E-ISSN: 2997-9382
American Journal of Technology Advancement
https://semantjournals.org/index.php/AJTA

| Research Article

Check for updates

Predicting Failure in Crude Oil Pipelines Using First Degree Polynomial Model and Reduced Second Degree Polynomial Model, Validating the Models Using Goodness of Fit Indices

Igbesi Friday Chukwuyem, Colllins Tobore Atumah
Department of Mechanical Engineering, Delta State Polytechnic, Otefe-Oghara

Abstract

To mitigate these risks, pipelines require constant monitoring and maintenance to detect and rectify defects such as corrosion before they lead to failure. However, the regular monitoring and non-destructive testing of pipelines incur substantial costs. Consequently, there is a growing interest in research focused on predictive corrosion monitoring of pipelines based on easily measurable operational parameters. This study aims to employ a model-based condition monitoring and failure avoidance approach to predict failures in crude oil pipelines. Secondary data on mean corrosion rates, mean pH levels, mean temperatures, mean pressures, and mean aqueous CO2 partial pressures were collected from an oil and gas multinational company spanning the years 2007 to 2011. Polynomial regression models were developed. The models' validity was assessed using Goodness of Fit Indices (GFI).

For the first-degree polynomial model, the following results were obtained for both training and testing datasets: Coefficient of Determination (R2) = 0.3022/0.2948, Root Mean Square Error (RMSE) = 0.0050/0.0040, Mean Biased Error (MBE) = 0.0000/0.00010, Mean Absolute Biased Error (MABE) = 0.0037/0.0026, Mean Percentage Error (MPE) = 0.0320/0.3633, and correlation coefficient (r) = 0.5497/0.5891. Furthermore, the goodness of fit for the reduced second-degree model for both training and testing datasets provided the following results: (R2) = 0.9859/0.9341, (RMSE) = 0.0007/0.0012, (MBE) = 0.0000/0.0001, (MABE) = 0.0004/0.0008, (MPE) =

0.0006/0.0390, and (r) = 0.9929/0.9676. This research contributes valuable insights into the technological applications and policy implications of predictive corrosion monitoring for pipelines.

This is an open-access article under the CC-BY 4.0 license

1. INTRODUCTION

The progressive deterioration of pipelines over time stems from their exposure to natural forces in their surroundings and the dynamic pressures exerted by the transported fluids. If left unaddressed, this degenerative process culminates in failure, accompanied by grave consequences, encompassing environmental contamination, health hazards, and substantial economic losses (El-Abbasy *et al*., 2015). The repercussions of pipeline failures extend to encompass severe economic,

social, and environmental devastations, resulting in substantial repair expenses, widespread disruptions to daily life, extensive environmental pollution, and even human injuries. The extensive network of pipelines that crisscrosses Nigeria, as highlighted by Ekpu and Ehighelua (2004), serves as the primary conduit for transporting the majority of the nation's petroleum products, covering a vast expanse of approximately 3106 miles. Consequently, the issue of pipeline failures, stemming from factors such as corrosion, material defects, third-party activities, mechanical damage, and environmental factors, has been extensively documented within the Nigerian context. Furthermore, when compared to developed nations, Nigeria has experienced an unacceptably high rate of pipeline failures in recent years. This research endeavor represents a significant step in this direction, focusing on model-based maintenance strategies to minimize the risks of pipeline failure and its associated catastrophic consequences. It is crucial to highlight that pipeline failures not only result in significant economic losses but also pose threats to human lives and the environment. To put the gravity of pipeline failures into perspective, one can recall the tragic incident in Lagos, Nigeria, in 2006, where a pipeline explosion claimed the lives of over two hundred people (Dawotola, 2012). Such harrowing incidents have left indelible marks on the histories of many other nations as well. Addressing this issue is paramount to safeguard lives, protect the environment, and promote sustainable development. Consequently, employing these model-based methods presents the most viable avenue for implementing condition monitoring and failure control across the extensive oil and gas pipeline network in Nigeria. The aim of this research is to further enhance the predictive accuracy of state-of-the-art methodologies and adapt the emerging new techniques to the Nigerian context, thereby establishing a robust framework for addressing the critical issue of pipeline failures in the country. The primary challenge in achieving optimal and efficient pipeline integrity management lies in ensuring effective and comprehensive inspection, which can significantly increase inspection costs. Traditional in-line inspection techniques, such as magnetic flux leakage and ultrasound, are not only prohibitively expensive but also timeconsuming. Previous research efforts have explored model-based approaches to address the issue of revenue loss (Sinha and Pandey, 2002). In light of this, our research aims to leverage a model-based approach to mitigate both revenue loss and safety concerns in pipeline oil management. This approach is predicated on the notion that by making decisions regarding inspection, repair, or replacement based on model-derived assessments of pipeline condition, we can reduce revenue loss by directing attention to the appropriate sections at the right time, thereby preempting failures. The twin challenges of revenue loss, which concerns investors, and safety, which is a concern for investors, the government, and the public, serve as the primary motivators for this study (Stephens and Nessim, 2006; De Silva *et al*., 2006; Maes *et al*., 2008; Thacker, 2010; Li and Mohamoodian, 2013; Mohamoodian and Li, 2015).

