Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/NA10.20250372
Published: June 17, 2025

Principal Component Analysis Machine Learning Technique On The Underground Big Diameter Steel Pipelines Condition Assessment and Related Data: South African Bulk Water Distribution Utilities

Mthunzi Lushozi
School of Mechanical, Industrial and Aeronautical Engineering
University of Witwatersrand
Johannesburg, South Africa
mthunzi.lushozi@gmail.com

Gbeminiyi John Oyewole
School of Mechanical, Industrial and Aeronautical Engineering
University of Witwatersrand
Johannesburg, South Africa
gbeminiyi.oyewole@wits.ac.za

Abstract

The structural integrity of big-diameter underground steel pipelines is a critical determinant of the long-term viability and operational resilience of bulk water utilities distribution networks, particularly in regions grappling with aging infrastructure and constrained resource. This study explores the application of factor analysis, a multivariate statistical technique grounded in machine learning principles, to assess the condition assessment data gathered from these pipelines managed by a major South African bulk water utility. A diverse condition assessment campaign was undertaken, integrating diverse diagnostic methodologies including External Corrosion Direct Assessment, Direct Current Voltage Gradient, Guided Ultrasonic Testing, and acoustic leak detection technologies (SmartBall and Sahara). These were complemented by asset registry attributes such as pipeline age, wall thickness, joint configuration, and coating condition, culminating in a robust dataset for machine learning analysis. Employing Principal Component Analysis (PCA) followed by Exploratory Factor Analysis (EFA), the study identified four latent constructs that encapsulate the underlying dimensions of pipeline degradation: (1) *Structural Aging and Joint Vulnerability*, (2) *Cathodic Protection Efficacy*, (3) *Leak Incidence and Defect Density*, and (4) *Material Durability in Corrosive Contexts*. The factor structure was substantiated through orthogonal rotation (Varimax), yielding a statistically robust model. The final factor model was delineated as follows:

- Parallel Analysis 1: Age and Integrity – encapsulating variables such as pipeline age, wall thickness, and historical leak frequency, indicative of cumulative structural fatigue.
- Parallel Analysis 2: Construction and Design – emphasising design-centric attributes, including joint type and wall configuration, which modulate mechanical resilience.
- Parallel Analysis 3: Material and Protection – reflecting the interplay between protective coatings, environmental corrosivity, and cathodic protection performance.
- Parallel Analysis 4: Performance and Risk – aggregating operational metrics such as defect frequency, pipeline length, and fault incidence, offering a predictive lens on system reliability.

Model fit indices affirm the solution's adequacy: $\chi^2 = 14.5$, $p = 0.63$; RMSEA = 0.00; TLI = 1.043; RMSA = 0.03, indicating excellent model-data congruence despite a modest Kaiser-Meyer-Olkin (KMO) measure. This research contributes literature, integrative analytical framework for underground big-diameter steel pipeline infrastructure diagnostics, enabling data-driven prioritisation of maintenance and rehabilitation in resource-constrained water utility environments.

Keywords
Factor Analysis, Machine Learning, Underground Big Diameter Steel Pipeline and Condition Assessment

1. Introduction
Worldwide, underground steel pipelines remain a reliable means of conveying vital resources, such as water, necessary for sustaining and advancing national economies (Mielke et al., 2023). Their research highlights the durability and adaptability of steel pipeline systems in addressing the needs of swiftly expanding urban populations. However, it is reported that the structural integrity of these assets progressively deteriorates over time, resulting in significant corrosion failure due to aging and extended exposure to adverse environments, such as soil corrosivity (Shipilov and May, 2005).

Sarwar et al. (2024) assert that efficient pipeline integrity management is essential for guaranteeing the safety, reliability, and durability of these vital assets. Deo et al. (2023) corroborate this by asserting that effective management via condition assessment and prompt repair or renewals is essential to mitigate the significant economic burden resulting from underground steel pipeline failures. However, a concerning aspect is that Fan and Yu (2023) recognised that the data on pipe failures is typically constrained. Recent academic research has increasingly demonstrated the efficacy of integrating machine learning techniques with statistical dimensionality reduction methods to enhance the predictive modeling of underground steel pipeline failures.

For instance, Liu and Meng (2025) studied pressure failure processes in corroded pipelines, utilising machine learning methods to model and forecast failure thresholds across different corrosion conditions. Their research highlights the capacity of data-driven methodologies to enhance conventional engineering evaluations. Phan and Duong (2020) asserted that the integration of Principal Component Analysis (PCA) with Adaptive Neuro-Fuzzy Inference Systems (ANFIS) can markedly enhance the precision of burst pressure forecasts in flawed subterranean steel pipelines. Their hybrid framework demonstrates the efficacy of integrating statistical feature extraction with intelligent inference techniques to address the nonlinearities seen in pipeline degradation processes.

Zhang et al. (2021) used PCA to identify leakage events in underground steel pipelines. Their findings illustrate that dimensionality reduction increases computing efficiency and improves the signal-to-noise ratio in sensorbased monitoring systems. They underscore the efficacy of PCA in preparing high-dimensional diagnostic data for real-time anomaly detection. Zhu (2023) utilised Artificial Neural Networks (ANNs) to predict the structural behavior of corroded underground steel pipelines, demonstrating the ability of deep learning architectures to generalize intricate deterioration patterns from empirical data. Zhu's research enhances the expanding literature supporting the incorporation of machine learning in infrastructure health monitoring, especially in scenarios where conventional inspection techniques are hindered by accessibility or financial limitations. These findings highlight a move towards data-centric approaches in pipeline integrity management, where machine learning and statistical modeling converge to offer more detailed, scalable, and predictive insights into asset performance under deteriorating conditions. However, we are unaware of a study that considers various condition assessment data combined with pipeline parameters to determine the key factors affecting the pipeline's useful life.

Therefore, to bridge that gap, one of the water utilities in Africa has conducted a comprehensive condition assessment of its infrastructure, including the underground steel pipeline. The techniques used on pipeline condition assessment included ultrasound techniques and dive-in cameras for leak detection and internal lining; External Corrosion Direct Assessment (ECDA) for soil corrosivity; Guided Ultrasonic (GUL) for pipeline thickness; and Direct Current Voltage Gradient (DSVG) for coating defects.

We are unaware of the evidence that any water utility in the world has carried out a thorough condition assessment utilising this variety of techniques to obtain data on their underground large-diameter steel pipeline, at least at the same time. This condition assessment data, combined with the pipeline parameters such as thickness, type of joints, material of construction, etc., will make this study unique. The previous studies such Jiang (2022) research on pipeline corrosion and machine learning; Tavakoli (2020) examination of asset remaining useful life and predictive models; and Zhou et al. (2017) analysis of hydraulics and predictive models were deficient in the condition assessment data utilised, and their emphasis was not primarily on predictive models but rather on the analysis of failed pipe samples.

Consequently, taking into consideration the challenges that water distribution utilities in South Africa face, which include deteriorating infrastructure, a lack of available water, and financial constraints (Ruiters and AmadiEchendu, 2022), Factor Analysis has the potential to be an efficient instrument for improving maintenance and investment strategies (Kurita, 2020). Suppose the principal variables that contribute to pipeline deterioration are understood. In that case, the utilities can concentrate their resources on the most critical sites and carry out targeted measures to extend the lifespan of their assets. The goal is to assist the South African Water Utilities in improving the precision and efficacy of pipeline status evaluations, facilitating proactive maintenance and minimising the likelihood of unforeseen failures.

