*" = logit formulated.

The variable "z' is computed using the following equation:

$$z = (\text{Logit}) = w_1x_1 + w_2x_2 + \cdots + w_nx_n \quad (3)$$

From equations 2 and 3 it can be concluded that the logit is the log of odds, as presented in Equation 4 (Agresti 2007).

$$Logit = \ln\left(\frac{p(y=1)}{1+p(y=1)}\right) = w_1x_1 + w_2x_2 + \cdots + w_nx_n \quad (4)$$

The assumptions and limitations of logistic regression can be summarized as follows (Burns and Burns 2008);

- This regression type is not based on a linear relationship between the response and the independent variables.
- The response can undergo a dichotomy (two-phase) condition.
- The categories must be mutually exclusive.

- Large sample sizes are required.
- The independent variables cannot not be intervals.
- The independent variables cannot not be normally or linearly distributed.

A multinomial logit (MNL) model is an extension of logistic regression. The Dependent Variable (DV), which has many categories, has one value of its values set as a reference category. Then, the membership probability of the other categories of the DV is compared to the one corresponding to the reference category. For a DV with *M* categories, *M – 1* equations are required to present a relationship between the response and the explanatory variables (Agresti 2007).When there are more than two categories for a dependent variable, the following equation is used to calculate the outcome probability for *m=2* to *M*:

$$p(y_i = m) = \frac{\exp\,(z_{mi})}{1+\sum_{h=2}^{M} \exp\,(z_{hi})} \quad (5)$$

In addition, the probability for the reference category is computed using the following equation (Agresti 2007):

$$p(y = 1) = \frac{1}{(1+\sum_{h=2}^{M} e^{z_{hi}})} \quad (6)$$

Where "z" = Logit as defined in Equation 3.

In this research, the dependent variable (i.e., failure type) is nominal while all independent variables except pipeline age and diameter are discrete. Therefore, the basic form of logistic regression, i.e. multinomial logit (MNL), is applied for a binary response. MNL models can handle categorical variables. Moreover, a probability equation for each output category is obtained which provides a means to identify the failure that threatens the pipeline.

***Artificial neural networks (ANN)***

Over the past decade, artificial neural networks (ANNs) have been applied extensively in

engineering applications due to their ability to solve complex problems. ANN technique mimics the human brain's techniques for learning and recalling patterns. It is useful in problems where a solution is not clearly identified and the relationship between explanatory and response variables is not adequately defined (Al Barqawi 2007). Neurons are randomly connected in three main different layers, namely, input, hidden and output layers to form an artificial neural network. The hidden layers connect the input and the output layers and have no connection to the outside world (Zayed and Halpin 2005). Depending on the learning paradigms, there are three main types of neural networks, i.e. supervised, unsupervised and hybrid types. In a supervised network, the network is provided with a correct answer (output) for each pattern and the weights are assigned to help the network calculate outputs that are as close as possible to the real values. Unsupervised learning does not require the provision of an output to the network. Instead, it perceives the underlying correlations of data patterns and organizes these patterns into categories. Hybrid networks combine the supervised and unsupervised learning processes to provide part of the weights using supervised learning while the remaining weights are provided through unsupervised learning (Abdrabou 2012).

The most commonly applied neural network is the multi-layer feed (MLF) which is trained with a back propagation-learning algorithm (Svozil et al. 1997). Back propagation neural networks (BPNNs) learn by examples, which makes them effective in prediction (Al Barqawi 2007). A BPNN training algorithm is commonly used to supervise neural networks where the output is provided to instruct the ANN. The learning process of supervised ANNs using a BPNN learning algorithm is accomplished by providing the network with datasets that include both inputs and outputs to be trained. The network pattern is introduced and then the output pattern is estimated using random weights. The generated output is then compared to the actual value, and the corresponding error between the two outputs is backward propagated into the network to adjust

the connections weights. This procedure is repeated until an allowable error or the maximum number of epochs is reached, or any other stopping condition is satisfied (Al Barqawi 2007). This process is called back propagation since it involves taking this derivative and adding it to the corresponding weight starting from the output layer back to the input layer (Abraham 2005). Once the neural network is trained, it can be used to predict the output for other input patterns through the connection's weights calculated during the learning process.