2.0 METHODOLOGY

The techniques that are applied in this work for data-driven modelling of corrosion rate are described in detail. The methods include the following: secondary data on mean PH, mean temperature, mean pressure, mean aqueous CO_2 partial pressure and mean corrosion rate were obtained **2.1. Data Collection**

The secondary corrosion data such as mean PH, mean temperature, mean pressure, mean aqueous CO_2 and mean corrosion rate were obtained from an oil and gas multinational company in Nigeria. However, for data privacy and confidentiality reason, the company name is not provided. The data set were collected from 2007 to 2011.

2.2. Method of Data Analysis and Model

The methods of data analysis used in this work are Polynomial Regression.

2.3. Polynomial Regression Analysis (PRA)

The much more advanced generalized approach which cover general-order polynomial analyses is effectively handled on the framework of vector/matrix algebra. For such analysis the response is generally related to the predictors as follows (Levin, 1998);

$$oy(\boldsymbol{x}) = [\boldsymbol{a}(\boldsymbol{x})]^T\{\sum_{i=1}^{n}\boldsymbol{a}(\boldsymbol{x}_i)[\boldsymbol{a}(\boldsymbol{x}_i)]^T\}^{-1}\sum_{i=1}^{n}\boldsymbol{a}(\boldsymbol{x}_i)y_i + e \quad (1)$$

Where: $\boldsymbol{a}(\boldsymbol{x}) = \{a_1(x)a_2(x) \dots a_r(x)\}^T$ is the polynomial of basis vectors

$a_j(x)$ represent any of all the possible interactions of the independent variables from zero to order p.

n is the number of experimental/sampling runs d is the number of independent variables.

The length r of $\boldsymbol{a}(x)$ is generally given as $r = \frac{(d+p)!}{d!p!}$. In the polynomial model, the numerical term

$$\boldsymbol{b} = \{\sum_{i=1}^{n}\boldsymbol{a}(\boldsymbol{x}_i)[\boldsymbol{a}(\boldsymbol{x}_i)]^T\}^{-1}\sum_{i=1}^{n}\boldsymbol{a}(\boldsymbol{x}_i)y_i \quad (2)$$

represent the vector of the model coefficients in terms of the data sets of the predictors and the targets. A good understanding of how to handle equation (2) is all that is need in polynomial regression. For illustration with some of the simplest multiple linear regression cases, consider two predictors, x_1 and x_2.

For the first degree (1stPRA) case

$$\boldsymbol{a}(\boldsymbol{x}) = \begin{Bmatrix} 1 \\ x_1 \\ x_2 \end{Bmatrix} \quad (3)$$

and

$$\boldsymbol{b} = \left[\sum_{i=1}^{n}\begin{Bmatrix} 1 \\ x_{1i} \\ x_{2i} \end{Bmatrix}\{1\ x_{1i}\ x_{2i}\}\right]^{-1}\sum_{i=1}^{n}\begin{Bmatrix} 1 \\ x_{1i} \\ x_{2i} \end{Bmatrix}y_i \quad (4)$$

where the summation signs in equation (3.4) apply element-by-element to the matrix $\begin{Bmatrix} 1 \\ x_{1i} \\ x_{2i} \end{Bmatrix}\{1\ x_{1i}\ x_{2i}\}$ and the vector $\begin{Bmatrix} 1 \\ x_{1i} \\ x_{2i} \end{Bmatrix}y_i$. Therefore

$$\boldsymbol{b} = \begin{bmatrix} \sum_{i=1}^{n}1 & \sum_{i=1}^{n}x_{1i} & \sum_{i=1}^{n}x_{2i} \\ \sum_{i=1}^{n}x_{1i} & \sum_{i=1}^{n}x_{1i}^2 & \sum_{i=1}^{n}x_{1i}x_{2i} \\ \sum_{i=1}^{n}x_{2i} & \sum_{i=1}^{n}x_{1i}x_{2i} & \sum_{i=1}^{n}x_{2i}^2 \end{bmatrix}^{-1}\begin{Bmatrix} \sum_{i=1}^{n}y_i \\ \sum_{i=1}^{n}x_{1i}y_i \\ \sum_{i=1}^{n}x_{2i}y_i \end{Bmatrix} \quad (5)$$

For the quadratic case

$$\boldsymbol{a}(\boldsymbol{x}) = \begin{Bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_1x_2 \\ x_2^2 \end{Bmatrix} \quad (6)$$

and

$$\boldsymbol{b} = \left[\sum_{i=1}^{n} \begin{Bmatrix} 1 \\ x_{1i} \\ x_{2i} \\ x_{1i}^2 \\ x_{1i}x_{2i} \\ x_{2i}^2 \end{Bmatrix} \{1\ x_{1i}\ x_{2i}\ x_{1i}^2\ x_{1i}x_{2i}\ x_{2i}^2\} \right]^{-1} \sum_{i=1}^{n} \begin{Bmatrix} 1 \\ x_{1i} \\ x_{2i} \\ x_{1i}^2 \\ x_{1i}x_{2i} \\ x_{2i}^2 \end{Bmatrix} y_i \tag{7}$$