The main objective of this study is to utilise machine learning techniques, particularly factor analysis, to analyse and interpret condition assessment datasets related to big-diameter underground steel pipelines. The pipelines, essential to Utilities' bulk water distribution infrastructure, experience degradation over time due to environmental, operational, and material pressures. The research aims to identify the fundamental latent elements contributing to pipeline degradation and generate predictive insights regarding possible failure mechanisms among various condition assessment and pipeline parameters. The study employs component analysis to diminish the dimensionality of intricate, multivariate datasets while retaining the most significant variance. This method enables the identification of primary degradation factors that are not readily observable but are deduced from data patterns. The primary objective is to improve pipeline condition evaluations' accuracy, reliability, and interpretability, thus empowering South African water utilities to make informed decisions about maintenance prioritising, rehabilitation strategies, and risk management. The study will pursue the following objectives to attain the primary goal.

a) To ascertain the significance of factors in pipeline failure. This objective entails a thorough statistical analysis to assess the relative impact of each observed variable (e.g., defect rate, pipeline thickness, pipe age, etc.) on pipeline failure. Initial factor loading evaluations will be utilized to measure the impact of each component. This stage is essential for determining which variables should be included in the factor model and comprehending the hierarchical structure of degradation drivers.

b) To derive the principal latent factors by four-factor analysis. This sub-objective is to utilise exploratory factor analysis (EFA) to identify four interpretable latent components that represent the most critical dimensions of pipeline degradation. The quantity of factors will be ascertained by eigenvalue criteria (e.g., Kaiser's rule), scree plot evaluation, and parallel analysis. Each element will be analyzed according to the loading patterns and designated depending on the predominant characteristics it signifies.

c) To formulate and authenticate a Four-Factor Predictive Model. The ultimate sub-objective is to develop a four-factor model specifically designed for practical implementation by South African water utilities, derived from the identified themes. The model will be tested by rotation approaches, including Varimax (orthogonal), to improve interpretability and verify the resilience of the factor structure. The verified model will function as a decision-support instrument, allowing South African water utilities to predict failure risks and enhance asset management techniques.

2. Literature Review

2.1 Principal Component Analysis

Principal Component Analysis (PCA) has become an essential preprocessing method in evaluating underground steel pipeline conditions, specifically for improving the efficacy of machine learning models that forecast corrosion rates and assess pipeline systems' remaining useful life (RUL). PCA diminishes noise and redundancy by converting high-

dimensional input data into a reduced set of uncorrelated principal components, enhancing prediction models' accuracy and computational efficiency (Chen et al. 2020; Yang et al. 2020). This dimensionality reduction is advantageous for managing intricate datasets obtained from various inspection methods and environmental factors. Migenda et al. (2021) showed that PCA alleviates the computing demands of extensive data processing and improves model generalisation by concentrating on the most salient features. This results in more resilient and comprehensible models that detect pipeline surveillance degradation patterns and failure indicators. In addition to predictive modelling, PCA has been utilised to detect operational irregularities.

Świercz and Mroczkowska (2020) demonstrated the application of PCA in monitoring departures from standard operating conditions, including oscillations in pressure and flow rate, which may signify leaks or structural irregularities. PCA is reported to identify minor anomalies that may remain undetected in unprocessed sensor data by establishing a normative behaviour baseline. Numerous studies have investigated the utilisation of PCA in leak detection and metal loss assessment. For instance, Ji et al. (2021) and Zhou et al. (2019) utilised PCA-based frameworks to discern leak signatures in pipeline systems, exhibiting enhanced detection sensitivity and diminished false alarm rates. Ordóñez et al. (2015) employed PCA to estimate metal loss in corroded underground steel pipelines, demonstrating that principal components could accurately represent the fundamental degradation patterns derived from inspection data. These findings highlight the adaptability of PCA as a diagnostic and prognostic instrument in pipeline integrity management. Its capacity to synthesise intricate, multidimensional facts into meaningful insights renders it essential for the development of intelligent infrastructure monitoring systems.

Prior to executing Factor Analysis, the Kaiser-Meyer-Olkin (KMO) measure must be conducted to evaluate the appropriateness of the data for factor analysis (Hu et al., 2015; Song et al., 2023;Kurita,2020; Zhu, 2023). The KMO test assesses the ratio of common variance, with elevated values signifying enhanced appropriateness (Malibary et al., 2019). Generally, a KMO value exceeding 0.6 is typically regarded as satisfactory (Zhu 2023).

Furthermore, literature suggest that after the initial PCA, rotation is necessary to enhance the interpretability of the primary components. It is argued that it seeks to identify a more straightforward structure in which each variable exhibits significant loading on only a limited number of components. For instance, in support of this is Kavengik (2025) as he posits that it improves interpretability, as a variable that initially loads moderately on several major components can subsequently load significantly on a single factor post-rotation. Rohe and Zeng (2023) concur, asserting that factor rotation frequently enhances interpretability and that varimax rotation facilitates statistical inference. Varimax is defined as a conventional rotation technique in PCA that aims to maximize the variance of the squared loadings for each component, hence simplifying the components by elevating high loadings and diminishing low loadings (Akhtar-Danesh 2023). However, if it is anticipated that the underlying variables are interrelated, Promax is indicated (Akhtar-Danesh 2023).

It argued that these validation methods alone are not enough, and it always be advised instruments such as scree plot, eigen values and biplot also used to enhance the model validly. According to Team (2018), a scree plot illustrates the extent of variance captured by each Principal Component from the data and serves as a diagnostic instrument to evaluate the efficacy of PCA on the dataset. Nonetheless, Ledesma et al. (2015) acknowledge its usefulness but contend that it possesses certain limitations when utilised in isolation; hence, it should be employed alongside other methodologies such as eigen values, biplot etc. Ryu et al. (2022) suggest that eigenvalues measure the amount of variation, and their advantage is that can be presented as numeric and visuals, while Team (2018), assert that biplot can be used to read how strongly each characteristic influences the principal components.

2.2 Underground Big Diameter Steel Pipelines And Condition Assessment

According to Kostryzhev et al. (2010) a big-diameter pipeline has a diameter of greater than 400 mm. The ageing of underground big-diameter steel pipelines is challenging for many countries globally. For instance, Shou and Huang (2020)has reported that in Taiwan, ageing and overused underground big-diameter steel pipelines are prone to damage by corrosion. They further assert that a condition assessment and risk management can provide an understanding of the mechanical behaviour of damaged underground pipelines. This was supported by Hassan et al., (2019) They maintain that regular condition assessments are critical to avoiding underground pipeline deterioration, as rehabilitation and maintenance require a considerable budget and time-consuming planning process.

A condition assessment of underground steel pipelines is critical for engineers working in the realm of buried infrastructure (Muggleton et al., 2016), and can assist with proactive warnings of imminent failures Numerous non-destructive condition assessment technologies are used to gather quantitative data relating to the condition of underground pipelines, such as the Sahara system, SmartBall, magnetic flux leakage, remote field eddy current, and ground-penetrating radar (Wand, 2017). This study will focus mainly on leak detection (using SmartBall and Sahara) and external corrosion direct assessment (ECDA).