**Data collection**

To develop the failure prediction models, failure records reported by CONCAWE in 2010 were collected. The reported failure records cover over 37 years of spillage data since 1971 and 35000 km of pipelines transporting 780 million $m^3$ of crude oil and petroleum products across Europe. In this research, the collected data corresponds to 467 spillage accidents for cross-country oil pipelines, which satisfy several criteria (Davis et al. 2010). The first criteria is that pipelines transport crude oil or petroleum product while the second one is that pipelines have a minimum length of 2 kilometers in the public domain. The third criterion is that pipelines run across the country, including short estuary or river crossings. On the other hand, the fourth criteria is that accidents cover pump stations and intermediate storage facilities but ignore terminal facilities and tank farms. Finally, the fifth criterion is that accidents involve a minimum spillage size of 1 $m^3$, unless exceptional safety or environmental consequences are considered (Abdrabou 2012).

To develop the proposed failure prediction models, the records related to explanatory and dependent variables were collected in two datasets. Dataset 1 includes 289 accident records whose failure type falls into the category of mechanical, corrosion, and third party. On the other hand, dataset 2 includes 225 accidents corresponding to only corrosion and third-party failures. Accidents due to operational and natural hazards failures were not collected in this research due to

their small number. The explanatory variables considered include pipeline age, diameter, service type, facility type, and land use. The collected data for each model variable is detailed in Table 1 (Abdrabou 2012).

i) Pipeline age: Pipeline age is one of the main factors with direct influence on corrosion failure. This variable values range between 1 and 40 years.

ii) Pipeline diameter: Pipelines with a relatively small diameter are more vulnerable to third party damages such as excavation. This variable values range between 1 and 60 inches

iii) Service type: The variable corresponds to the type of transferred product, i.e. crude oil, white product, or fuel oil.

iv) Facility type: The variable describes the vertical location of a pipeline, i.e. above-ground or underground.

v) Land use: The variable corresponds to the type of land use where the accident occurred. The records related to three land use types namely, residential, agricultural, and industrial/commercial were collected in this research.

**Failure model development**

Figure 3 presents the methodology used to develop failure prediction models of oil pipelines. Multinomial logistic regression and neural networks were used to treat categorical variables as nominal rather than continuous. The models were developed using historical failure records of cross-country oil pipelines found in CONCAWE database. Two main models, namely Models A and B, were developed. Model A predicts mechanical, corrosion, or third party failure types. On the other hand, Model B predicts corrosion and third party types failure types. More specifically, the models A.1 and A.2 predict the failure type using artificial neural network and multinomial

278 logit regression approaches, respectively. On the other hand, model B.1 predicts the failure type
279 using artificial neural network approach. To develop these models, the datasets 1 and 2 were
280 randomly divided into 80% and 20% for training and validation purposes, respectively.

281 ***ANN Model Development***

282 A back-propagation procedure was used to develop the models A.1 and B.1 using SPSS 19
283 statistical software with 289 and 225 records, respectively. The corresponding learning rate and
284 momentum of the neural network were chosen as 0.05 and 0.9, respectively. Moreover, Tanh
285 activation function was used between the input and hidden layers with a stopping criterion of 10
286 steps without any error decrease. Depending on the model, the network output can vary between
287 mechanical, corrosion and third party failures for model A.1 or corrosion and third party failures
288 for model B.1. On the other hand, the model input includes pipeline age, facility type, land use,
289 diameter, and service type as presented Table 1. One neuron is used to represent each subcategory
290 of each nominal variable (i.e., land use, facility and service type). On the other hand, each
291 continuous variable (age and diameter) is represented by one neuron. Model A.1 consists of an
292 input layer with 10 neurons, a hidden layer with 35 neurons, and an output layer with 3 neurons.
293 On the other hand, model B.1 consists of an input layer with 10 neurons, a hidden layer with 26
294 neurons, and an output layer with 2 neurons. The importance factor analysis results show that the
295 ranking of the model A.1 predictors (i.e., from the most to the least important) is: 1) service type,
296 2) diameter, 3) facility type, 4) land use, and 5) age. On the other hand, Table 2 shows that the
297 ranking of the model B.1 predictors (i.e., from the most to the least important) is: 1) service type,
298 2) age, 3) land use, 4) diameter, and 5) facility type. Moreover, the training results show that the
299 models A.1 and B.1 correctly classified 73.8% and 73.3% of the records, respectively. The area
300 under the ROC curve of the dependent output is another diagnostic measure for the ANN accuracy.
301 Table 3 shows that the area under the ROC curve for each dependent variable category in both

neural networks models is generally more than 0.75 for each output, which is an indication of an acceptable accuracy.