This becomes

$$\boldsymbol{b} = \begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{2i}^2 \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{1i}^3 & \sum_{i=1}^{n} x_{1i}^2x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i}^2 \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{2i}^2 & \sum_{i=1}^{n} x_{1i}^2x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i}^2 & \sum_{i=1}^{n} x_{2i}^3 \\ \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}^3 & \sum_{i=1}^{n} x_{1i}^2x_{2i} & \sum_{i=1}^{n} x_{1i}^4 & \sum_{i=1}^{n} x_{1i}^3x_{2i} & \sum_{i=1}^{n} x_{2i}^2x_{2i}^2 \\ \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{1i}^2x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i}^2 & \sum_{i=1}^{n} x_{1i}^3x_{2i} & \sum_{i=1}^{n} x_{1i}^2x_{2i}^2 & \sum_{i=1}^{n} x_{2i}^3x_{2i} \\ \sum_{i=1}^{n} x_{2i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i}^2 & \sum_{i=1}^{n} x_{2i}^3 & \sum_{i=1}^{n} x_{2i}^2x_{2i}^2 & \sum_{i=1}^{n} x_{2i}^3x_{2i}^2 & \sum_{i=1}^{n} x_{2i}^4 \end{bmatrix}^{-1} \begin{Bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1i}y_i \\ \sum_{i=1}^{n} x_{2i}y_i \\ \sum_{i=1}^{n} x_{1i}^2y_i \\ \sum_{i=1}^{n} x_{1i}x_{2i}y_i \\ \sum_{i=1}^{n} x_{2i}^2y_i \end{Bmatrix} \tag{7}$$

Many regression models have more than two predictor variables. For such systems, it can be inferred from the foregoing that the response in linear regression is simply represented in the form of simple summation as follows;

$$y = a_0 + \sum_{i=1}^{d} a_i x_i + e \tag{8}$$

Where: a_i are the coefficients, x i is the ith of d predictors e is the error of the analysis which is assumed to be normally distributed.

Excluding the error and substituting n measured/sampled data points in equation (8) becomes

$$\begin{aligned} y^{(1)} &= a_0 + \sum_{i=1}^{d} \mathrm{a}_i x_i^{(1)} \\ y^{(2)} &= a_0 + \sum_{i=1}^{d} \mathrm{a}_i x_i^{(2)} \\ &\vdots \\ &\vdots \\ y^{(n)} &= a_0 + \sum_{i=1}^{d} \mathrm{a}_i x_i^{(n)} \end{aligned} \tag{9}$$

On minimizing the sum of square errors between the left- and right-hand-side of equation (3.9), the normal equations

$$\sum_{j=1}^{n} y^{(j)} = na_0 + \sum_{i=1}^{d} a_i \sum_{j=1}^{n} x_i^{(j)} \tag{10a}$$

$$\sum_{j=1}^{n} x_1^{(j)} y^{(j)} = a_0 \sum_{j=1}^{n} x_1^{(j)} + \sum_{i=1}^{d} a_i \sum_{j=1}^{n} x_1^{(j)} x_i^{(j)} \tag{10b}$$

$$\sum_{j=1}^{n} x_2^{(j)} y^{(j)} = a_0 \sum_{j=1}^{n} x_2^{(j)} + \sum_{i=1}^{d} a_i \sum_{j=1}^{n} x_2^{(j)} x_i^{(j)} \tag{10c}$$

$$\vdots$$

$$\sum_{j=1}^{n} x_d^{(j)} y^{(j)} = a_0 \sum_{j=1}^{n} x_d^{(j)} + \sum_{i=1}^{d} a_i \sum_{j=1}^{n} x_d^{(j)} x_i^{(j)} \tag{10d}$$

are generated. These normal equations are solved simultaneously to give the coefficients a_0 and a_i.

The general quadratic model for the problem deriving from equation (3.1) reads

$$y = a_0 + \sum_{i=1}^{d} a_i x_i + \sum_{i=d+1}^{2d} a_{i,i} x_i^2 + \sum_{i \neq k} a_{i,k} x_i x_k \quad (11)$$

Substituting n measured/sampled data points in equation (3.8) becomes

$$\begin{aligned} y^{(1)} &= a_0 + \sum_{i=1}^{d} a_i x_i^{(1)} + \sum_{i=d+1}^{2d} a_{i,i} x_i^{(1)2} + \sum_{i \neq k} a_{i,k} x_i^{(1)} x_k^{(1)} \\ y^{(2)} &= a_0 + \sum_{i=1}^{d} a_i x_i^{(2)} + \sum_{i=d+1}^{2d} a_{i,i} x_i^{(2)2} + \sum_{i \neq k} a_{i,k} x_i^{(2)} x_k^{(2)} \\ &\vdots \\ &\vdots \\ y^{(n)} &= a_0 + \sum_{i=1}^{d} a_i x_i^{(n)} + \sum_{i=d+1}^{2d} a_{i,i} x_i^{(n)2} + \sum_{i \neq k} a_{i,k} x_i^{(n)} x_k^{(n)} \end{aligned} \quad (12)$$

