

## Simon Holcombe

Simon has over ten years' experience in the banking finance industry utilising a range of analytics in the design and administration of infrastructure, including the development of classification models to detect fraud. He also has over 20 years experience in the education sector where he is currently the Academic Director of the highly ranked Master of Business Analytics (MBA) at Melbourne Business School, the graduate business school of the University of Melbourne, a position he has held for 7 years. As part of his role as academic director, he is responsible for student selection, curriculum, learning outcomes and liaising with business partners to coordinate, on average, 15 student internships each year. In addition to this he regularly teaches statistics and mathematics related subjects in the MBA and the MBA.

## Executive Summary

This report provides a comprehensive analysis of historical data on volumetric assets within the Powercor, CitiPower, and United Energy networks to forecast asset replacement volumes for the years 2026 to 2031. This analysis was undertaken by Wei Chang and Junze Li, alumni of the Master of Business Analytics degree at Melbourne Business School, the graduate business school of the University of Melbourne, and was completed in October 2023. The report focuses on uncovering insights from historical data to identify variables influencing asset replacement patterns and to forecast the replacement volume. Volumetric assets are categorised into two primary business units: Line Assets and Plant Assets, each containing various asset categories with distinct characteristics. The data is further segmented into specific groupings for independent analysis, such as high voltage (HV) and low voltage (LV) crossarms, and different types of transformers (indoor, kiosk, pole top, and ground).

The analysis follows a conventional analytics pipeline, including data preprocessing, feature selection, model building, model evaluation, and forecasting. As the data availability and historical trends varies across each asset and sub-asset classes, multiple methodology such as linear regression, averages, gradient boosting etc. were tested to determine the best fitting model for the corresponding group of assets. Line Assets has a richer dataset allowing for more sophisticated modelling, yielding deeper insights. In contrast, the limited data availability for Plant Assets led to relatively higher prediction errors in the forecasted volume numbers.

## Line Asset

For wood poles, replacement volumes are forecasted using a decay rate model, which predicts each wood pole's diameter and sound wood thickness based on features such as age, maintenance history, and species. This new model allows for individual predictions for each wood pole and accounts for changes in the decay rate as the pole ages. This represents a significant improvement over the original methodology, which relies mainly on categorized wood poles, and provides additional insights into asset decay behaviours. The utilization of rich data and advanced modelling highlights the importance of regular inspections and data collection.

Replacement volumes for crossarms, insulators, and services are predicted using defect rates modelling. The model breaks down asset defect performance by different locations over time, allowing trends to be observed across various areas for each asset group. By considering location and time, the new model provides more confidence in predicting replacement volumes. The observed insights also consolidate the business's understanding of each asset group, allowing targeted inspections and maintenance plans for specific locations.

## Plant Asset


The analysis of transformers and switchgear focuses on predicting asset defect rates and asset populations of each asset subclass, a two-step prediction before getting to the replacement volume. The reliability of this statistical analysis heavily depends on the variability of the limited data that was available. For transformers, the historical defect rate for the PAL network demonstrates more consistency compared to the CP network, leading to a more reliable prediction with lower relative error. When comparing transformer subclasses, pole top transformers show the lowest and most consistent defect rate, providing a more dependable result. Further investigation into oil leak related defects reduced variation, serving as a reliable lower bound for replacement volume forecast.

Switchgear followed a similar analysis although due to limited data an average of the 3-year historical defect rate was used for all switchgear subclasses.

To reduce the two-step prediction error, asset population estimates can be enhanced by incorporating additional information, such as upcoming projects and future housing data. This approach helps mitigate the errors identified in the two-step prediction process. Additionally, leveraging data related to equipment specifications, environmental conditions, and loading conditions will further enhance the predictive capabilities for transformer and switchgear defects.

Time series analysis was conducted for underground asset classes using data spanning 72-months. The monthly aggregated data did not show a clear trend.

Therefore, while several predictions were obtained within the plant asset class, careful attention should be paid to model validity, and adopting the recommendations outlined in the report for future refinement is encouraged.

AUTHOR		
Name	Signature	Date:
Simon Holcombe		August 30, 2024





# EDPR FORECAST MODELING


Line Asset & Plant Distribution Asset

Junze Li, Si-Wei Chang

## Table of Contents

<b>1. Background</b>	<b>2</b>
<b>2. Data Exploration and Methodology</b>	<b>3</b>
2.1. Data Exploration	3
2.2. Analytics Pipeline	4
2.3. Data Preprocessing	5
2.4. Feature Selection	7
2.5. Model Building	8
2.6. Model Evaluation	11
2.7. Model Insight: Feature Importance	12
<b>3. Modelling Process and Empirical Results</b>	<b>13</b>
3.1. Line Assets: Wood Poles	13
3.2. Line Assets: Crossarms, Insulators, Services	15
3.3. Plant Assets: Transformers, Switchgears	19
3.4. Plant Assets: Underground Asset	25
<b>4. Key Findings and Analysis</b>	<b>27</b>
4.1. Line Assets: Decay Rate Summary	27
4.2. Line Assets: Find Rate Comparisons	27
4.3. Plant Assets: Asset Defect Rate	30
4.4. Plant Assets: Defect Volume	31
<b>5. Recommendations</b>	<b>31</b>
Plant Distribution	32
Line Assets Comparison	32
Compliance with AER Requirements & Recommendations for Improvement	32
<b>6. References</b>	<b>33</b>
<b>7. Appendices</b>	<b>34</b>

<b>AUTHORS</b>		
<b>Name</b>	<b>Signature</b>	<b>Date:</b>
Si-Wei Chang		August 29, 2024
Junze Li		August 29, 2024

<b>REVIEWED</b>		
<b>Name</b>	<b>Signature</b>	<b>Date:</b>
Simon Holcombe		August 30, 2024

# 1. Background

This report undertakes a comprehensive analysis of historical data pertaining to volumetric assets within the Powercor, CitiPower, and United Energy networks. The objective is to forecast asset replacement volumes for the years 2026 to 2031, a crucial element for the upcoming submission and review with the Australian Energy Regulator (AER). Network assets naturally deteriorate over time and through use. This deterioration can compromise the condition of the asset and increase the risk of asset failure leading to disruptions affecting safety and network reliability. Timely replacement is essential to maintain networks' reliability.

The aim of this analysis is to harness insights from historical data, thereby uncovering the insights and variables that influence asset replacement patterns. The volumetric asset is segmented into two distinct business units: Line Assets and Plant Assets. Within each business unit, diverse asset categories exist, where variations of characteristics are also present within each asset class. Consequently, the data is further organized into groupings, facilitating independent analysis. For instance, within the category of wooden crossarms, further subdivisions consider high voltage (HV) and low voltage (LV) crossarms, while transformers are categorized into indoor, kiosk, pole top, and ground transformers. Each of these groupings is subject to individualized analysis, with their findings thoroughly documented in the report.

The outcome variable varies depending on the data availability for each asset grouping. Ultimately, the overarching objective remains constant—forecasting asset replacement volumes. The analysis approach differs, contingent upon the nature of the asset. For instance, assets like poles are scrutinized at an individual equipment level, with a focus on understanding the decay rate. In contrast, assets subject to rigorous inspection programs are assessed in terms of the defect find rate.