***Development of multinomial logit model***

As previously mentioned, model A.2 uses multinomial logit regression to predict mechanical, corrosion, or third party failures. The multinomial regression performs the analysis by computing the probability of occurrence for each failure type, where the highest probability is set as the predicted value. The logit model pairs each category to a baseline category. In this context, the last output category (third party failure) is set as the baseline category and the maximum likelihood estimate (MLE) method is used to measure the performance level of the MNL model. MLE is the value of the parameter that makes an observed data most likely. It is usually reported as the Log-likelihood or the initial Log-likelihood function that is equal to -2 Log-likelihood (-2LL) (Menard 2002). The initial likelihood function (-2 Log Likelihood) is a statistical measure similar to the total sum square in linear regression. Table 4 presents the initial likelihood function value of 498.091 for the model with no independent variables (constant only) and the initial likelihood value of 387.59 for the model with all the independent variables. The reduction of this value indicates the improvement in model's prediction due to the presence of independent variables. The difference between the two values is the Chi-Squared (101.49), which has a significance of less than 0.0001. Therefore, it can be concluded that there is a significant relationship between the explanatory and response variables (Menard 2002).

As diagnostic measures for multinomial regression models, several pseudo R squares are used to evaluate model's goodness of fit. These pseudo R-squares have a scale similar to that of $R^2$ in a linear regression. Cox and Snell Pseudo R-Square, McFadden Pseudo R-Square and Nagelkerke Pseudo R-Square are derived using equations 7 to 9 as diagnostic measures. Table 5 summarizes pseudo-square results.

$$R^2_{\text{Cox and Snell}} = 1 - \{\frac{\mathrm{L}(M_{\text{intercept}})}{\mathrm{L}(M_{\text{Full}})}\}^{2/N} \quad (7)$$

$$R^2_{McFadden} = 1 - \frac{lnL(M_{Full})}{ln\ L(M_{intercept})} \quad (8)$$

$$R^2_{Nagelkerke} = \frac{1-\{\frac{L(M_{intercept})}{L(M_{Full})}\}^{2/N}}{1-L(M_{intercept})^{2/N}} \quad (9)$$

Where, $M_{full}$= model with predictors; $M_{intercept}$= model without predictors; L=estimated likelihood; and N= number of observations

To identify the importance of each predictor, SPSS software calculates the initial likelihood value for the reduced model, which is the one that contains all predictors except the one under study. This likelihood is compared to the one achieved by the model when all predictors are present (full model). The Chi-square is then calculated for each model by subtracting the full model value from the reduced model value. The predictor with a high Chi-square and a low significance is considered as an important variable. Table 6 summarizes the SPSS analysis results, which show that the type of service is the most important variable while pipeline age is of the least important.

As discussed earlier, a multi nominal logit regression model is based upon the calculation of the probability of occurrence in each failure source. The logit of each dependent variable is calculated using equation 10 to compute the probability of each dependent variable.

$$Logit = Z = B + B_1X_1 + B_2X_2 + \cdots + B_nX_n \quad (8)$$

Where $X_1$ to $X_n$ = predictor values and B, $B_1$, …, $B_n$ = variable coefficients as shown in Table 7. The SPSS program generates the coefficients for outputs one and two. Therefore, the coefficient for output three (the reference category) can be calculated using the following equation:

$$P3 = 1- (P1+P2). \quad (11)$$

Finally, the Logit equations for each output are listed in equations 12, 13, 1nd 14.

$$Z1 = 2.005 + 0.0097D - 0.008A - 1.453S1 - 1.493S2 - 2.351F1 - 0.193L1 + 0.044L2 \quad (9)$$

$$Z2 = 3.949 + 0.008D + 0.016A - 3.358S1 - 4.5S2 - 0.307F1 - 1.213L1 + 1.823L2 \quad (10)$$

Z3 = 0 (Reference Category) (14)