Also, for the quadratic (2ndPRA) case, on minimizing the sum of square errors between the left- and right-hand-side of equation (3.12), the normal equations

$$\sum_{j=1}^{n} y^{(j)} = n a_0 + \sum_{i=1}^{d} a_i \sum_{j=1}^{n} x_i^{(j)} + \sum_{i=d+1}^{2d} a_{i,i} \sum_{j=1}^{n} x_i^{(j)2} + \sum_{i \neq k} a_{i,k} \sum_{j=1}^{n} x_i^{(j)} x_{kj)} \quad (13a)$$

$$\sum_{j=1}^{n} x_1^{(j)} y^{(j)} = a_0 \sum_{j=1}^{n} x_1^{(j)} + \sum_{i=1}^{d} a_i \sum_{j=1}^{n} x \qquad {}_1^{(j)} x_i^{(j)} + \sum_{i=d+1}^{2d} a_{i,i} \sum_{j=1}^{n} x_1^{(j)} x_i^{(j)2} + \sum_{i \neq k} a_{i,k} \sum_{j=1}^{n} x_1^{(j)} x_i^{(j)} x_k^{(j)} \quad (13b)$$

$$\sum_{j=1}^{n} x_2^{(j)} y^{(j)} = a_0 \sum_{j=1}^{n} x_2^{(j)} + \sum_{i=1}^{d} a_i \sum_{j=1}^{n} x \qquad {}_2^{(j)} x_i^{(j)} + \sum_{i=d+1}^{2d} a_{i,i} \sum_{j=1}^{n} x_2^{(j)} x_i^{(j)2} + \sum_{i \neq k} a_{i,k} \sum_{j=1}^{n} x_2^{(j)} x_i^{(j)} x_k^{(j)}$$

(13c):

$$\sum_{j=1}^{n} x_d^{(j)} y^{(j)} = a_0 \sum_{j=1}^{n} x_d^{(j)} + \sum_{i=1}^{d} a_i \sum_{j=1}^{n} x_d^{(j)} x_i^{(j)} + \sum_{i=d+1}^{2d} a_{i,i} \sum_{j=1}^{n} x_d^{(j)} x_i^{(j)2} + \sum_{i \neq k} a_{i,k} \sum_{j=1}^{n} x_d^{(j)} x_i^{(j)} x_k^{(j)}$$

(13d) are generated and solved simultaneously to give the coefficients a_0 a_i $a_{i,i}$, , and $a_{i,k}$.

For four-predictor system, which specially apply to this study, this is expanded to read

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + a_{1,2} x_1 x_2 + a_{1,3} x_1 x_3 + a_{1,4} x_1 x_4 + \quad a_{2,3} x_2 x_3 + a_{2,4} x_2 x_4 + a_{3,4} x_3 x_4 + a_{1,1} x_1^2 + a_{2,2} x_2^2 + a_{3,3} x_3^2 + a_{4,4} x_4^2 \quad (14)$$

Then, the arising normal equations are

$$\sum_{j=1}^{n} y^{(j)} = a_0 n + a_1 \sum_{j=1}^{n} x_1^{(j)} + a_2 \sum_{j=1}^{n} x_2^{(j)} + a_3 \sum_{j=1}^{n} x_3^{(j)} + a_4 \sum_{j=1}^{n} x_4^{(j)} + a_{1,2} \sum_{j=1}^{n} x_1^{(j)} x_2^{(j)} + a_{1,3} \sum_{j=1}^{n} x_1^{(j)} x_3^{(j)} + a_{1,4} \sum_{j=1}^{n} x_1^{(j)} x_4^{(j)} + a_{2,3} \sum_{j=1}^{n} x_2^{(j)} x_3^{(j)} + a_{2,4} \sum_{j=1}^{n} x_2^{(j)} x_4^{(j)} + a_{3,4} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_3^{(j)} x_4^{(j)} + a_{1,1} \sum_{j=1}^{n} x_1^{(j)2} + a_{2,2} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_2^{(j)2} + a_{3,3} \sum_{j=1}^{n} x_3^{(j)2} + a_{4,4} \sum_{j=1}^{n} x_4^{(j)2}$$

$$\sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} y^{(j)} = a_0 \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} + a_1 \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_1^{(j)} + a_2 \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_2^{(j)} + a_3 \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_3^{(j)} + a_4 \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_4^{(j)} + a_{1,2} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_1^{(j)} x_2^{(j)} + a_{1,3} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_1^{(j)} x_3^{(j)} + a_{1,4} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_1^{(j)} x_4^{(j)} + a_{2,3} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_2^{(j)} x_3^{(j)} + a_{2,4} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_2^{(j)} x_4^{(j)} + a_{3,4} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_3^{(j)} x_4^{(j)} + a_{1,1} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_1^{(j)2} + a_{2,2} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_2^{(j)2} + a_{3,3} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_3^{(j)2} + a_{4,4} \sum_{j=1}^{n} x_i^{(j)} x_k^{(j)} x_4^{(j)2}$$

(15)

2.4 Goodness of Fit Indices (GFI)

Error is included in equation 1) to highlight the expected error in empirical modelling. The close to zero the error is, the better the model is said to be. The closeness of error to zero is judged using goodness of fit indices.

In this thesis, the predicted responses are compared to the targets using the following statistical indices; the Coefficient of Determination (R^2), the Root Mean Square Error (RMSE), the Mean Biased Error (MBE), the Mean Absolute Biased Error (MABE), the Mean percentage error (MPE), and Correlation coefficient (r). The correlation is a measure of how the response targets as presented in Figure 3.3 agree with the response predictions.