This report is focused on exploring and identifying the most suitable forecasting method or model for each asset classes. These techniques are carefully outlined in the context of its capabilities and constraints. Furthermore, to achieve the most dependable and accurate forecast values, the associated error with each methodology is presented. This transparency allows for a thorough understanding of the reliability and precision of the forecasting process, contributing to well-informed decisions. It's important to note that this report concentrates on the methodology, and the output data may evolve as improved data is incorporated.

## 2. Data Exploration and Methodology

### 2.1. Data Exploration

There are two major categories of assets: Line Assets and Plant Assets. Each category exhibits distinct characteristics, is managed by a different team, and follows a unique data collection process. Due to these differences, Line Asset and Plant Asset are processed and analysed separately. The following sections provide a high-level summary of each asset group and outline the key characteristics of each asset.

#### 2.1.1. Line Assets

Line assets comprise four types: wood poles, crossarms, insulators, and service lines, which are distributed and managed by CP, PAL, and UE.

For wood poles, detailed equipment feature data is available. This includes measurements of sound wood thickness (available for CP, PAL, and UE) and wood pole original diameters (available for CP and PAL), tracked over time. The features considered to have a significant impact on equipment durability and decay behavior include:

- Pole species (e.g. messmate)
- Pole type (e.g. wood untreated dress)
- Pole classification (e.g. durability class 1)
- Age group (e.g. 51-60 years)

For crossarms, insulators, and service lines, equipment feature data, inspection data, and defect notification data are accessible for CP and PAL. UE lacks inspection data at asset class level; however, the inspections of these assets can be correlated with the pole inspections, for which data is available. For these three assets, impactful features include:

- Location (maintenance planner group)
- Year

Forecast inspection volumes are also available for all line assets.

#### 2.1.2. Plant Assets

Plant Assets comprise of three main types: transformers, switchgear, and underground assets, which are distributed and managed by CP, PAL, and UE.

For transformers and switchgear, all three networks have access to

- Defect Notification Data
- Asset Population Data

For CP and PAL, these datasets provide the necessary granularity to classify assets into their respective subclasses. Each subclass exhibits distinct characteristics, and thus analysed separately. However, UE lacks the subclass granularity, thus was analysed on an asset level without diving down into the subclasses.

Asset Type	Asset Subclass
Transformer	Pole top, Kiosk, Indoor, Ground
Switchgear	ACR, Air Break- Indoor, Air Break - Pole, Gas Switch - Pole, HV Isolator, LV Circuit Breaker, RMU (Ring Main Unit)

In transformer analysis, defect notification data provides insights into the reason behind defects. Since oil leakage can have a significant implication for transformers, the defect notifications that were linked to leakage have a separate analysis. Similarly, this detailed analysis only applies for CP and PAL as the UE data did not provide this level of detail.

For the switchgear analysis, it is important to note that the asset population data offers a snapshot as of 2023 for all three networks. Details regarding how the absence of the population data is dealt with are outlined in the Data Cleaning in 2.2.1. All switchgear defect notifications regardless of the reason are treated equally in the analysis for all networks.

Underground assets comprise various associated classes: subtransmission cable, HV cable, LV cable, pits, and pillar. Defect notification data serves as the foundation for analysis across all networks where a time series analysis was conducted to forecast the replacement volume. Subtransmission cable analysis pertains only to CP and PAL networks.

### 2.2. Analytics Pipeline

Taking into consideration variations in data availability and asset characteristics, we explored various outcome variables to predict replacement volumes. Each outcome variable requires a distinct methodology to achieve the goal of forecasting replacement volume.

We followed the conventional analytics pipeline (Figure 1) to forecast the respective outcome variable, which will be discussed in this section. This process includes key stages such as data preprocessing, feature selection, model building, model evaluation, and finally forecasting and extracting insights from the chosen best model. Adjustments are made to align with the unique data and characteristics of each asset, utilizing diverse techniques tailored to their specific contexts. The following section will offer an in-depth exploration of each technique, outlining their advantages and limitations.

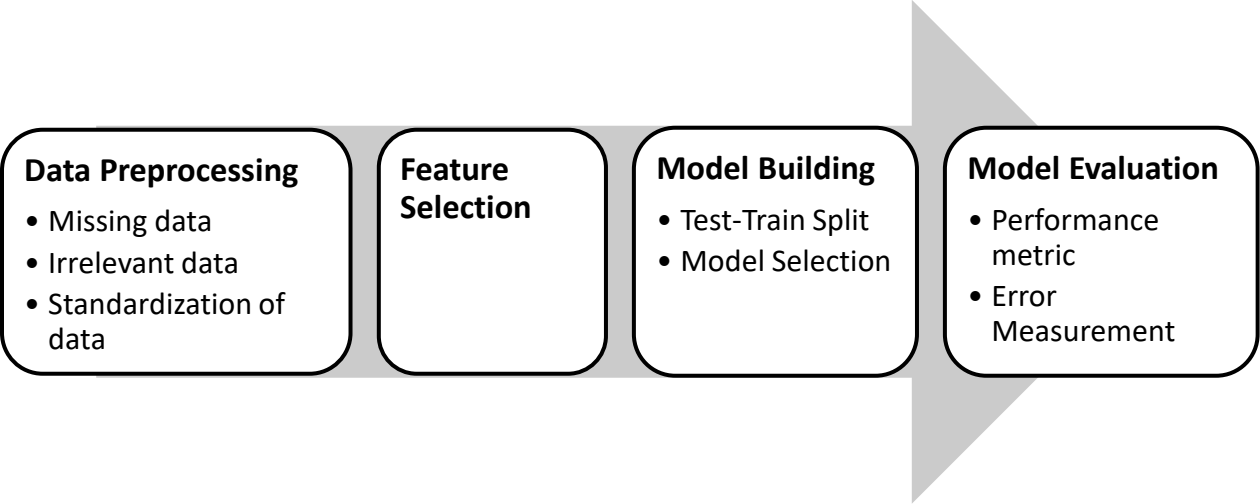


Figure 1: Data Analytics Pipeline

#### 2.2.1. Outcome Variable

##### *Decay Rate*

The decay rate signifies the unit/percentage decrease in measurement per year, specifically relevant to wood poles. Utilizing historical data to understand the features contributing to decay provides insights



into the factors influencing deterioration. Applying this understanding to the entire population enables the forecasting of replacement volumes for wood poles.

#### *Defect Find Rate*

The defect find rate is defined as the number of defects found per number of inspections conducted. This outcome variable is applicable to asset classes with a standardized inspection program and documented inspection data. By considering the defect find rate in conjunction with future inspections, we can estimate the replacement volume for the specific asset class of interest.

#### *Asset Defect Rate*

The asset defect rate is defined as the number of defects per total asset population. This outcome variable is relevant for asset classes with available asset population data and a consistent trend in the change of asset population. To determine the replacement volume based on the asset defect rate, the defect rate is multiplied by the estimated future asset population. It's crucial to highlight that the future asset population is either forecasted or explicitly stated within the business context.