2.3 Leak Detection Techniques

In Europe, at least 25% of water is being leaked through underground pipeline networks, while the reported amount of up to 50% in some developed countries such as Spain, Britain, Australia, France, and the United States (Al-Kadi et al., 2013). The leaks in underground steel pipelines can result in serious challenges such as sinkholes and drinking-water scarcity; therefore, they must be detected and monitored (Ali and Choi, 2019). Leakage management-related methods are broadly classified as leakage assessment (quantification of water loss) and leakage detection (detection of leakage hotspots) methods (Atef et al., 2016).

Leakage Detection Systems are important aspects of pipeline technology because their primary purpose is to assist pipeline controllers in detecting and localising leaks (Al-Kadi et al., 2013). Leak detection is grouped into two main categories such as static and dynamic leak detection (El-Zahab and Zayed, 2019). Numerous tools and techniques have been proposed for monitoring, detecting, and preventing leakages in underground water pipelines (Ali and Choi, 2019).

Amongst these tools and techniques, Williams et al. (2019) assert that acoustic leak detection inspection tools have become a common technique used for underground pipeline leakage detection. SmartBall is an acoustical inspection technology that is currently widely used for leak detection, and it is described as an un-tethered inspection tool that can inspect many kilometres of pipeline during a single deployment (Chapman 2012). It is a free-swimming tool that is enclosed within a foam ball which is equipped with a highly sensitive acoustic sensor that can detect leaks on pressurized underground steel pipelines (Livingston et al. 2019).

2.4 External Corrosion Direct Assessment (ECDA)

Underground steel pipelines are generally subject to corrosion where there are inadequate levels of cathodic protection or any other corrosion protection mechanisms(Nicholson 2006). Consequently, the External Corrosion Direct Assessment (ECDA) was developed to proactively prevent external corrosion and ensure the integrity of underground steel pipelines (McDonnell and Onnuoha 2015). They assert that, this condition assessment technique normally follows the four steps i.e. pre-assessment, indirect inspection, direct examination and post-examination.

ECDA is used to identify hotspot areas where coating defects have already formed and can ascertain where cathodic protection (CP) is insufficient for the underground steel pipelines(Onuoha et al. 2015). According to Corrpro (2010), ECDA has been used to assess the condition of thousands of underground steel pipelines over the years. The corrosion and leaks in underground steel pipelines can result in high consequences for water utilities (Mcdonnell and Onuoha, 2015). Therefore, ECDA condition assessments can assist water utilities to identify corrosion hot spots before major repairs are required (Onuoha et al. 2021). Severe pipeline deterioration can trigger impairment of a portion of the pipeline, which could affect the water utility's balance sheet.

2.5 Pipeline Coating and Cathodic Protection

External corrosion is a major threat to the integrity of underground steel pipelines for water utilities; therefore, coating the pipeline and cathodic protection are the main methods used to protect the underground pipelines (Gao and Song, 2009). The United States alone has over 4 million km of underground pipelines, and approximately 8% of pipeline incidents were caused by external corrosion (Kim et al. 2021).

Therefore, the protection coating is normally used as a first line of defense to avoid corrosion attacks for underground steel pipelines, (Chen et al., 2009). Kim et al. (2021) concur and assert that the underground steel pipeline forms an electrochemical system where a steel pipeline is an electrode and soil an electrolyte. Therefore, the coating is applied

as a barrier to separate the steel pipeline from the electrolyte to avoid corrosion. McDonnell and Onuoha (2015) assert that ECDA coating defects can be inspected using various techniques, including the direct current voltage gradient (DCVG). According to Onuoha et al. (2021) the DCVG and soil corrosivity data collected by water utilities are used to select areas susceptible to external corrosion for underground steel pipelines. This study collected data using the DCVG technique. Where there was suspected coating damage, excavations were done to conduct physical inspections, and in some areas, wall thickness tests were also conducted.

According to McDonnell and Onuoha, (2015) it is generally accepted that the results are better when two complementary technologies are applied simultaneously for corrosion control than when applied individually. They further suggest that when cathodic protection and coatings are used in an underground pipeline, the cathodic protection system can complement the coating by providing protection on holidays. Generally, the natural corrosion process for underground steel pipelines is prevented by including a physical layer between the underground steel pipeline and the soil. This is normally achieved by applying coatings on the external wall of the underground pipeline or the addition of an external electrical current source, such as cathodic protection (Kim et al. 2021). CP assessments are critical in the evaluation of an underground steel pipeline's condition and in assessing the possible risk of failure (McDonnell and Onuoha 2015).

3. Methods

This chapter addresses study design and methodology, quantitative investigations, statistical analysis techniques, and machine learning approaches, as summarised in Figure 1. This figure outlines the methodology approach adopted by this study. This research employed quantitative confirmatory design. The confirmatory approach is suitable because this study depends on empirical data to evaluate the correlations between pipeline condition indicators and failure probability (Creswell and Creswell 2018). Research design is a strategic framework that guides the study and bridges research questions and implementation (Blanche et al. 2007). According to Andrew et al. (2011), research can progress using several strategies, such as quantitative, qualitative, or mixed methods approaches. A researcher adopts a quantitative approach to respond to research questions requiring numerical data, a qualitative approach to research questions requiring textural data, and a mixed methods approach to research questions requiring both numerical and textural data (Williams 2007).

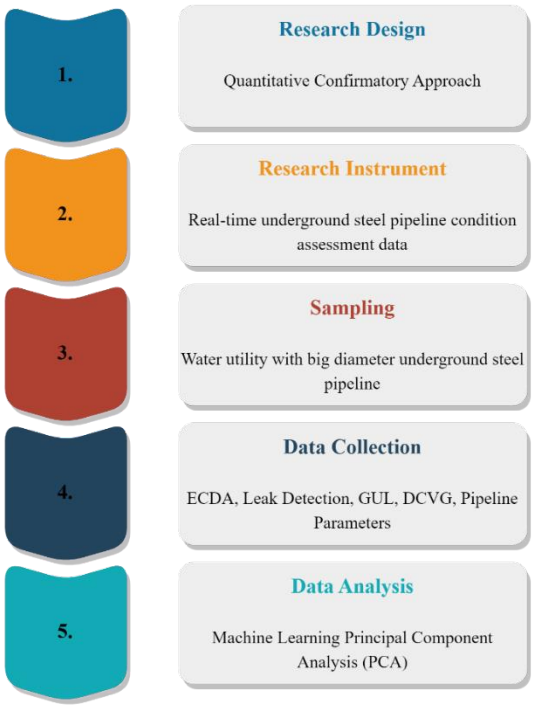

Figure 1. Research Methodology Overview

The study instrument consists of a condition assessment dataset gathered from a water utility managing bigdiameter underground steel pipelines. This dataset comprises non-destructive assessment (NDE) methodologies, including External Corrosion Direct Assessment (ECDA), Guided Ultrasonic Testing (GUL), Direct Current Voltage Gradient (DCVG), and acoustic leak detection devices. These techniques offer detailed insights into the integrity, corrosion activity, and structural anomalies of big-diameter underground steel pipes.

The sampling was purposeful, focusing on a water utility with a recorded history of pipelines condition obtained through extensive infrastructure condition assessment technologies. This guarantees the accessibility of superior, longitudinal data appropriate for machine learning applications. The emphasis on big-diameter pipelines is warranted due to their essential function in bulk water distribution and the significant economic and social repercussions of their failure. The data collection incorporated many condition evaluation techniques, each providing distinct diagnostic insights. ECDA and DCVG deliver electrochemical assessments of corrosion activity, whereas GUL and leak detection systems furnish structural and acoustic diagnostics. Additional pipeline parameters, including age, material characteristics, and environmental conditions, were documented to enhance the data.