Where, D = pipeline diameter in inches, A = pipeline age by year, S = type of service, F = facility and L = land use. The probability of occurrence of each failure source in oil pipelines are presented in the following equations:

$$P_1 = (Mechanical\ Failure) = \frac{e^{z1}}{e^{z1}+e^{z2}+e^{z3}} \quad (11)$$

$$P_2 = (Corrosion\ Failure) = \frac{e^{z2}}{e^{z1}+e^{z2}+e^{z3}} \quad (12)$$

$$P_3 = (Third\ Party\ \ Failure) = \frac{e^{z3}}{e^{z1}+e^{z2}+e^{z3}} \quad (13)$$

**Model validation**

The developed models were validated using the remaining 20% data. The input of the validation dataset was fed into the trained model to obtain the predicted failure type values. The average invalidity percentage (AIP) and the average validity percentage (AVP) for the testing dataset were computed using equations 18 and 19, respectively.

$$AIP = (\sum_{i=1}^{n} \left|1 - (\frac{E_i}{C_i})\right|)/n \quad (14)$$

$$AVP = 1 - AIP \quad (15)$$

Where, "$E_i$" = predicted value while "$C_i$" = actual value.

The validation results for models A.1, A.2 and B.1 are summarized in Table 8. Moreover, the validation outputs and actual failure sources are shown in Figure 4.

Table 8 shows that the percentage of correct predictions in all the models is approximately 70%, which is acceptable. The results obtained are more reliable than those published in similar studies that considered the dependent variable as numerical. Such consideration is highly important since failure source identification is a classification problem with discrete dependent variables. On the

other hand, some of the contributing factors in the failure of oil and gas pipelines are missing in the database including thickness, operating pressure, and yield strength. Such limitations can be a leading factor in reducing the prediction accuracy of the models. Moreover, access to large number of failure records is complicated in the oil and gas industry due to confidentiality matters.

In this study, the accuracy of all the three models is almost similar. Moreover, ANN and multinomial regression models exhibited identical performances. However, the prediction of the failure source using two failure sources (corrosion and third party) leads to better results compared to the one with three failure sources (corrosion, third-party and mechanical damage). The multinomial logit model was used herein to estimate the probability of occurrence for each major failure source that may target pipelines, given five pipeline attributes. By knowing this probability, the most probable failure source is identified through probability measures. Even though the validation results using neural network and multinominal logit regression were similar, the latter has the following two important advantages:

- MN model presents an equation, which can be used in practice.
- MNL model provides the probability of each failure source, which helps operators to have a better awareness of most critical failure sources.

**Sensitivity analysis**

A sensitivity analysis was carried out to examine the effect of variation in explanatory variables on the model response. An accident, where the diameter, service type, facility type, age, and land use were set equal to 11, 2, 1, 40 and 1, respectively, was selected. Next, each predictor under study was altered between its maximum and minimum values while keeping the other predictors constant. The procedure was repeated for each input variable. When there was no change of the output variable related to a particular input, other inputs were then maintained at a different value,

one at a time to study the correlation between inputs. The sensitivity analysis results for model A.1 almost confirmed the inputs' importance that was previously determined by the importance factor analysis where 'service' was the most sensitive predictor while 'age' had the smallest effect. The following subsection presents the sensitivity analysis results for this model.

***Service Type Variations:***

Figure 5(a) presents the effect of service type variations on failure causes. The results show that the failure source would be third party if the service type were white oil. However, if the service type changes to crude oil or hot products, the failure would originate from a corrosion source due to the associated heating and the presence of impurities.

***Facility Type Variations:***

Figure 5(b) shows the variations in failure source due to variations of facility location. If the pipeline is buried, the failure source is predicted to be third party. On other hand, the predicted failure source changes to mechanical failure for aboveground pipelines.

***Land Use Variations:***

For the selected accident record, no variations in the resulting failure source were observed due to land use variations. Therefore, other input variables were altered one at a time to identify the correlation between the inputs. Figure 5(c) shows a significant correlation between land use and age, diameter, and service.

***Diameter Variations:***

Figure 6(a) represents the effects of diameter variation on the failure source. The pipelines with small diameters were vulnerable to third party failure. On the other hand, a variation in failure source occurred when the pipeline diameter increased.