#### *Defect Volume*

Defect volume is the number of defects identified. This is suitable for asset classes where only the defect notification is available. The defect volume is particularly relevant for underground assets, and the forecasted defect volume is utilized as an indicator for replacement volume.

#### *Defect (binary variable)*

The "defect" is a binary variable used to predict the probability of a defect based on a set of asset characteristics. This outcome variable is well-suited for asset classes with a comprehensive set of features. Setting a threshold for the probability of a defect allows the forecasting of the replacement volume for the specific asset class of interest.

Given the characteristics of the assets and the available data, the table below outlines the specific outcome variable modeled for each asset class.

<b>Asset Class</b>	<b>Outcome Variable</b>
Line Asset: Wood Poles	Decay Rate
Line Asset: Crossarms, Insulators, Services	Defect Find Rate, Defect (binary)
Plant Asset: Transformers	Asset Defect Rate
Plant Asset: Switchgear	Asset Defect Rate
Plant Asset: Underground assets	Defect Volume

*Table 1: Outcome Variable Selection for each Asset Class*

## 2.3. Data Preprocessing

To prepare the data for analysis, the data preprocessing is a crucial step to ensure the integrity and reliability of the dataset. Key steps include standardizing data for uniformity, imputing missing values for dataset completeness, and managing irrelevant or outlier data to ensure robustness in subsequent analyses. These measures collectively contribute to a clean and high-quality dataset, fostering more accurate and reliable insights.

### 2.3.1. Line Asset

Wood Pole Decay Rate Model:

- Construction Year between 1900 and 2022
- Decay in measurements needs to be positive.
- Time difference between inspections for the same equipment needs to be at least one year.
- Measurement date needs to be later than construction year.
- Unknown wood pole class is replaced by Class 3.

For sound wood thickness, we required:

- Sound wood thickness needs to be less than 200.

For diameter:

- For CP and PAL diameter data, if there is inconsistency between measurements of the same equipment (i.e., when the measurement increases in one of its inspections), this record is removed.

Additionally, a parameter is set up such that:

- Records with more than a certain number of years in time duration but no measurement changes are removed. The default set up is 10 years.

### 2.3.2. Plant Asset

For transformers analysis, defects associated with cover damage were omitted, considering them as cosmetic issues that usually do not trigger the replacement of the asset. The transformer subclass was also remapped accordingly using the 'RIN\_Type' column in the Asset Population dataset. Additionally, to assess the influence of equipment features, defect notification data needed to be aligned with asset population data to determine corresponding features. Consequently, only 1327 notifications out of a total of 2840 defect notifications could be successfully mapped. Although the incompleteness of the data precluded the use of this analysis in the final forecasting method, it still provided valuable insights for asset management.

For switchgear analysis, the asset population data offers a snapshot as of 2023. For CP and PAL networks, it is assumed that equipment installed on or before 2017 represents the 2017 population count. Newer installations are considered additions to this count, as they are unlikely to be decommissioned. Approximately 6% of the equipment (1470 pieces) lacked subclass information, which was imputed proportionally based on each network's population. For UE network, asset population is assumed to be constant from 2017 to 2023.

For underground assets analysis, defect notifications were used. The number of defect notifications were aggregated monthly, and instances where the monthly defect volume exceeded the 95<sup>th</sup> percentile were addressed by replacing them with the average of the preceding and subsequent datapoints. If subsequent data was unavailable, the replacement was made using the preceding data.

## 2.4. Feature Selection

Feature selection involves choosing the most relevant variables to enhance model performance and interpretability. This process aids in simplifying complex datasets by retaining only the key features that contribute significantly to the predictive power of the model.

### 2.4.1. Variance Analysis

ANOVA was utilized to assess the impact of various factors on dataset variability by investigating whether means of different groups, categorized by these factors, are statistically significantly different. Specifically in this project, ANOVA rigorously evaluated the significance of each factor in explaining outcomes' variability, with F-statistics serving as a key metric to determine significant differences in observed variances among groups and test the hypothesis of equal group means.

The outcomes of ANOVA, particularly the F-statistics, were utilized for feature selection before modeling. These results offered insights into the relative impacts of various factors on dataset variability. The identified influential factors guided the decision-making process, ensuring that subsequent analysis and interpretations were based on robust statistical evidence.

### 2.4.2. Hypothesis testing, F-test, T-test and P-value

Hypothesis testing was central to the Variance Analysis process, providing a structured methodology to assess the statistical significance and validity of the model's results. This approach included the use of F-tests, T-tests, and P-values, each offering distinct insights into the model's robustness and the reliability of its parameters.

#### F-test

The F-test was utilized primarily in the context of ANOVA to assess whether the variances across multiple groups were equal. It was instrumental in evaluating the overall significance of the models, helping to ascertain whether the explanatory variables as a group were impactful in predicting the dependent variable.

#### T-test

T-tests were employed to assess the significance of individual parameters or coefficients in the model. It allowed for the evaluation of whether each variable significantly contributed to explaining variations in the dependent variable, offering a granular perspective on the model's components.

#### P-value

P-values were calculated in conjunction with the F-tests and T-tests, providing a metric to gauge the evidence against the null hypothesis. A low P-value indicates that there is a lower probability that model's accuracy happens by chance, supporting the relevance and impact of the evaluated parameters and the model. A significance level of 0.1 is used in this analysis. This means there is a 10% chance that the observed results, or more extreme ones, could occur if the null hypothesis were true. In simpler terms, this suggests moderate evidence that the predictive model is working.

While a p-value of 0.1 is not conventionally considered strong evidence (as p-values of 0.05 or lower are typically sought after), in the complex and often unpredictable realm of asset prediction, where perfect predictability is nearly impossible due to data availability and numerous influencing factors, a p-value of

0.1 can still be indicative of a potentially useful model. It suggests that the model has a decent chance of capturing a true effect, which, in the challenging field of asset prediction, can be a significant step towards making informed decisions.

Through the application of these hypothesis testing techniques, a rigorous validation of the models was conducted, ensuring that the findings are both statistically sound and practically insightful, enhancing the confidence in the models' predictive capabilities and overall reliability.

## 2.5. Model Building

Model building is the phase where predictive models are constructed based on the available data. It involves critical steps such as selecting appropriate algorithms, dividing the dataset into training and testing sets using techniques like test-train split, and employing cross-validation to assess the model's performance across multiple subsets. The objective is to create a robust and generalizable model that can accurately predict outcomes on new, unseen data.

### 2.5.1. Test – Train Split

The test-train split is a fundamental methodology applied in this project to validate the performance and generalizability of our models. This technique involves partitioning the dataset into two subsets: a training set used for fitting the model, and a test set used for evaluating its predictive accuracy.

By adopting this approach, an unbiased evaluation of the model's performance was achieved, ensuring its capacity to make reliable and accurate predictions on new, unseen data, and confirming that the model was not overfitting to the noise or idiosyncrasies of the training data.