The expression for these indices are

$$R^2 = 1 - \left(\sum_{n=1}^{N}(T_n - \bar{y})^2\right)^{-1} \sum_{n=1}^{N}(T_n - y_n)^2 \quad (16)$$

$$\text{RMSE} = \left(\frac{1}{N}\sum_{n=1}^{N}(y_n - T_n)^2\right)^{0.5} \quad (17)$$

$$\text{MBE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - T_n) \quad (18)$$

$$\text{MABE} = \frac{1}{N}\sum_{n=1}^{N}|(y_n - T_n)| \quad (19)$$

$$\text{MPE} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{y_n - T_n}{T_n}\right)100 \quad (20)$$

$$r = \left(\sum_{n=1}^{N}(y_n - \bar{y})^2 \sum_{n=1}^{N}(T_n - T)^2\right)^{-\frac{1}{2}} \sum_{n=1}^{N}(y_n - \bar{y})(T_n - \bar{T}) \quad (21)$$

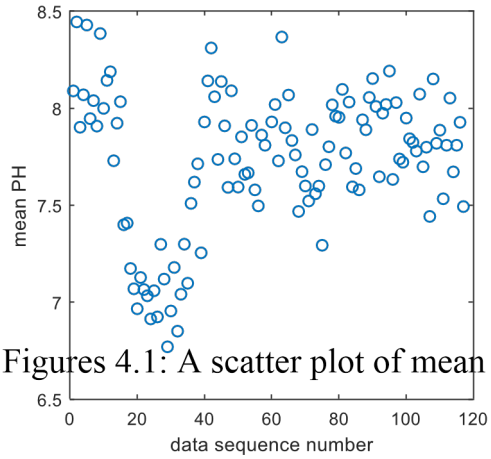
where: T represents a target response y represents the predicted response.

The closer to unity R^2 is, the better the predictions. The smaller the RMSE, MBE, MABE and MPE, the more accurate the model. The correlation coefficient r is, better when it is closer to one since it measures the linearity between the predicted and the target responses.

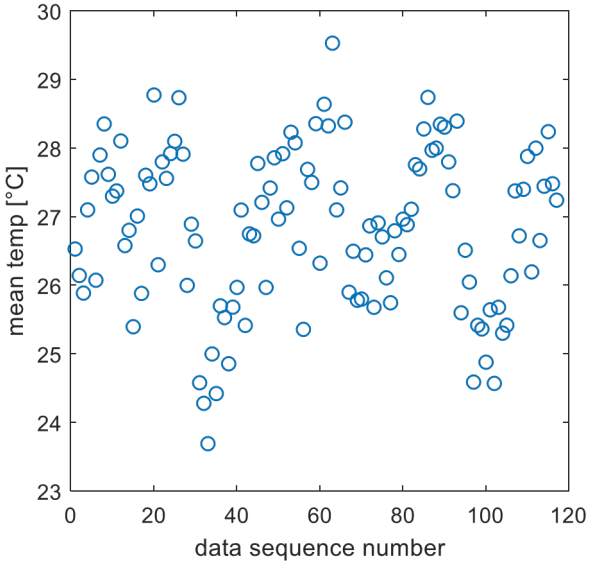
3.0 RESULTS AND DISCUSSIONS

In this chapter, the data for modelling is described, the Polynomial Regression Analysis (PRA). The goodness-of-fit metrics of the models are computed and used to explain the adequacy of the models. **3.1. The Corrosion Data**

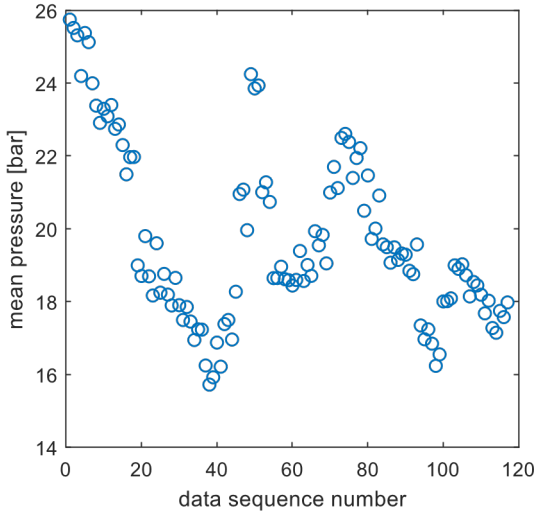
The corrosion data is sourced from an oil and gas multinational in Nigeria. For data privacy and confidentiality reasons, the name of the multinational company is not provided. The data is made up of rows of predictors (mean PH, mean temperature, mean pressure and mean aqueous CO_2 partial pressure) and a row of the response (*Mean Corrosion Rate*). The data set were collected from 2007 to 2011.