The study utilises machine learning methodologies for data analysis, specifically focusing on Principal Component Analysis (PCA) for dimensionality reduction and feature selection. PCA facilitates the extraction of underlying patterns from high-dimensional data, enhancing model interpretability and computational efficiency (Jolliffe and Cadima, 2016). The validation techniques, Kaiser-Meyer-Olkin **and** Bartlett's tests were performed before PCA to validate the dataset's suitability for dimensionality reduction. Varimax rotation was employed subsequent to PCA to elucidate the structure of the main components, facilitating the interpretation of the condition assessment variables (e.g., ECDA, GUL, DCVG) that most significantly contribute to each latent factor. The validations were performed to confirm that the PCA findings are statistically valid and realistically interpretable, hence reinforcing the basis for later machine learning models.

4. Data Collection

The integrity and operational reliability of big diameter underground steel pipeline infrastructure are essential for the safe and efficient transportation of water across long distances. Figure 2 illustrates a summary of typical condition assessment data that was collected and used for the study, which mainly quantitative diagnostic approaches. This figure summaries the collected data into four interconnected domains: Pipeline Parameters, Leak Detection, External Corrosion Direct Assessment (ECDA), and Pipeline Defect Characterisation. Each area encompasses a distinct array of diagnostic instruments and data categories that collectively interpret the structural integrity and functionality of underground steel pipeline network.

Figure 2. Data Collection – condition assessment

The Pipeline Parameters section records fundamental design and construction characteristics, including wall thickness, material composition, and junction type, which are critical for baseline integrity modeling. Leak Detection utilised sophisticated acoustic and sensor-based technologies, such as Smart Ball and Sahara systems, to detect active leaks with great spatial resolution. The ECDA domain emphasises the electrochemical and environmental factors affecting the pipeline, such as soil resistivity, pH levels, and the functionality of cathodic protection systems—elements crucial for evaluating corrosion vulnerability. The Pipeline Defect section incorporates non-destructive evaluation (NDE) methods, including Guided Ultrasonic Testing (GUL) and Direct Current Voltage Gradient (DCVG) surveys, to identify wall thinning and coating irregularities, respectively. This integrative approach promotes condition monitoring granularity while supporting predictive maintenance strategies and risk-informed decision-making in pipeline asset management.

5. Results and Discussion

A range of multivariate statistical approaches was utilised in R Studio to investigate the dataset's underlying structure and evaluate inter-variable interactions for the underground big diameter steel pipeline dataset. The analytical procedure commenced with a correlation plot to illustrate the strength and direction of linear relationships among variables, offering an early assessment of potential multicollinearity and clustering patterns, as shown in Figure 3, Subsequently, a scree plot was presented, enabling the identification of the optimal number of components or factors to maintain by analysing the eigenvalue distribution and pinpointing the inflexion point as shown in Figure 4.

Moreover, a biplot was created to concurrently depict observations and variable loadings in a reduceddimensional space, providing insights into each variable's contribution to the principal components and the spatial distribution of instances. Lastly, a four-factor solution was ultimately derived based on theoretical rationale and empirical requirements, including eigenvalues exceeding one and the interpretability of the factor structure. This solution was analysed to clarify the underlying constructions of the observed variables and to evaluate the coherence and discriminant validity of the identified factors. Collectively, these visual and statistical outputs establish a thorough basis for understanding the data's dimensionality and internal structure, which in turn informs the subsequent final four-factor model presented in Figure 7.

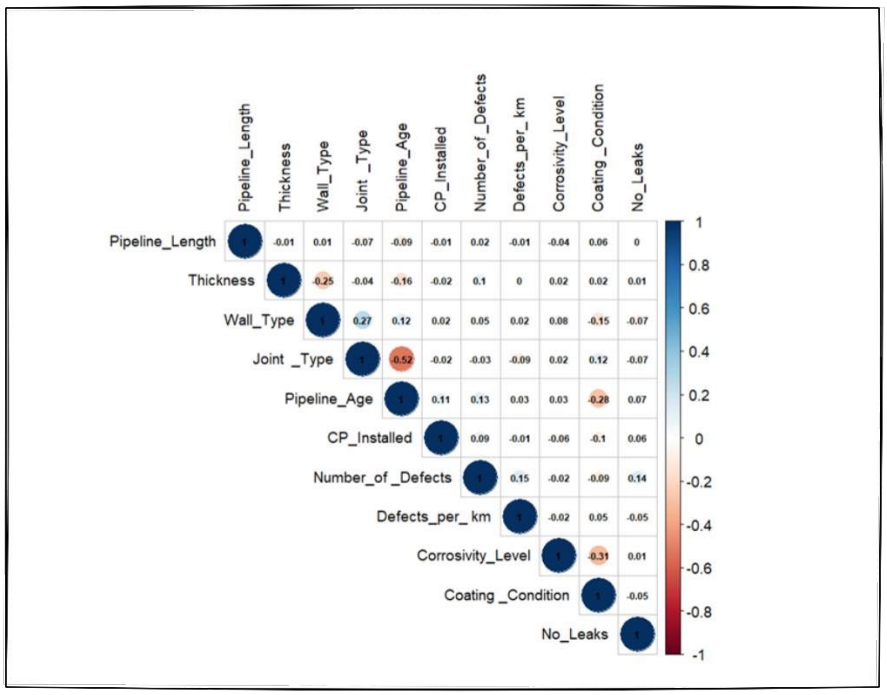

Figure 3. Correlations plot

The pipeline length and number of defects have a strong positive correlation. This means that longer pipelines tend to have more defects. Moreover, the pipeline age and soil corrosivity have a strong positive correlation. Therefore, this means older pipelines may be in more corrosive environments. Cathodic Protection installation and the number of defects have a strong negative correlation. Therefore, this means that installing cathodic protection reduces the number of defects. Moreover, coating condition and soil corrosivity levels also strongly correlate negatively. This, therefore, implies that better coating conditions are associated with lower corrosivity. Factor analysis has traditionally been utilised to investigate critical success factors for water infrastructure projects (Dithebe et al. 2019). Additionally, factor analysis is utilised to investigate the variables affecting competition in bottled drinking water sales (Yulianto et al. 2020).

5.1 Optimal Number of Factors to Retain

The scree plot on the left side of figure 4 demonstrates the variance captured by each of the first 10 principal components (PCs) in a dataset (Ledesma et al., 2015; Team, 2018). The y-axis represents the amount of variance explained, while the x-axis lists the principal components in order. This plot demonstrates steep decline in variance from PC1 to PC3 or PC4 because after PC4, the curve begins to flatten, indicating that additional components contribute progressively less to explaining the total variance. The elbow rule is a heuristic used to determine the optimal number of principal components to retain. It identifies the point at which the marginal gain in explained variance drops sharply—this point resembles an "elbow" in the curve (Nguyen and Holmes, 2019; Yoo,2024.