In the implementation of this project, a conventional 80-20 split was employed, allocating 80% of the data for training and the remaining 20% for testing. For non-time-series data, a random split approach was utilized to create the training and test subsets, ensuring a diverse and representative sampling for model validation. In the case for time-series data, a chronological split was maintained to preserve the temporal integrity of the data. (Figure 2)

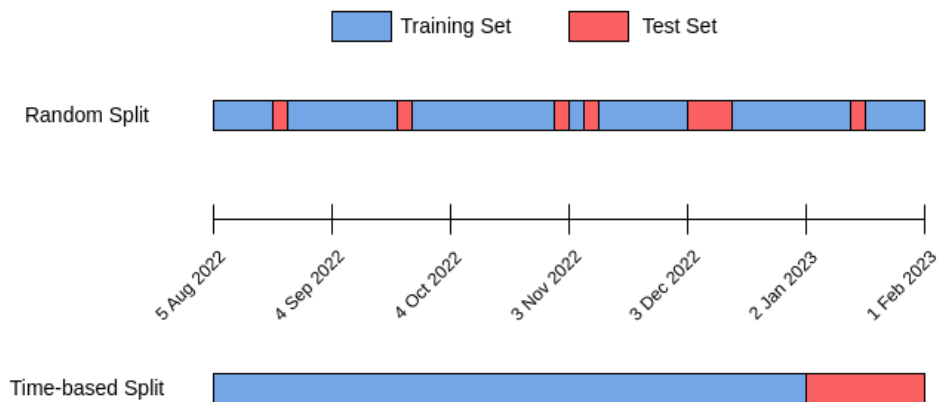


Figure 2 <https://www.pi.exchange/knowledgehub/how-are-the-models-evaluated-in-the-customer-churn-prediction-template>

### 2.5.2. Cross-Validation

Cross-validation was used to enhance the robustness and generalizability of the models' performance. It entails partitioning the dataset into multiple subsets, iteratively training the model on a combination of

these subsets and validating it on the remaining data. This process circulates multiple times, providing a comprehensive evaluation of the model across various segments of the data.

In contrast to a single train-test split, cross-validation mitigates the risk of overfitting and biases associated with a specific, random partitioning of the data. This is particularly relevant in complex models where there is a risk of overfitting, ensuring that the model retains its predictive accuracy and generalizability across diverse data subsets.

In this project, cross-validation was predominantly applied in the context of decay rate models, where advanced modeling techniques such as multiple linear regression, random forest, and gradient boosting were employed. Given the complexity of these models, a 20-fold cross-validation was adopted. This intensive cross-validation process facilitated the optimal selection of hyperparameters, ensuring the models' robustness and enhancing their predictive integrity and generalizability.

### 2.5.3. Model Selection

#### *Averages*

Utilizing arithmetic averages of historical data points for future forecasting is a fundamental but powerful method. This method is suitable when data is limited or appears randomly distributed without clear trends or patterns. By computing the average, we extract a central tendency from historical data, enabling straightforward yet crucial forecasts. Despite its simplicity, this method reveals valuable insights, providing a stabilizing influence amid the complexity and variability of data.

However, there are inherent limitations. While the average serves as a robust measure of central tendency, it may not entirely capture the nuanced variations in the data, especially those linked to equipment characteristics. As a result, the forecasts produced might not be sensitive enough to underlying anomalies or subtle changes in asset conditions, potentially missing crucial aspects of the data's behavior.

#### *Linear Regression*

Linear regression holds a central role in statistical and data analysis, offering a systematic approach to understand the impact of various factors and their interactions.

In this project, both simple (with a single independent variable) and multiple linear regressions (examining the influence of multiple independent variables) were used to model relationships between variables, assuming normally distributed data residuals. One of the prominent advantages of linear regression lies in its transparency and interpretability. The model's coefficients provide a direct interpretation of each factor's unit impact, thereby enhancing the clarity and usability of the model's outputs.

However, it is essential to underscore that linear regression comes with inherent assumptions, such as linearity and the normal distribution of residuals. These assumptions may not align with all datasets, necessitating a thorough examination and validation process to ensure the model's appropriateness and reliability in capturing underlying data patterns and relationships.

#### *Machine Learning Algorithm*

Machine Learning algorithms such as Gradient Boosting and Random Forest have been instrumental in enhancing the predictive accuracy and robustness of our models. These algorithms leverage historical data, learning intricate patterns and relationships, to make powerful predictions and uncover nuanced insights in asset behavior. Gradient Boosting and Random Forest are popular and powerful algorithms, for they are effective in handling diverse datasets and delivering strong predictive performance.

## Gradient Boosting

Gradient Boosting is a powerful ensemble technique that constructs multiple decision trees one after the other, with each tree correcting the errors of the one before it. This method is effective in handling diverse complexities and nuances in the data, enabling the model to capture intricate relationships and dependencies successfully.

In this project, Gradient Boosting was employed due to its capability to minimize bias and variance, delivering models that are both accurate and generalize well to new data. It allows for the consideration of various influential factors, enhancing our understanding and predictive prowess regarding asset decay rates and defect occurrences.

## Random Forest

Random Forest is another ensemble learning method that operates by constructing multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees for unseen data. It offers a robust approach, particularly beneficial due to its natural ability to handle non-linear relationships and interactions between variables, as well as its resilience against overfitting.

In this project Random Forest was used to model and predict decay rates based on the historical data and features of the assets.

Both Gradient Boosting and Random Forest were subjected to a rigorous evaluation process, ensuring model reliability and robustness. The selection of these algorithms was crucial in navigating the complexities of the data, providing a solid foundation for deriving actionable insights and informed decision-making based on the forecasted outcomes.

## *Simple Exponential Smoothing*

Simple exponential smoothing is a time series forecasting method for univariate data without apparent trend or seasonality. In this project, simple exponential smoothing is applied to forecast defect volumes by applying it to the historical defect notifications data.

Simple exponential smoothing prioritises recent data by assigning it more weight through a smoothing factor ( $\alpha$ ). This results in older data having a diminishing influence due to the use of exponentially decreasing weights for past observations. The choice of  $\alpha$  is crucial and requires a model selection process to achieve a reliable forecast. The smoothing factor ranges from 0 to 1, such that when  $\alpha = 1$ , the forecast closely follows the most recent observation, like a naïve forecast, indicating past observation is not indicative of the future. When  $\alpha = 0$ , the model produces a forecast that resembles an average of historical data. The  $\alpha$  provides an understanding of the influence of time for the data (Galit Shmueli and Lichtendahl, 2018).

However, there are inherent limitations such that its predictive power yields a point forecast which is most suitable for one-step ahead forecast and not necessarily for long forecasting horizon. It also assumed that the underlying patterns in the data will persist. Nevertheless, its ability to adapt based on recent data can enhance accuracy compared to a simple average.

### *Logistic Regression*

Logistic Regression is a powerful statistical technique that can be applied to the models where the outcome variable is binary – in this project, specifically, determining whether a defect is present or not. This method is particularly adept at handling categorical outcome variables, making it an ideal choice for this type of analysis.

Logistic Regression is a valuable tool for modeling the probability of a defect occurring, providing in-depth insights into the connections between predictors and the likelihood of a defect. It enables the exploration of complex relationships between explanatory variables and binary outcomes, offering a nuanced understanding of the factors influencing the presence of defects.