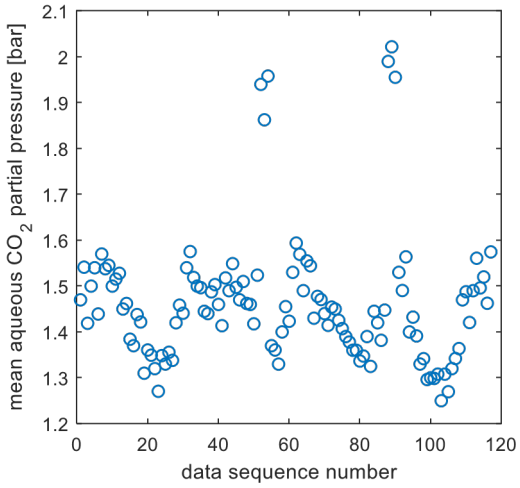
Figures 4.1: A scatter plot of mean PH



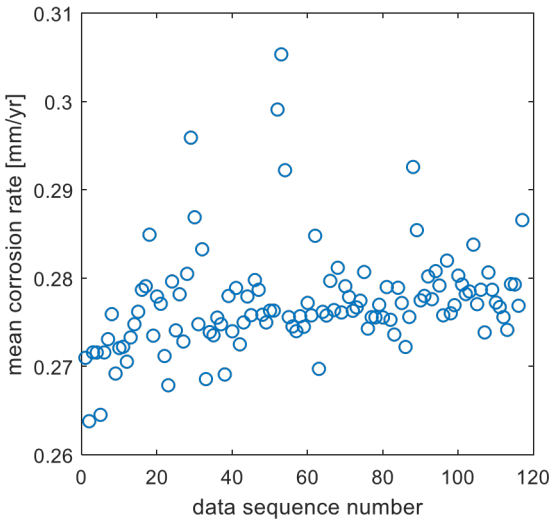
Figures 1: A scatter plot of mean temperature



Figures 2: A scatter plot of mean pressure



Figures 3: A scatter plot of mean aqueous CO_2 partial pressure



Figures 4: A scatter plot of mean corrosion rate [mm/yr]

4.2. The Regression Results

The data are inserted in the normal equations (10) for the multiple linear regression of the first-degree polynomial model. The inserted response data $y^{(j)}$ which is the mean corrosion rate are as given where the inserted predictor data $x_i^{(j)}$ which are the mean PH, mean temperature, mean pressure and mean aqueous CO2 partial pressure are as given in Figures 1 to 4. These normal equations are solved simultaneously in a supervised learning to give the coefficients a_0 and a_i. It can be seen that the first-degree polynomial linear regression gives predictions that fairly correlate with the input data as can be seen from the fairly positive trend about the ideal line. The results for both training and testing are graphed on the same axes for comparative virtual judgement. Even though the model captures the responses poorly, the predicted testing results validate the predicted training results showing that the developed model will perform to similar extend when independent field data are introduced as when the model is trained. The training subset are randomly sampled 80% of the full data set while the testing subset are the remainder.

The parameters of the model are summarized table 1. It can be seen from the table that the model is specifically given as

$$y = 0.29841 - 0.0049549x_1 - 0.00029507x_2 - 0.00034207x_3 + 0.021861x_4 \quad (22)$$

The model coefficients are tested for significance and the numerical measures of the significance are also summarized in table 1. It can be seen from the p-values that the intercept is the most significant predictor in the developed first-degree polynomial model and that, at 95% confidence interval, the coefficients of mean temperature and mean pressure are not significant in the model.

The goodness-of-fit indices of the first-degree model are summarized in table 2 to compare directly the predicted responses against the measured responses. It can be seen that the R^2 is far from unity at 0.3022 and the correlation coefficient is far from unity at 0.5497. These indicate that the higher-degree polynomial regressions should be tried.

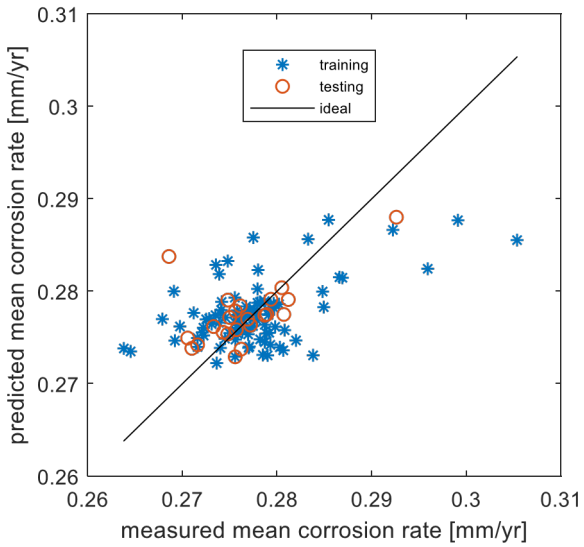


Figure 5: The correlation of the predictions with the measured data for the first-degree regression model

Table 1: The first-degree model and the hypothesis testing results

	Variable Name	**Coefficient**	**SE**	**TStat**	**pValue**
Intercept		0.29841	0.015265	19.548	5.7992e-34
x_1	mean PH	-0.0049549	0.0014461	-3.4265	0.00092731
x_2	mean temperature	-0.00029507	0.0004891	-0.6033	0.54785
x_3	mean pressure	-0.00034207	0.00024701	-1.3848	0.16957
x_4	mean aqueous CO_2 partial pressure	0.021861	0.0039907	5.478	3.9499e-07

Table 2: The goodness-of-fit indices of the first-degree model

GFI	**Training**	**Testing**
R^2	0.3022	0.2948
RMSE	0.0050	0.0040
MBE	0.0000	0.0010
MABE	0.0037	0.0026
MPE	0.0320	0.3633
R	0.5497	0.5891