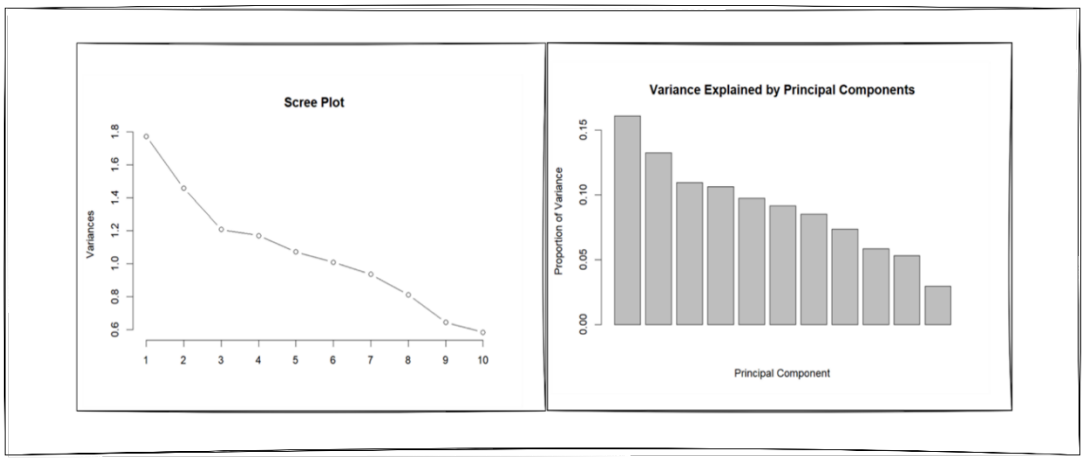

Figure 4. Scree Plot and Principal Components Variance Explained

Therefore, it can be concluded from the curve that retaining the first 3 or 4 components would capture most of the meaningful variance in the data. Consequently, beyond this point, the inclusion of additional components yields diminishing returns and may introduce noise or overfitting. As outlined on the literature review section, the elbow rule can be subjective (Yoo, 2024), and the scree plot may at times be ambiguous and open to interpretation (Ledesma et al., 2015). Hence, variance explained curve and biplot are used.

The first bar on the curve on the right of Figure 4 has the highest value, approximately 0.15. This indicates that the first principal component explains about 15% of the total variance in the data. This curve demonstrates that the first few principal components capture a significant portion of the variance. This is typical in PCA, where the initial components are the most informative. The diminishing heights of the bars indicate that the additional variance explained becomes minimal after a certain number of components. These results align with a scree plot as they both suggest that retaining the first few principal components, such as the first three or four, would capture most of the critical information in the data while reducing its dimensionality. Therefore, four factors are justified. Moreover, biplot as presented in Figure 4, was also used as it is considered an effective instrument for visualising the interactions between variables and observations in factor analysis (Hashemi et al., 2018). They also suggest that it enables the visualisation of variable clustering and the positioning of observations in relation to these clusters.

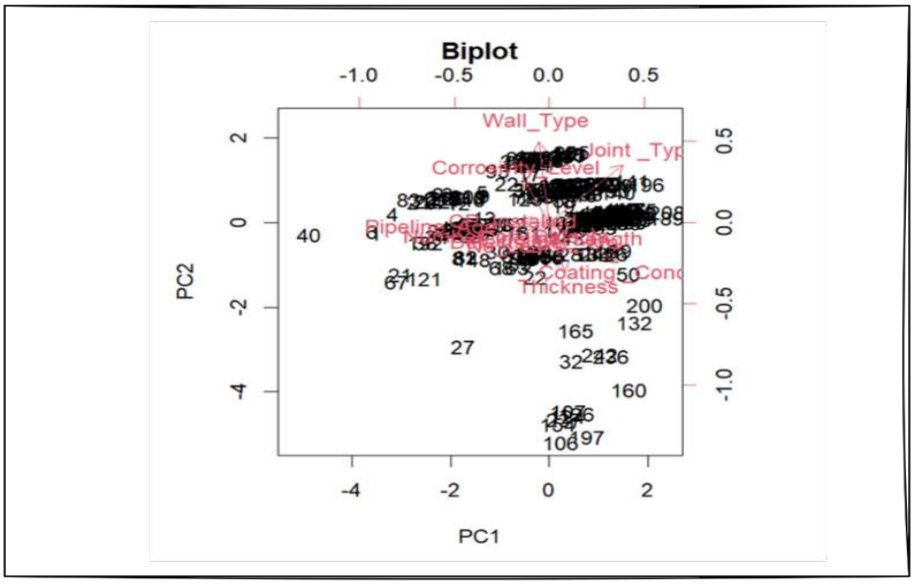

Figure 5. Biplot

Factor analysis is a potent instrument; however, it is susceptible to misuse (Streiner 1994). Consequently, it is essential to meticulously evaluate the assumptions and constraints of the technique prior to interpreting the results. It may be beneficial to examine how other researchers have employed factor analysis in analogous circumstances (Silva et al., 2017). The inspection of Figure 5.3 suggests that pipeline thickness and external coating are positively correlated, while wall type, joint type, and corrosivity are positively correlated. According to this Biplot, there are two principal components: PC1 and PC2. The wall type, joint type, and corrosivity contribute positively to PC2.

5.2 Final Model

The integrity and reliability of pipeline infrastructure are essential for the safe and efficient distribution of bulk water. As worldwide water demand escalates, the necessity for resilient underground big-diameter pipeline systems capable of enduring environmental, operational, and pressure conditions also increases. Big-diameter underground steel pipelines, frequently traversing extensive and diverse landscapes, are susceptible to numerous deterioration mechanisms, such as corrosion, mechanical wear, and environmental exposure. Thus, the evaluation and upkeep of pipeline networks have emerged as a primary focus for engineers, regulators, and asset managers around the globe. This study tackles the issue of pipeline integrity evaluation by utilising factor analysis to identify and classify the most significant elements impacting pipeline performance.

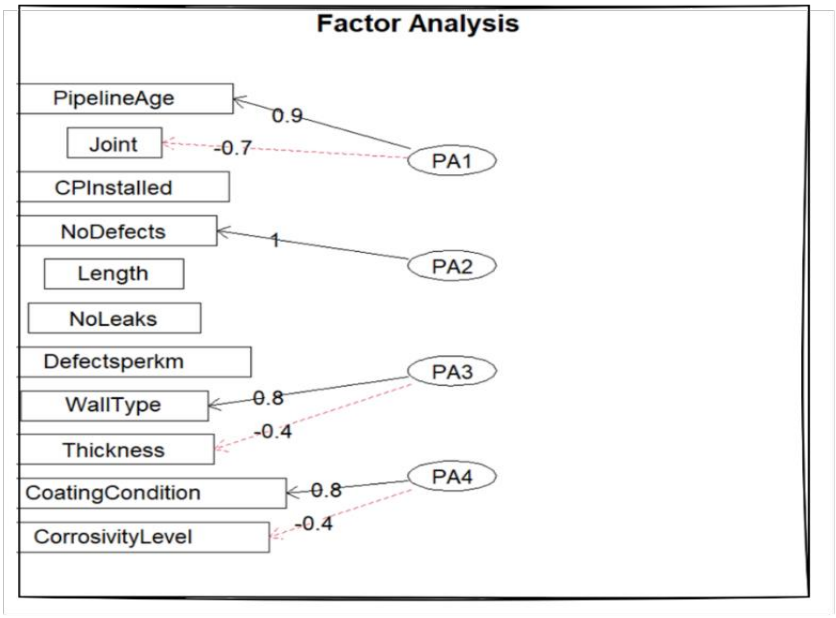

Figure 6. Four Factor Analysis Diagram

According to Figure 6, **Parallel Analysis 1**: Pipeline Age: Loading of 0.9; Joint: Loading of -0.7. This suggests that PA1 is strongly associated with the age of the pipeline and the type of joints used. A positive loading for Pipeline Age suggests that older pipelines are a significant factor, while a negative loading for Joint indicates that certain types of joints might be less favorable.