The coefficients in the Logistic Regression model are interpreted as the log odds, providing a robust framework for understanding the multiplicative change in the odds of the outcome due to a one-unit change in the predictor. This approach enhances the model's interpretability, allowing for a more straightforward translation of results into actionable insights.

### *Weibull Survival Distribution*

Weibull Survival Analysis stands as a robust statistical methodology, particularly illuminating when exploring the relationships between defect rates and the age of assets in our study. This technique, characterized by its flexibility and adaptability, allows for a nuanced modeling of the time until an event occurs, in this case, the manifestation of defects.

Weibull Survival Analysis can be instrumental in dissecting the intricate interplay between the age of assets and the occurrence of defects. Its application can give a deeper understanding of how the likelihood of defects evolves over time, enabling a more precise characterization of equipment durability and longevity.

The Weibull model, with its shape and scale parameters, offers insightful interpretations regarding failure rates, allowing for a nuanced exploration of defect patterns across different asset age groups. However, careful attention is required to ensure that the model's assumptions and the data's characteristics are harmoniously aligned, ensuring the reliability and validity of the analytical insights derived.

In this project, emphasis was placed on validating the Weibull model's accuracy, previously developed by the Line Asset team in Excel. The focus was concentrated on confirming the model's fidelity, comparing it against alternative models, and rigorously testing the necessary assumptions to ensure analytical robustness.

## **2.6. Model Evaluation**

Model evaluation is a crucial step in the analytical process, ensuring that the developed models are robust, reliable, and generalizable to new, unseen data. Various metrics and techniques were utilized to validate the models, providing a comprehensive evaluation of their performance and reliability.

### **2.6.1. Error Measurement: Root Mean Squared Error (RMSE)**

The Root Mean Squared Error (RMSE) serves as a crucial metric in evaluating the predictive accuracy of our models. RMSE quantifies the average difference between the model's predicted values and the actual outcomes, providing a reliable indicator of the model's predictive performance. A lower RMSE signifies a higher accuracy in the model's predictions, reflecting a closer alignment between predicted values and actual results.

An essential characteristic of RMSE is that it is expressed in the same units as the original data, facilitating a straightforward interpretation of the error magnitude. However, it's crucial to note that RMSE values are inherently dataset-specific and are most informative when used to compare models applied to the same dataset.

In this project, RMSE played a crucial role in selecting the most suitable models by identifying superior performance and accuracy, guiding the choice for subsequent forecasting tasks.

### 2.6.2. R-squared Value

The R-squared value was instrumental in assessing the explanatory power of linear regression model. It quantifies the proportion of variance in the dependent variable that is predictable by the independent variables, serving as a robust indicator of model fit and predictive capability.

An R-squared value encapsulates the percentage of variation in the data that the model successfully explains. Higher R-squared values signify models with enhanced explanatory power, indicating a more substantial proportion of variability being accounted for by the model.

Various thresholds can be used to evaluate the adequacy of the R-squared value, each applicable based on different considerations and contexts. In this project, a threshold of 0.5 was adopted. This criterion was based on the rationale that a model explaining over 50% of the data variation offers valuable predictive insights, substantiating its utility beyond mere random guessing, and thereby warranting its application in forecasting.

## 2.7. Model Insight: Feature Importance

Feature importance was a crucial aspect of our model evaluation process, predominantly applied to the machine learning models utilized in this project. Unlike linear regression models where coefficients offer direct interpretations of the variables' impacts, machine learning models often lack such straightforward interpretability. Consequently, feature importance becomes instrumental in understanding and interpreting the influence of various variables on the model's predictions.

In machine learning, feature importance varies by model due to their distinct methodologies. In Random Forest, importance is often expressed as decimals, representing the normalized reduction in criterion (like Gini impurity) brought by a feature. This is calculated based on the feature's contribution to the homogeneity of nodes in the decision trees. In contrast, Gradient Boosting models typically report feature importance as whole numbers, indicating the count of times a feature is used in split points of trees. While the numbers are presented in different scales, the higher the importance value, the more important the variable is.

In our approach, feature importance was meticulously utilized to discern which variables wielded the most significant influence on the model's outcomes. This not only facilitated a deeper understanding of the model's decision-making process but also assisted in validating the relevance and significance of the variables included in our models.

By evaluating feature importance, we gained invaluable insights into the relative contributions of each feature, ensuring that the models were operating with optimally relevant and impactful variables. This approach enhanced the robustness and reliability of our models, providing a solid foundation for the interpretation and application of their predictions and insights.



## 3. Modelling Process and Empirical Results

### 3.1. Line Assets: Wood Poles

For wood poles, the outcome variables are the decay rate of Sound Wood Thickness (SWT) and the diameter—crucial metrics used during inspections to assess the wood pole's condition. The line asset team employs a condition-based model that effectively translates SWT and diameter measurements into discernible wood pole conditions, facilitating the estimation of replacement volumes.

The historical measurements of the SWT and diameter are readily available from cyclic inspections. These measurements, coupled with the wood pole features serve as the basis for modeling process. The primary focus is on modeling the decay behavior of SWT and diameter, with an emphasis on capturing decay rates, enhancing prediction accuracy, and enabling flexibility for future improvements in the condition-based model.

The historical measurements were analyzed, calculating the time-differentiated changes in SWT and diameter to ascertain decay rates. Definitions applied include:

- For SWT, decay rate is defined as millimeters per year.
- For diameter, decay rate is defined as a percentage change per year.

#### 3.1.1. Modelling Process: Decay Rates

In modeling decay rates, we considered factors such as wood pole species, type, classification, and age group. With a comprehensive historical dataset available, we tested Linear Regression, Random Forest, and Gradient Boosting (refer to 2.5.3 for model specifications). To ensure robust model evaluation, train-test split, and 20-fold cross-validation were applied, facilitating a comprehensive comparison of model accuracies and the derivation of confidence intervals. The comprehensive assessment of the model performance was then evaluated using Root Mean Squared Error (RMSE) obtained from the 20 cross-validations and one standard deviation to indicate variability.

Historical measurements of Sound Wood Thickness (SWT) are available for assets from CP, PAL, and UE; however, diameter measurements are only accessible for CP and PAL. Given this, all diameter forecasts will be primarily based on predictive models derived from CP and PAL data. As UE doesn't have initial diameter measurements for wood poles, the average decay rate from the CP/PAL model will be applied for forecasting.

Utilizing the best-performing models, as identified through cross-validation, we proceeded to forecast the decay rates for individual equipment. This approach accounted for the progressive aging of equipment over time, ensuring a nuanced consideration of temporal influences on decay rates. Subsequently, the forecasted decay rates were applied to the measurements, facilitating the determination of future values for diameters and Sound Wood Thickness (SWT).

These refined estimates were then integrated into the condition-based model, enabling a precise calculation of anticipated defect volumes, thereby enriching the model with forecasted data and enhancing its predictive accuracy and utility.