The observation that two second-degree variables (the interaction of mean pressure and mean aqueous CO_2 partial pressure and of the latter) are not relevant to the model and can be excluded without significantly affecting the model can be verified by reworking the regression to seen the impact on the model performance indicators. The results are presented in equation 4.3, Figure 4.9, tables 4.17 and 4.18 to show the veracity of the observation in terms similarity of the values with the corresponding indicators for the full model in equation 4.2, Figure 4.8, tables 4.15 and 4.16.

$$y = -0.93980 + 0.087521x_1 + 0.043510x_2 + 0.0136527 x_3 + 0.2274035x_4 - 0.00316755x_1x_2 - 0.00157643x_1x_3 - 0.11742780x_1x_4 + 0.0242314756x_2x_4 + 0.0026338902 x_3x_4 + 0.012384649x_1^2 - 0.000997065x_2^2 - 0.0001334789x_3^2 \quad (23)$$

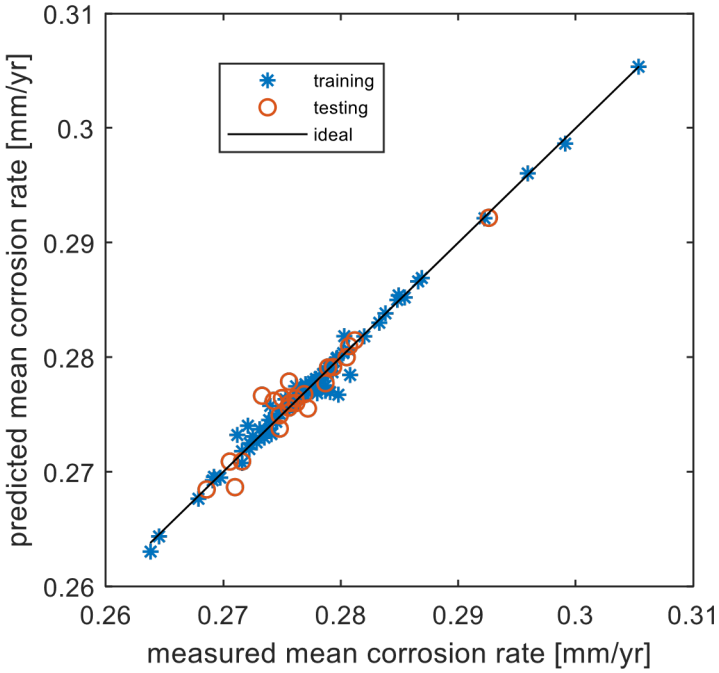


Figure 6: The correlation of the predictions with the measured data for the reduced second-degree regression model

Table 3: The reduced second-degree model and the hypothesis testing results

	Coefficient	**SE**	**tStat**	**pValue**
Intercept	-0.93980	0.06432344	-14.6106096	2.04989253e-24
x_1	0.087521	0.0097881	8.94162774	1.0363354e-13
x_2	0.043510	0.00311802	13.9544395	2.9785469e-23
x_3	0.0136527	0.00101308	13.4763858	2.1650894e-22
x_4	0.2274035	0.02198834	10.3420047	1.8162327e-16
x_1x_2	-0.00316755	0.000210252	-15.065517	3.3101078e-25
x_1x_3	-0.00157643	0.000137267	-11.4843421	1.1233545e-18
x_1x_4	-0.11742780	0.002667005	-44.02983497	2.4276929e-58
x_2x_4	0.0242314756	0.0007811680	31.0195436	1.0304365e-46
x_3x_4	0.0026338902	0.0005068729	5.19635296	1.4876765e-06
x_1^2	0.012384649	0.0005852508	21.1612679	9.5914389e-35
x_2^2	-0.000997065	6.238412378e-05	-15.9826743	9.0802652e-27
x_3^2	-0.0001334789	1.712118786e-05	-7.79612124	1.891749815e-11

Table 4: The goodness-of-fit indices of the reduced second-degree model

GFI	Training	Testing
R^2	0.9859	0.9341
RMSE	0.0007	0.0012
MBE	0.0000	0.0001
MABE	0.0004	0.0008
MPE	0.0007	0.0390
r	0.9929	0.9676

These results demonstrate that corrosion is not only affected by the predictors but also by their self and cross interactions as wide difference in predictive accuracy between the firstorder and second order polynomial models.

i. First-degree polynomial linear regression gives predictions that fairly correlate with the input data as can be seen from then fairly positive trend about the ideal. Even though the model captures the responses poorly, the predicted testing results validate the predicted training results showing that the developed model will perform to similar extend when independent field data are introduced as when the model is trained.

ii. It can be seen that all parameters of the model are significant at 95% confidence interval except the coefficients of the interaction of mean pressure and mean aqueous CO_2 partial pressure (x_2x_3) and the square of mean aqueous CO_2 partial pressure (x_4^2). These mean that the two second-degree variables are not relevant to the model and can be excluded without significantly affecting the model.

iii. The observation that two second-degree variables (the interaction of mean pressure and mean aqueous CO_2 partial pressure and of the latter) are not relevant to the model and can be excluded without significantly affecting the model can be verified by reworking the regression to seen the impact on the model performance indicators.

iv. These results demonstrate that corrosion is not only affected by the predictors but also by their self and cross interactions as wide difference in predictive accuracy between the first-order and second order polynomial models.