***Age Variations:***

Figure 6(b) shows the impact of age variations on failure source. Variation of pipeline age did not alter the failure source. However, Figure 6(c) shows that for larger diameter (i.e. 30 inches) pipelines, failure source variations occurred due the age variation. In other words, pipeline age has more impact on failure source for pipelines with larger diameters.

## Conclusions

This research proposes several models for failure source prediction in oil pipelines. These models were built using historical data from real failure incidents. They can help oil pipeline operators and decision makers to plan for the actions required to maintain pipeline safety by estimating pipeline failure sources. These models predict the source of failure targeting a pipeline based on specific physical, operational, and environmental characteristics. The first two models (A.1 and B.1) were developed using ANN technique. Model A.1 was able to predict three failure sources, namely, mechanical, corrosion, third party, with an accuracy of 68.5%. On the other hand, model B.1 was able to predict two failure sources, namely, corrosion or third party, with an accuracy of 72.2%. This marginal accuracy increase may be due to the inclusion of mechanical failures in the prediction model. Therefore, model A.1 is more beneficial, as it can predict failure among three different failure sources, which count for more than 87% of accidents in oil pipelines.

An MNL approach was also used to develop the third model (model A.2) for the prediction of three failure sources with a prediction accuracy of 68.4%. It is worth noting that the results obtained using model A.2 are close to those obtained from using model A.1. The MNL model calculates the probability of occurrence for each failure source, which can be helpful for decision makers in identification of the most probable and critical failure sources.

The product type and pipeline age have the most and the least impact on the failure source,

respectively. The developed models will help pipeline operators avoid performing excessive inspections and prioritize them. These models provide a clear view of the failure sources that threaten a pipeline, allowing decision makers to take the actions required to mitigate risks and maintain the pipeline in a safe condition.

## References


Abdrabou, B. (2012). "Failure Prediction Model for Oil Pipelines". Master of science. Concordia University, Montreal, Canada.

Abraham, A. (2005). "Rule-Based expert systems, *Handbook of Measuring System Design*." *John Wiley & Sons,* .

Agresti, A. (2007). ***An Introduction to Categorical Data Analysis***. Wiley-Interscience, Hoboken, New Jersey, USA.

Al Barqawi, H. (2007). "Condition rating models for underground infrastructure : sustainable water mains.". Concordia University, Montreal, Canada.

Beavers, J. A., and Thompson, N. G. (2006). "External corrosion of oil and natural gas pipelines." ASM International Materials Park, Ohio, USA, 1015-1025.

Bersani, C., Citro, L., Gagliardi, R., Sacile, R., and Tomasoni, A. (2010). "Accident occurrance evaluation in the pipeline transport of dangerous goods." *Chemical Engineering Transactions,* 249-254.

Bertolini, M., and Bevilacqua, M. (2006). "Oil pipeline spill cause analysis: A classification tree approach." *Journal of Quality in Maintenance Engineering,* 12(2), 186-198.

Bertuccio, I., and Moraleda, M. B. (2012). "Risk assessment of corrosion in oil and gas pipelines using fuzzy logic." *Corrosion Engineering, Science and Technology,* 47(7), 553-558.

Burns, B., and Burns, R. (2008). *Business research methods and statistics using SPSS.* SAGE, London.

Caleyo, F., Velázquez, J., Valor, A., and Hallen, J. (2009). "Probability distribution of pitting corrosion depth and rate in underground pipelines: A Monte Carlo study." *Corrosion Science,* 51(9), 1925-1934.

Davis, M., Dubois, J., Gambardella, F., and Uhlig, F. (2010). "Performance of European cross-country oil pipelines: Statistical summary of reported spillages in 2008 and since 1971." CONCAWE Oil Pipelines Management Group, Special Task Force, Brussels.

Dey, P., Ogunlana, S., and Naksuksakul, S. (2004). "Risk-based maintenance model for offshore oil and gas pipelines: a case study." *Journal of Quality in Maintenance Engineering,* 10(3), 169-183.

Dundulis, G., Žutautaitė, I., Janulionis, R., Ušpuras, E., Rimkevičius, S., and Eid, M. (2016). "Integrated failure probability estimation based on structural integrity analysis and failure data: Natural gas pipeline case." *Reliability Engineering and System Safety,* 156 195-202.