#### 3.1.2. Empirical Results: Decay Rates

Below is a summary of the model accuracies. In selecting the most effective model for each measurement, considerations were made regarding both the average RMSE and the confidence range. A comparative analysis was conducted across the outputs of all three models for each measurement. The model with the smallest average RMSE coupled with the narrowest confidence range is selected as the most accurate and

robust model. As shown in Table 2, Random Forest model demonstrated superior performance for measurements 1) and 2), while Gradient Boosting proved to be the best model for measurement 3).

Measurement	Model	Cross-validation RMSE
1) Diameter (CP, PAL)	Linear Regression	0.2212 ± 0.2441
	Random Forest	0.2193 ± 0.2434
	Gradient Boosting	0.2220 ± 0.2361
2) Sound Wood Thickness (UE)	Linear Regression	4.5180 ± 0.4230
	Random Forest	4.4684 ± 0.4131
	Gradient Boosting	4.4881 ± 0.4210
3) Sound Wood Thickness (CP, PAL)	Linear Regression	6.1118 ± 1.7079
	Random Forest	6.1113 ± 1.7127
	Gradient Boosting	6.1036 ± 1.7097

Table 2: Decay Rates Model Performance

To gain deeper insights into the model and enhance interpretability, a focused analysis was conducted to determine the significance of each variable in influencing the measurement decay rates. Utilizing the best-performing models, feature importance was computed for each measurement, identifying the variables that had the most influence in determining decay rates. The following Tables show the top ten most influential features for each model, offering a detailed perspective on the variables that predominantly drive the measurement decay rates in the studied models:

Feature	Importance
Serial Number_WOOD CREOS IMPREGNATED	0.501851
decay duration	0.451777
Serial Number_WOOD SALT IMP (GREEN)	0.015662
Maintenance planner group [MPG]_GEE	0.013506
Maintenance planner group [MPG]_CP	0.006839
Model number [EQP NavAtt]_BB-BLACKBUTT	0.005931
Maintenance planner group [MPG]_BEN	0.001369
Maintenance planner group [MPG]_SUN	0.001240
Equipment Type [EQP NavAtt]_P_WOOD_CL1	0.000759
Model number [EQP NavAtt]_GG-MOUNTAIN GREY GUM	0.000511

Table 3: Diameter (CP, PAL) - Random Forest

#### Sound Wood Thickness (UE) – Random Forest

Feature	Importance
age	0.548261
Pole Type_Wood Creos Impregnated	0.136072
Class Of Pole_Class 1	0.112379
Pole Species_White Stringybark	0.040904
Pole Type_Wood Untreated Dressed	0.019489
Pole Species_Mountain Greygum	0.017164
Class Of Pole_Class 3	0.016223
Pole Species_Grey Gum	0.009830
Pole Type_Wood Untreated Round	0.009675
Pole Species_Grey Ironbark	0.009517

## Sound Wood Thickness (CP, PAL) – Gradient Boosting

Feature	Importance
age	137
Planner Group_CP	33
Model No./Species_ZZ-WOOD UNKNOWN	19
ManufSerialNo._WOOD SALT IMP (GREEN)	18
ManufSerialNo._WOOD UNTREATED ROUND	16
Model No./Species_IB-IRONBARK	15
Planner Group_SUN	12
Object Type/Equipment Type_P_WOOD_CL3	11
ManufSerialNo._WOOD CREOS IMPREGNATED	8
Model No./Species_GI-GREY IRONBARK	7

The analysis reveals distinctive contributions of various features across different measurements. Notably, the features incorporated in the model largely demonstrate relevance, with ‘age (duration)’ emerging as a particularly influential factor across all decay rates.

### 3.2. Line Assets: Crossarms, Insulators, Services

The analysis for crossarms, insulators, and services primarily relies on defect notification data, with two outcome variables being modeled. The first approach involves forecasting the defect find rate, determining the number of identified defects per inspection each year. The second approach examines defects as binary variables, focusing on predicting the probability of defect occurrence for each piece of equipment.

#### 3.2.1. Modelling Process: Defect Find Rate

In analyzing the defect find rate for CP and PAL assets, emphasis was placed on two variables believed to significantly influence defect occurrence: the year of inspection and the equipment’s location. The significance of these variables in influencing defect occurrence within each asset group was assessed using Analysis of Variance (ANOVA), facilitating informed decision-making in the model-building process.

To predict the defect find rate from 2026 to 2031, it's crucial to assess the significance of time (year variable). The outcome of the ANOVA helps determine if the year contributed to the variation in the data. If the year shows a significant influence with p-value of 0.1 (refer to 2.6.2), it is then incorporated into the analysis. Given the limited number of data points (ranging from 3 to 5) and the presence of only one explanatory variable (year), linear regression was chosen to model and understand the trend. For asset classes where the year was not significant, an average historical defect find rate at the location level was used as the selected model to forecast the defect find rate.

The efficacy of this model is evaluated based on the R-squared value it yields. A model is deemed satisfactory if the R-squared value exceeds 0.5 (refer to 2.6.2), indicating that over half of the variation is explicable by the model. In scenarios where the model doesn’t meet this threshold, a different approach is adopted. Instead of relying on the regression model, we revert to using historical averages, proceeding to make predictions based on these averages. The final model performance was then evaluated using Root Mean Squared Error (RMSE).

### 3.2.2. Empirical Results: Defect Find Rate

#### *CP & PAL Network*

The table below shows the outputs from the ANOVA analysis applied to each asset group's historical data. The F-statistics were utilized to test the hypotheses, with significance assessed based on p-values. A variable was considered statistically insignificant at a 10% level if the p-value exceeded the chosen threshold, indicating a lack of substantial evidence to confirm its influence on the defect find rate.

		<b>F-statistics</b>	<b>P-Value(&gt;F)</b>	<b>Variance proportion (%)</b>
<b>HV Crossarm</b>	Year	1.2151	3.2569e-01	2.70%
	Location	13.4295	3.5959e-07	79.53%
	Residual			17.77%
<b>LV Crossarm</b>	Year	3.3718	3.4976e-02	7.67%
	Location	12.2112	8.6315e-07	74.12%
	Residual			18.21%
<b>HV Insulator</b>	Year	19.6356	3.7919e-07	73.28%
	Location	0.7043	6.8484e-01	5.26%
	Residual			21.46%
<b>LV Insulator</b>	Year	2.5756	0.0774	8.80%
	Location	7.0069	0.0001	63.86%
	Residual			27.34%
<b>Service</b>	Year	5.2731	2.2311e-03	12.42%
	Location	14.5869	1.0408e-08	68.73%
	Residual			18.85%

The model outputs indicate that year and location together explain over 70% of the data's variation. Location emerges as a significant determinant for all assets, except for HV insulators. Segmenting the forecasts based on location not only enhances explanatory power but also refines the granularity of defect volume predictions. Consequently, the decision was made to proceed with analyses focused on datasets grouped by location.

The defect find rate plots below illustrate the clarification of location and year effects, highlighting the advantages of their separate modeling. This approach allows for the capture of diverse influences on defect find rates. For instance, location 'SHE' shows a notably distinct trend compared to other locations across various assets. Similarly, 'CP' exhibits a considerably lower defect find rate in numerous assets, attributed to geographic advantages. These nuanced differences and trends are more accurately captured and explained when year and location variables are meticulously incorporated into our analyses.