4.0 CONCLUSION

First-degree polynomial modelling without considering the interaction effects of predictors, this work bridges the gap by studying the interaction effects and excluding the non-significant ones. These results demonstrate that corrosion is not only affected by the predictors but also by their self and cross interactions as deduced from the wide difference in predictive accuracy between the first-order and second order polynomial models. The full first-degree model was found to be grossly inadequate for modeling the data. The full second-degree polynomial is found to be very accurate: for training/testing, the R^2 is almost unity at 0.9869/0.9361, the RMSE is very low at 0.0007/0.0012, the MBE is very low at 0.0000/0.0002, the MABE is very low at 0.0004/0.0008, the MPE is very low at

0.0006/0.0636 and the correlation coefficient is almost unity at 0.9934/0.9689. The second-degree model coefficients are tested using hypothesis testing for significance and it is concluded that all parameters of the model are significant at 95% confidence interval except the coefficients of the interaction of mean pressure and mean aqueous CO_2 partial pressure (x_2x_3) and the square of mean aqueous CO_2 partial pressure (x_4^2).

These not-significant interactions were excluded from the model to give the reduced model which expectedly performed as well as the full model as follows: for training/testing, the R^2 is almost unity at 0.9859/0.9341, the RMSE is very low at

0.0007/0.0012, the MBE is very low at 0.0000/0.0001, the MABE is very low at 0.0004/0.0008, the MPE is very low at 0.0007/0.0390 and the correlation coefficient is almost unity at 0.9929/0.9676. This confirms the non-significance of the excluded interactions. A constrained optimization of the reduced second-degree polynomial model within the lower and upper bounds of all the predictors is carried using genetic algorithm. The calculated optimal is 0.191498 mm/year occurring at mean PH = 8.446 °C, mean temperature = 23.692, mean pressure = 15.725 bar and mean aqueous CO_2 partial pressure = 2.022 bar.

References

1. Dawotola (2012) '*Risk Based Maintenance of Petroleum Pipelines*'. Delft University of Technology, The Netherlands.
2. De Silva, D.; Moglia, M.; Davis, P.; Burn, S. (2006) '*Condition assessment to estimate failure rates in buried metallic pipelines.*' J. Water Supply. Res. Technol. -AQUA 55 179-191.
3. Ekpu, A. O. O. and Ehighelua, I. (2004) 'Sabotage in the Oil Industry: The Role of the Law', *Environmental Policy and Law*, 34(6), pp. 274–280. Available at:
4. https://www.scopus.com/inward/record.uri?eid=2-s2.0-
5. 11844252563&partnerID=40&md5=37e9fb83201c04f8ff4ea24c1a55ac9f.
6. El-Abbasy, M. S. *et al.* (2015) 'A condition Assessment Model for Oil and Gas Pipelines Using Integrated Simulation and Analytic Network Process', *Structure and Infrastructure Engineering*, 11(3), pp. 263–281. doi: 10.1080/15732479.2013.873471.
7. Levin, D. (1998) 'The Approximation Power of Moving Least-Squares', *Mathematics of Computation*, 67(224), pp. 1517–1532. doi: 10.1090/S0025-5718-98-00974-0.
8. Li, C. Q., Mohamoodian, M. (2013). '*Risk based service life prediction of undreground pipelines subjected to corrosion.*' J. Reliah. Eng. Syst. Saf. 119, 102-108.
9. Maes, M.; Dann, M.; Salama, M. M. (2008). '*Influence of grade on the reliability of corroding pipelines.*' Reliah. Eng. . Syst. Saf. 93, 447-455.
10. Mahmoodian, M., Li, C. Q. (2016). '*Structural integrity of corrosion affected cast iron water pipe using a reliability-based stochastic analysis method.*' Strut. infrasruct. Eng 12 (10)
11. Mahmooodian, M., Li, C. Q. (2015). '*Stochastic failure analysis of defected oil and gas pipelines Handbook of Materials Failure Analysis with Case Studies from the Oil and Gas Industry.*' Elsevier.
12. Senouci, A., Elabbasy, M., Elwakil, E., Abdrabou, B., Zayed, T.(2014) 'A Model for Predicting Failure of Oil Pipelines', *Structure and Infrastructure Engineering*, 10(3), pp. 375–387. doi: 10.1080/15732479.2012.756918.
13. Senouci, A., El-Abbasy, M. S. and Zayed, T. (2014) 'Fuzzy-Based Model for Predicting Failure of Oil Pipelines', *Journal of Infrastructure Systems*, 20(4), p. 04014018. doi: 10.1061/(ASCE)IS.1943-555X.0000181.
14. Sinha, S. K. and Pandey, M. D. (2002) 'Probabilistic Neural Network for Reliability Assessment of Oil and Gas Pipelines', *Computer‐Aided Civil and Infrastructure Engineering*, 17(5), pp. 320–329. doi: 10.1111/1467-8667.00279.

15. Thacker, R. H., Light, G. M., Dante, J. F., Trillo, E., Song, F., Popellar, C.F., Coulter, K. E., Page, R. A. (2010). Corrosion Control In Oil and Gas Pipelines vol. 237. Southwest Research Institute, San Anthonio,S TX.