Hasan, A. (2016). "Security of Cross-Country Oil and Gas Pipelines: A Risk-Based Model." *Journal of Pipeline Systems Engineering and Practice,* 04016006.

Kabir, G., Sadiq, R., and Tesfamariam, S. (2016). "A fuzzy Bayesian belief network for safety assessment of oil and gas pipelines." *Structure and Infrastructure Engineering,* 12(8), 874-889.

Li, J., Zhang, H., Han, Y., and Wang, B. (2016). "Study on failure of third-party damage for urban gas pipeline based on fuzzy comprehensive evaluation." *PLoS One,* 11(11), e0166472-1.

Li, X., Yu, R., Zeng, L., Li, H., and Liang, R. (2009). "Predicting corrosion remaining life of underground pipelines with a mechanically-based probabilistic model." *Journal of Petroleum Science and Engineering,* 65(3), 162-166.

Liao, K., Yao, Q., Wu, X., and Jia, W. (2012). "A numerical corrosion rate prediction method for direct assessment of wet gas gathering pipelines internal corrosion." *Energies,* 5(10), 3892-3907.

Ma, B., Shuai, J., Liu, D., and Xu, K. (2013). "Assessment on failure pressure of high strength pipeline with corrosion defects." *Engineering Failure Analysis,* 32 209-219.

Menard, S. (2002). *Applied logistic regression analysis.* SAGE publications, California, US.

Muhlbauer, W. K. (2004). *Pipeline risk management manual: ideas, techniques, and resources.* Gulf Professional Publishing, Amsterdam.

Omidvar, B., and Kivi, H. K. (2016). "Multi-hazard failure probability analysis of gas pipelines for earthquake shaking, ground failure and fire following earthquake." *Natural Hazards,* 82(1), 703-720.

Panetta, D., Diaz, A., Pappas, A., Taylor, T., Francini, B., and Johnson, I. (2001). "Mechanical damage characterization in pipelines." *Rep. No. DE-AC06-76RLO1830,* Pacific Northwest National Laboratory, Operated by Battelle for the United States Department of Energy, .

Papavinasam, S., Doiron, A., and Revie, R. W. (2010). "Model to predict internal pitting corrosion of oil and gas pipelines." *Corrosion,* 66(3), 035006-035006-11.

Parvizsedghy, L., and Zayed, T. (2015a). "Consequence of Failure: Neurofuzzy-Based Prediction Model for Gas Pipelines." *Journal of Performance of Constructed Facilities,* 30(4), 04015073.1-04015073.10.

Parvizsedghy, L., and Zayed, T. (2015b). "Developing failure age prediction model of hazardous liquid pipelines." *International Construction Specialty Conference* , Canadian society of civil engineers (CSCE), Vancouver, BC., Canada, 285.1-285.10.

Parvizsedghy, L., and Zayed, T. (2013). "Failure prediction model of oil and gas pipelines." *14th International Conference on Civil, Structural and Environmental Engineering Computing*, Civil-Comp Press, Cagliari, Sardinia, Italy, 1.

Restrepo, C. E., Simonoff, J. S., and Zimmerman, R. (2009). "Causes, cost consequences, and risk implications of accidents in US hazardous liquid pipeline infrastructure." *International Journal of Critical Infrastructure Protection,* 2(1), 38-50.

Senouci, A., Elabbasy, M., Elwakil, E., Abdrabou, B., and Zayed, T. (2014). "A model for predicting failure of oil pipelines." *Structure and Infrastructure Engineering,* 10(3), 375-387.

Senouci, A., El-Abbasy, M., and Zayed, T. (2014). "Fuzzy-based model for predicting failure of oil pipelines." *Journal of Infrastructure Systems,* 20(4), 04014018.1-04014018.11.

Svozil, D., Kvasnicka, V., and Pospichal, J. (1997). "Introduction to multi-layer feed-forward neural networks." *Chemometrics and Intelligent Laboratory Systems,* 39(1), 43-62.

Wen, K., Gong, J., Chen, F., and Liu, Y. (2014). "A Framework for Calculating the Failure Probability of Natural Gas Pipeline." *Journal of Computer Science Technology Updates,* 1(1), 1-8.

Witek, M. (2016). "Gas transmission pipeline failure probability estimation and defect repairs activities based on in-line inspection data." *Engineering Failure Analysis,* 70 255-272.