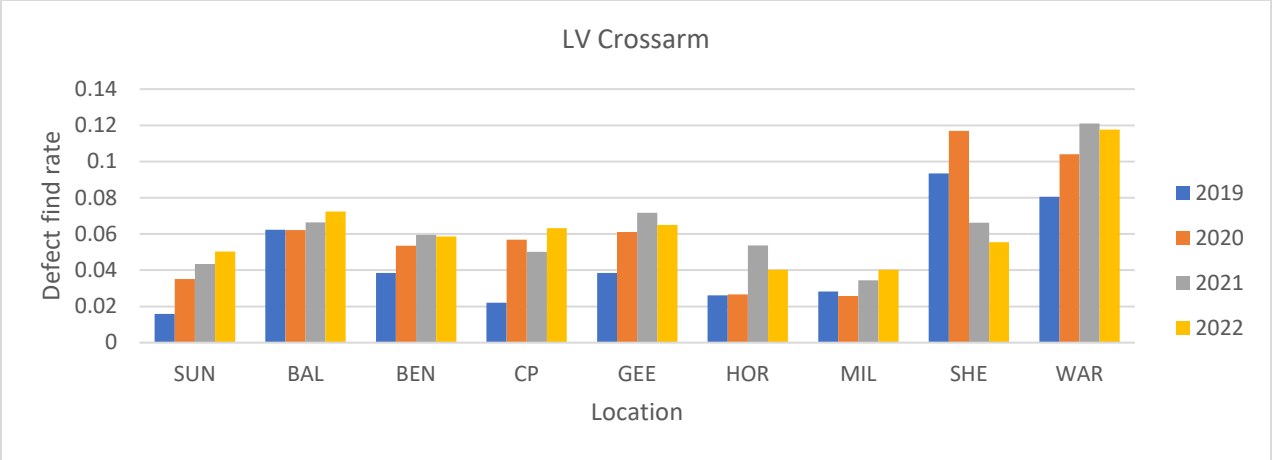
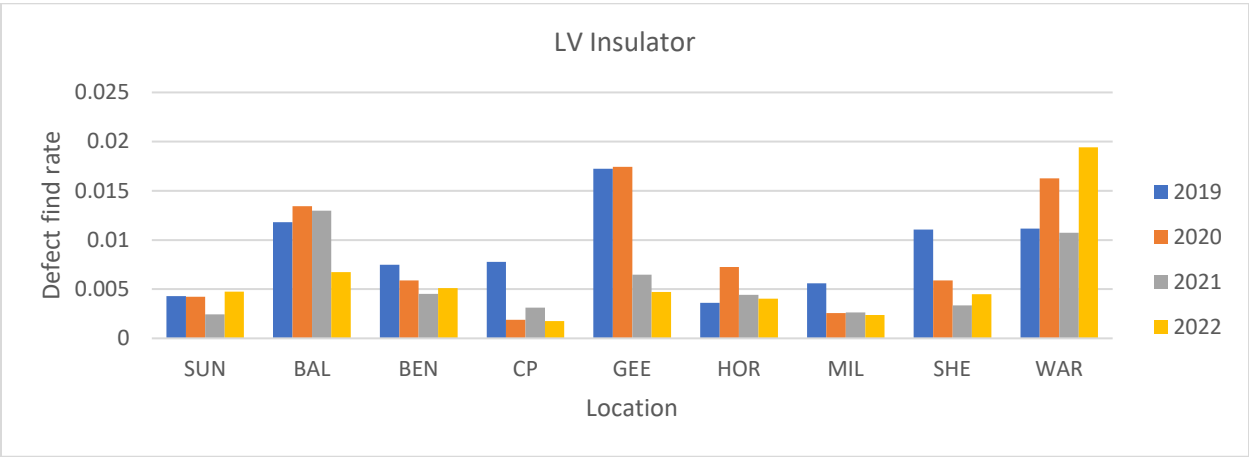
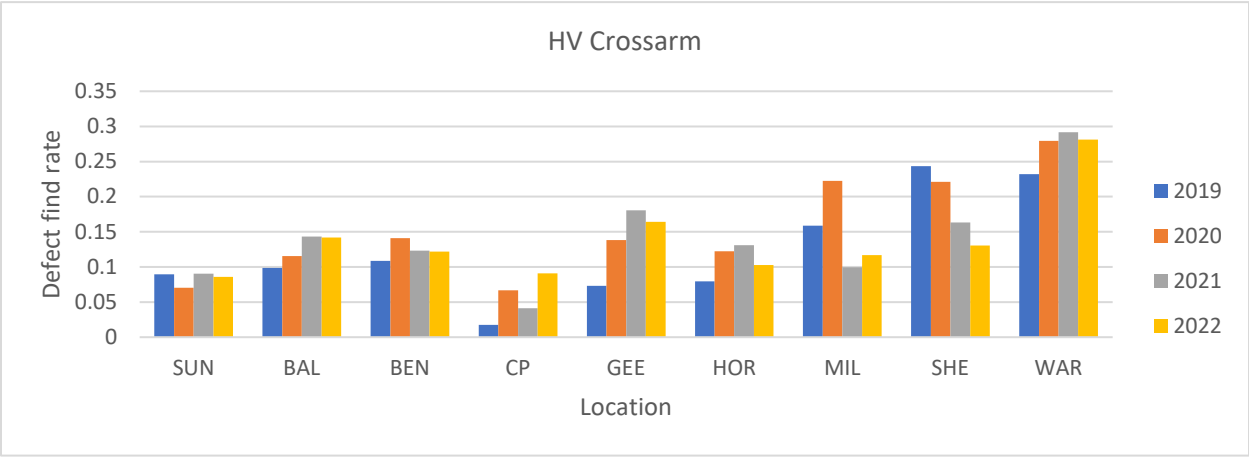
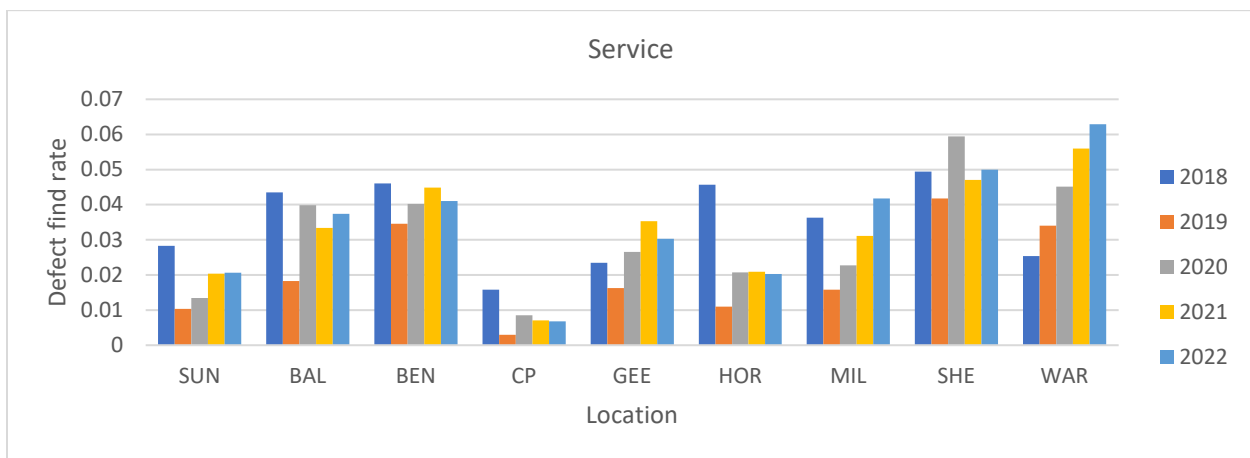
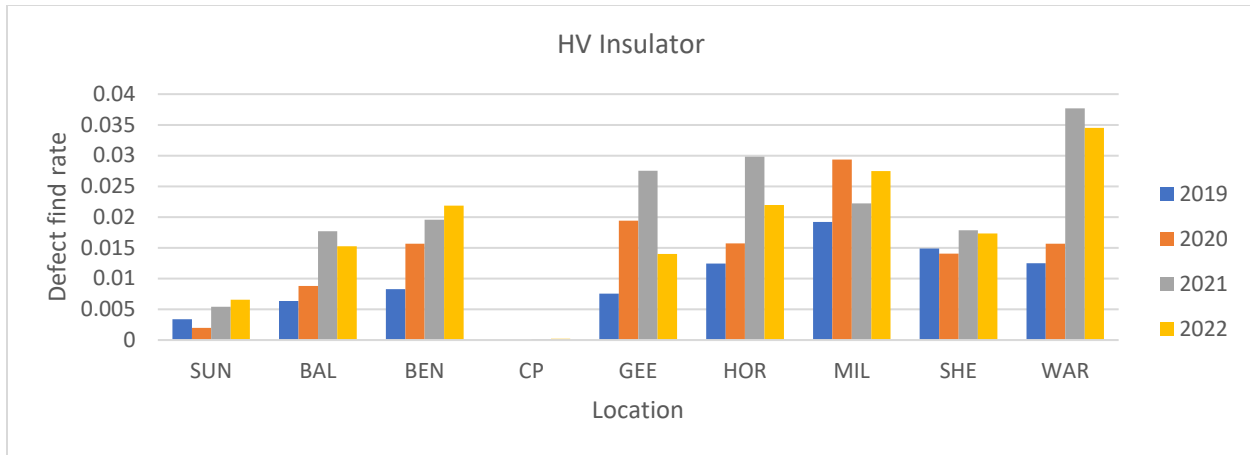