Parallel Analysis 2: Cathodic Protection Installed: Loading of 1 and Number Defects: Loading of 0.8. This suggest that PA2 is highly influenced by the presence of cathodic protection (CP) installations and the number of defects. A perfect loading for CP Installed suggests that having CP installed is crucial, and a high positive loading for No Defects indicates that fewer defects are beneficial.

Parallel Analysis 3: Length: Loading of -0.6, No Leaks: Loading of 0.8 and Defects per km: Loading of -0.4. These results suggest that PA3 is related to the length of the pipeline, the absence of leaks, and the number of defects per

kilometer. A negative loading for Length and Defects per km suggests that longer pipelines and more defects per kilometer are less desirable, while a positive loading for No Leaks indicates that fewer leaks are advantageous.

5.3 Parallel Analysis 4: Wall Type: Loading of -0.8, Thickness: Loading of 0.4, Coating Condition: Loading of 0.6, and Corrosivity Level: Loading of -0.4. Therefore, this means PA4 is influenced by the type of wall, the thickness of the pipeline, the condition of the coating, and the level of corrosivity. Negative loadings for Wall Type, Coating Condition, and Corrosivity Level suggest that certain wall types, poor coating conditions, and high corrosivity levels are detrimental, while a positive loading for Thickness indicates that thicker pipelines are beneficial. To improve the integrity management of big-diameter underground steel pipelines, it is essential for the Water Utilities to comprehend the fundamental elements affecting pipeline performance. This study utilised factor analysis to identify and classify essential characteristics influencing underground big diameter steel pipelines into four primary assessment categories: Age and Integrity (PA1), Construction and Design (PA2), Material and Protection (PA3), and Performance and Risk (PA4) as outlined in Figure 6.

To develop the final model, the four factors were further grouped into themes using the thematic method. As a result, the model in Figure 7 resulted. These themes encompass several physical, operational, and environmental characteristics, including pipeline age, joint type, wall thickness, coating condition, and defect prevalence.

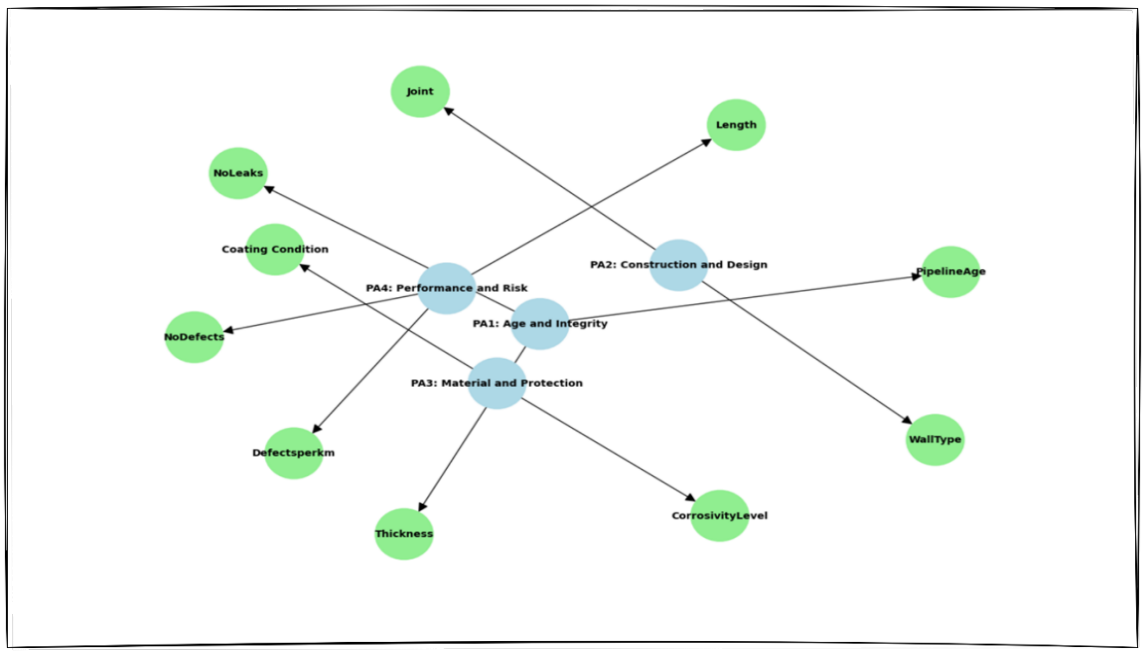

Figure 7. Final Model with Themes

The model offers a systematic framework that will assist them in assessing pipeline conditions by examining the strength and direction of correlations among these variables and their associated factors. This thematic classification enhances comprehension of pipeline vulnerabilities and aids in data-driven decision-making on maintenance prioritising and risk mitigation techniques. Therefore, this is the proposed model for Water Utilities to prioritise their scarce resources for underground steel pipeline condition assessment and improvement plans.

5.4 Proposed Improvements

If used alone as a machine learning technique, it is evident that Factor Analysis cannot produce a much-needed predictive model for the underground steel pipeline. The literature seems to suggest combining it with other techniques to advance its findings further. For instance, Phan and Duong (2020) state that Principal Component Analysis, in conjunction with machine learning approaches like Adaptive Neuro-Fuzzy Inference Systems, can be

employed to anticipate the burst pressure of defective subterranean steel pipelines. Ji et al. (2021) assert that underground steel pipeline leak detection can be integrated with Self-Organising Maps neural network techniques. In conclusion, Lapin et al. (2021) propose developing a continuous monitoring system, necessitating a predictive model that may be utilised to establish a system for the predictive analysis of pipeline degradation. It's against this background; therefore, it is suggested that this study be taken forward and a predictive model be produced using various machine learning techniques as indicated by the literature.

6. Validation

Validation is essential in confirming the sufficiency and comprehensibility of a Principal Component Analysis (PCA) model. Before component extraction, evaluating the dataset's appropriateness for dimensionality reduction is crucial. Two often utilised statistical tests for this objective are the Kaiser-Meyer-Olkin (KMO) measure of sample adequacy and Bartlett's test of sphericity. The KMO statistics assess the variance ratio across variables that may represent common variation, with values approaching 1 suggesting that PCA will likely produce unique and reliable components. Conversely, scores below 0.5 indicate that the data may be unsuitable for factor analysis. Bartlett's test enhances this by assessing if the correlation matrix significantly deviates from an identity matrix, thus validating the existence of underlying structures appropriate for extraction. Collectively, these assessments establish a solid basis for corroborating the assumptions of PCA and guaranteeing the dependability of the resultant components.

The overall KMO for this study is 0.47, indicating poor suitability, while values below 0.5 imply that the dataset may not be appropriate for factor analysis. Nonetheless, Bartlett's Test of Sphericity yielded a Chi-Square value of 246.90 with a p-value less than 0.001, suggesting that the correlation matrix significantly deviates from an identity matrix. According to 5.7, these results affirm the suitability of factor analysis despite the low
KMO.