Yuhua, D., and Datao, Y. (2005). "Estimation of failure probability of oil and gas transmission pipelines by fuzzy fault tree analysis." *Journal of Loss Prevention in the Process Industries,* 18(2), 83-88.

Zayed, T., and Halpin, D. (2005). "Pile Construction Productivity Assessment." *Journal of Construction Engineering and Management,* 131(6), 705-714.

Zhou, Q., Wu, W., Liu, D., Li, K., and Qiao, Q. (2016). "Estimation of corrosion failure likelihood of oil and gas pipeline based on fuzzy logic approach." *Engineering Failure Analysis,* 70 48-55.

534

535 **List of Figures:**



542

543 **List of Tables:**



552

Table 1. Data collected for model development

| Variable | Type | Unit | Scale/subcategories |
|---|---|---|---|
| **age** | Continuous | Year | 1 to 40 |
| **diameter** | Continuous | Inch | 1 to 60 |
| **Land use** | Nominal | - | residential<br>agricultural<br>Industrial/commercial |
| **Facility** | Nominal | - | aboveground<br>underground |
| **Service** | Nominal | - | crude oil<br>white product<br>fuel oil (HOT) |
| **Failure type** | Nominal | - | Mechanical<br>Corrosion<br>Third-party |

Table 2. Importance analysis of predictors in ANN

| | **Importance** | | **Normalized Importance** | |
|---|---|---|---|---|
| **Model Inputs** | **Model 1.A** | **Model 1.B** | **Model 1.A** | **Model 1.B** |
| **Service** | 0.27 | 0.418 | 100% | 100% |
| **Facility** | 0.193 | 0.93 | 71.70% | 22.2% |
| **Land use** | 0.142 | 0.166 | 52.80% | 39.8% |
| **Diameter** | 0.263 | 0.149 | 97.40% | 35.7% |
| **Age** | 0.132 | 0.174 | 48.90% | 41.8% |

563 Table 3. Area under ROC curve for ANN models

| | Area under ROC | |
|---|---|---|
| **Output categories** | **Model 1. A** | **Model 1. B** |
| **1-Mechanical Failure** | 0.75 | - |
| **2-Corrosion Failure** | 0.79 | 0.84 |
| **3-Third-Party Failure** | 0.78 | 0.84 |

564

565

566 Table 4. Fitting information for model A.2

| **Model** | **-2 Log Likelihood** | **Chi-Square** | **Sig.** |
|---|---|---|---|
| **Intercept Only** | 489.091 | N/A | N/A |
| **Final** | 387.592 | 101.499 | 0.000 |

567

568

569 Table 5. Pseudo R-Square Values for model B.1

| **Pseudo R-Square** | **Values** |
|---|---|
| Cox and Snell | 0.354 |
| Nagelkerke | 0.42 |
| McFadden | 0.205 |

570

571

572

573 Table 6. Likelihood Ratio Test for Model A.2

| Effect | -2 LL of Reduced Model | Chi-square | Significance |
|---|---|---|---|
| Intercept | 387.592 | 0 | . |
| Diameter | 399.623 | 12.031 | 0.02 |
| Age | 389.872 | 2.279 | 0.320 |
| Service | 436.608 | 49.016 | 0.000 |
| Facility | 404.405 | 16.813 | 0.000 |
| Land Use | 398.546 | 10.954 | 0.027 |

574

575

576 Table 7. Variable Coefficients in model A.2

| Variables | Coefficients (Output1) | Coefficients (Output2) |
|---|---|---|
| Intercept | 2.005 | 3.949 |
| Diameter | 0.097 | 0.008 |
| Age | -0.008 | 0.016 |
| Service 1 | -1.453 | -3.358 |
| Service 2 | -1.493 | -4.5 |
| Facility 1 | -2.351 | -0.307 |
| Land Use 1 | -0.193 | -1.213 |
| Land Use 2 | 0.044 | -1.823 |

577

578

579 Table 8. Results of the validation phase

| model | Correct Prediction (%) | AVP (%) | AIP (%) |
|---|---|---|---|
| A.1 | 68.5 | 73.7 | 26.3 |
| A.2 | 72.2 | 72.8 | 27.2 |
| B.1 | 68.5 | 73.7 | 26.3 |

580