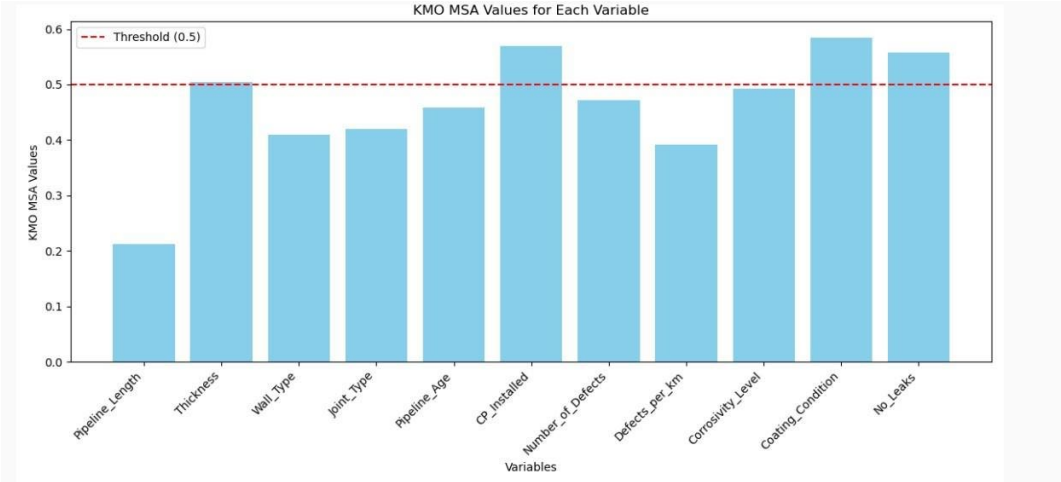

Figure 8. KMO Values Plot for All the Variables

According to figure 8, the minimum threshold deemed acceptable for factor analysis is indicated by the red dashed line at 0.5. Pipeline Length, Wall Type, Joint Type, and Defects per km are variables that do not meet the criteria, suggesting their potential inappropriateness for inclusion in a factor analysis. The elevated levels of specific variables, including Coating Condition, CP Installed, and No Leaks, surpass the threshold, while the Wall Thickness meets the threshold of 0.5, suggesting their appropriateness for factor analysis.

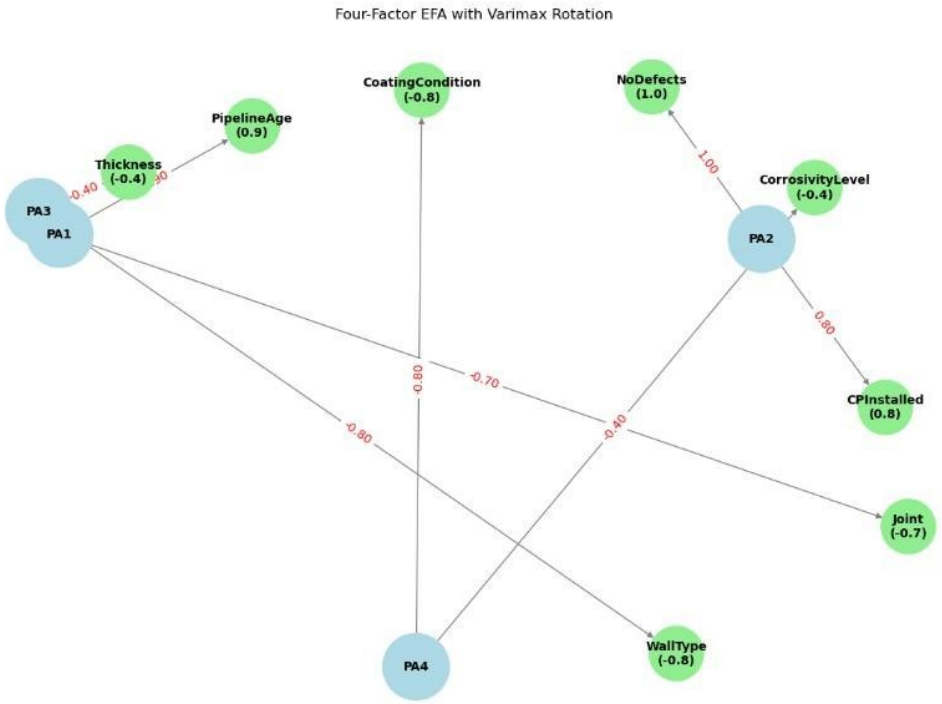

Figure 9. EFA Four Factor Final Model with Varimax Rotation

The inspection of figure 9 suggests that the Model Fit Indices for this Varimax rotated model are, Chi-Square = 14.5, p = 0.63, suggesting that the model fits well, RMSA = 0.03, TLI = 1.043, RMSEA = 0.00. Despite the low overall KMO, the model fits the data well statistically.

7. Conclusion

This study was commissioned to contribute to the comprehension and management of big-diameter underground steel pipelines using factor analysis to a condition assessment dataset collected from the Water Utility that boosting with have largest underground steel pipeline network. The research effectively achieved all its specified objectives, each of which is detailed below along with the relevant results:

Objective 1- To determine the importance of factors contributing to pipeline failure:
The objective was accomplished by a comprehensive statistical analysis of big-diameter underground pipeline parameters and a diverse condition assessment data. The correlation study demonstrated robust associations among critical variables: namely, a positive correlation between pipeline age and soil corrosivity, and a negative correlation between cathodic protection (CP) installation and defect frequency. The findings validated the significance of variables such as age, coating condition, and CP status in affecting pipeline degradation, hence affirming their incorporation in the factor model.

Objective 2 -To extract the primary latent variables by using four-factor analysis:
The study employed exploratory factor analysis (EFA) to derive four latent components that represent the principal characteristics of pipeline degradation. These included:

a) PA1: Age and Integrity — documenting pipeline age, wall thickness, and leak history.
b) PA2: Construction and Design – concentrating on joint type and wall configuration.
c) PA3: Material and Protection – including coating integrity and environmental corrosiveness.
d) PA4: Performance and Risk – including defect frequency, pipeline length, and fault history.

The scree plot and eigenvalue analysis justified the retention of four components, consistent with the elbow criterion and maintaining a balance between dimensionality reduction and interpretability.

Objective 3 **-To develop and validate a Four-Factor Model:**

The resultant model was evaluated by Varimax rotation, producing robust model fit indices (Chi-Square = 14.5, p = 0.63; RMSEA = 0.00; TLI = 1.043), signifying exceptional statistical strength. Although the overall KMO was low (0.47), Bartlett's Test of Sphericity validated the dataset's appropriateness for factor analysis. The model was then enhanced into thematic categories, providing a pragmatic decision-support instrument for South African water utilities to prioritise repairs and distribute resources efficiently.

This research presents a scalable, data-driven methodology for assessing the status of underground large-diameter steel pipelines, incorporating many diagnostic tools into a cohesive analytical model. It improves the interpretability of intricate underground steel large-diameter pipeline infrastructure data and facilitates proactive asset management. Future study could enhance the approach by:

a. Integrating predictive machine learning technologies, such as neural networks and ensemble models, to estimate failure rates.
b. Augmenting the dataset to encompass additional utilities and geographic areas for enhanced applicability.
c. Integrating real-time monitoring data to shift from periodic evaluation to continuous condition surveillance.

This study establishes groundwork for more comprehensive, resilient, and economical management of pipeline infrastructure in resource-limited settings.

References

Akhtar-Danesh, N. *Factor rotation techniques in principal component analysis*, Journal of Data Science, pp. 123–135, 2023.

Al-Kadi, A., Al-Mashaqbeh, I., & Al-Zoubi, A. *Water leakage detection in underground pipelines using acoustic techniques*, Water Resources Management, pp. 4121–4134, 2013.

Ali, M., & Choi, C. *Leak detection in water distribution networks: A review*, Water, pp. 1–15, 2019.

Andrew, M., Pedersen, P., & McEvoy, P. *Research methods and design in psychology*, SAGE Publications, 2011.

Atef, M., El-Zahab, S., & Zayed, T. *Leakage detection in water pipelines using data-driven techniques*, Journal of Performance of Constructed Facilities, pp. 04016042, 2016.

Blanche, M. T., Durrheim, K., & Painter, D. *Research in practice: Applied methods for the social sciences*, Juta and Company Ltd., 2007.

Chapman, D. *SmartBall technology for leak detection*, Pipeline Inspection Journal, pp. 34–39, 2012.

Chen, G., Zhang, Y., & Wang, H. *Corrosion protection of underground pipelines*, Corrosion Science, pp. 620– 628, 2009.

Chen, Y., Liu, J., & Wang, X. *Dimensionality reduction in pipeline condition monitoring using PCA*, Journal of Infrastructure Systems, pp. 04020045, 2020.

Corrpro. *External corrosion direct assessment (ECDA) for underground pipelines*, Corrpro Technical Bulletin, 2010.

Creswell, J. W., & Creswell, J. D. *Research design: Qualitative, quantitative, and mixed methods approaches*, 5th ed., SAGE Publications, 2018.

Deo, R., Singh, R., & Kumar, A. *Condition assessment of aging pipelines using machine learning*, Journal of Pipeline Engineering, pp. 45–58, 2023.

Dithebe, K., Musonda, I., & Makhetha, T. *Critical success factors for water infrastructure projects in South Africa*, Journal of Construction Project Management and Innovation, pp. 1895–1912, 2019.

El-Zahab, S., & Zayed, T. *Leak detection in water pipelines using dynamic modeling*, Automation in Construction, pp. 1–10, 2019.

Fan, Y., & Yu, H. *Data limitations in pipeline failure prediction*, Journal of Infrastructure Systems, pp. 04022045, 2023.

Gao, M., & Song, Y. *Cathodic protection and coating systems for underground pipelines*, Materials Performance, pp. 34–39, 2009.

Hassan, M., Ali, M., & Choi, C. *Condition assessment of underground pipelines*, Journal of Civil Engineering Research, pp. 1–10, 2019.

Hashemi, S., Ghasemi, M., & Zarei, M. *Biplot analysis in multivariate statistics*, Statistical Methods in Research, pp. 215–230, 2018.

Hu, Y., Zhang, L., & Wang, X. *KMO and Bartlett's test in factor analysis*, Journal of Applied Statistics, pp. 789– 802, 2015.

Ji, Y., Wang, L., & Zhang, H. *Leak detection in pipelines using PCA and neural networks*, Sensors, pp. 789, 2021.

Jiang, Y., Zhang, L., & Wang, X. *Machine learning-based corrosion prediction for buried pipelines using heterogeneous condition data*, Journal of Pipeline Systems Engineering and Practice, pp. 04023045, 2024. Jolliffe, I. T., & Cadima, J. *Principal component analysis: A review and recent developments*, Philosophical Transactions of the Royal Society A, pp. 20150202, 2016.

Kavengik. *Orthogonal rotation: Enhancing interpretability of principal components using the varimax technique*, Medium.com, 2025.

Kim, J., Lee, S., & Park, H. *Corrosion incidents in underground pipelines in the U.S.*, Corrosion Engineering, Science and Technology, pp. 123–132, 2021.

Kostryzhev, A., Pereloma, E., & Hodgson, P. *Mechanical behavior of large-diameter steel pipelines*, Materials Science and Engineering A, pp. 789–796, 2010.

Kurita, T. *Principal Component Analysis (PCA)*, Springer eBooks, 2020.

Ledesma, R. D., Valero-Mora, P., & Macbeth, G. *Scree plot use in factor analysis*, Psychological Methods, pp. 233–243, 2015.

Lapin, A., Zhang, Y., & Wang, X. *Predictive modeling for pipeline degradation*, Journal of Infrastructure Systems, pp. 04021045, 2021.

Liu, Y., & Meng, Q. *Machine learning for pressure failure prediction in corroded pipelines*, Engineering Structures, pp. 115456, 2025.

Livingston, D., Chapman, D., & Smith, R. *SmartBall acoustic leak detection*, Pipeline Technology Journal, pp. 45–52, 2019.

Malibary, H., Alghamdi, A., & Alzahrani, M. *KMO and Bartlett's test in PCA*, Journal of Applied Mathematics and Statistics, pp. 56–63, 2019.

McDonnell, J., & Onuoha, C. *ECDA for underground steel pipelines*, Pipeline Integrity Journal, pp. 22–29, 2015. Mielke, R. D., Keil, B. D., Cornwell, E., & Davidenko, G. *Construction of critical 96 and 84 steel water pipeline in Houston, Texas*, Pipelines 2023: Planning and Design, Orlando, Florida, USA, June 17–19, 2023.

Migenda, J., Zhang, Y., & Wang, X. *PCA in pipeline condition monitoring*, Journal of Infrastructure Systems, pp. 04021023, 2021.

Muggleton, J., Brennan, M., & Gao, Y. *Condition assessment of buried pipelines*, Journal of Sound and Vibration, pp. 1–16, 2016.

Nicholson, D. *Corrosion protection for buried pipelines*, Corrosion Engineering, pp. 45–52, 2006.

Nguyen, T., & Holmes, D. *Using the elbow method in PCA*, Journal of Data Science, pp. 345–359, 2019. Onuoha, C., McDonnell, J., & Smith, R. *DCVG and ECDA in corrosion assessment*, Pipeline Integrity Journal, pp. 33–41, 2021.

Ordóñez, F., et al. *PCA for metal loss estimation in pipelines*, Journal of Pipeline Engineering, pp. 45–56, 2015. Phan, T., & Duong, H. *PCA and ANFIS for burst pressure prediction*, Engineering Failure Analysis, pp. 104375, 2020.

Rohe, K., & Zeng, M. *Vintage factor analysis with Varimax performs statistical inference*, Journal of the Royal Statistical Society Series B, pp. 1037, 2023.

Ruiters, C., & Amadi-Echendu, J. *Challenges in South African water utilities*, Water SA, pp. 1–10, 2022.

Ryu, J., Kim, H., & Lee, S. *Eigenvalues and PCA visualization*, Journal of Data Visualization, pp. 89–101, 2022. Sarwar, M., Khan, A., & Ali, M. *Pipeline integrity management strategies*, Journal of Pipeline Systems Engineering and Practice, pp. 04024012, 2024.

Shipilov, S. A., & May, R. *Corrosion failure in underground pipelines*, Corrosion Science, pp. 1235–1248, 2005.

Silva, R., et al. *Factor analysis in infrastructure studies*, Journal of Civil Engineering Research, pp. 1–12, 2017