Figure 3: LV Cross Arm Defect Find Rate by Location across 2019 - 2022





The summary below outlines the specific models utilized for each asset type across various locations, reflecting the process detailed in 3.2.1 for determining the most suitable model for predictive analyses.

#### Final model decision

	BAL	BEN	CP	GEE	HOR	MIL	SHE	SUN	WAR
<b>HV Crossarm</b>	AVG	AVG	AVG	AVG	AVG	AVG	AVG	AVG	AVG
<b>LV Crossarm</b>	AVG	AVG	AVG	AVG	AVG	AVG	AVG	AVG	AVG
<b>HV Insulator</b>	AVG	Linear	Linear	AVG	AVG	AVG	Linear	Linear	Linear
<b>LV Insulator</b>	AVG	AVG	AVG	AVG	AVG	AVG	AVG	AVG	AVG
<b>Service</b>	AVG	AVG	AVG	Linear	AVG	AVG	AVG	AVG	Linear

#### Error measurement (RMSE)

	BAL	BEN	CP	GEE	HOR	MIL	SHE	SUN	WAR
<b>HV Crossarm</b>	0.0188	0.0116	0.0274	0.0409	0.0198	0.0473	0.0449	0.0080	0.0231
<b>LV Crossarm</b>	0.0016	0.0040	0.0087	0.0074	0.0113	0.0026	0.0240	0.0033	0.0069
<b>HV Insulator</b>	0.0023	0.0013	0.0000	0.0073	0.0046	0.0041	0.0010	0.0010	0.0052
<b>LV Insulator</b>	0.0027	0.0011	0.0024	0.0059	0.0014	0.0013	0.0030	0.0009	0.0036
<b>Service</b>	0.0088	0.0040	0.0042	0.0044	0.0116	0.0093	0.0057	0.0063	0.0010

### UE Network

For UE’s line assets, location information is not readily available, and there is no trending in defect find rate (see table below, all p-values are in-significant), thus averages of historical data are taken as forecasting rates.

	<b>HV Crossarm</b>	<b>LV Crossarm</b>	<b>HV Insulator</b>	<b>LV Insulator</b>	<b>Service</b>
<b>p-value of year</b>	0.5	0.9	0.1	0.6	0.7

In forecasting future defect volumes, a foundational assumption was made that inspections would occur on a consistent five-year cycle. This assumption implies that the inspection volumes will manifest a repetitive pattern every five years. Guided by this rule, projections for inspection volumes were formulated for the years 2023 through 2031. This was accomplished by replicating the inspection volumes recorded in the preceding five years.

Subsequent to establishing the projected inspection volumes, the forecasted defect find rates were meticulously applied to each asset at each respective location. This process facilitated the generation of comprehensive forecasts pertaining to defect volumes, ensuring that the projections were substantiated by a systematic application of historical patterns and forecasted rates.

#### 3.2.3. Modelling Process: Defect (Binary Variable)

For line assets (crossarms, insulators, and services), equipment feature data is available. This data can be integrated with defect notification data using the equipment number as a common key, enabling a detailed analysis based on combined datasets. Consequently, this allows for the exploration of a binary outcome—‘defect or not’—as the target variable, facilitating the development of models aimed at predicting the probability of a defect occurrence based on asset characteristics.

In this analytical endeavor, both logistic regression and Weibull distribution models were employed. Logistic regression was leveraged to ascertain the impact of each characteristic on the likelihood of a defect, providing insights into the utility of each feature in predicting the outcome. On the other hand, the Weibull distribution was utilized to model the survival behavior of the equipment, focusing on the temporal aspect of the defect occurrences.

While both models are intuitively appealing and theoretically pertinent for modeling the binary outcome, their practical performance was somewhat limited. A significant challenge encountered was the disproportionate representation of the defect occurrences within the dataset—defective equipment constituted a minor fraction compared to non-defective ones. This imbalance hindered the model's ability to glean meaningful insights and accurately characterize the defect behavior, ultimately affecting the robustness and reliability of the predictive models.

### 3.3. Plant Assets: Transformers, Switchgears

The analysis for transformers primarily relies on defect notification data using defect rate as the outcome variable. The first approach involves forecasting the asset defect rate and asset population (when appropriate) to determine the number of defects per asset population each year.

The analysis of switchgear follows a similar modelling process to the transformers, focused on forecasting the asset defect rate and asset population (when appropriate) to determine the number of defects per asset population each year

### 3.3.1. Modelling Process: Asset Defect Rate

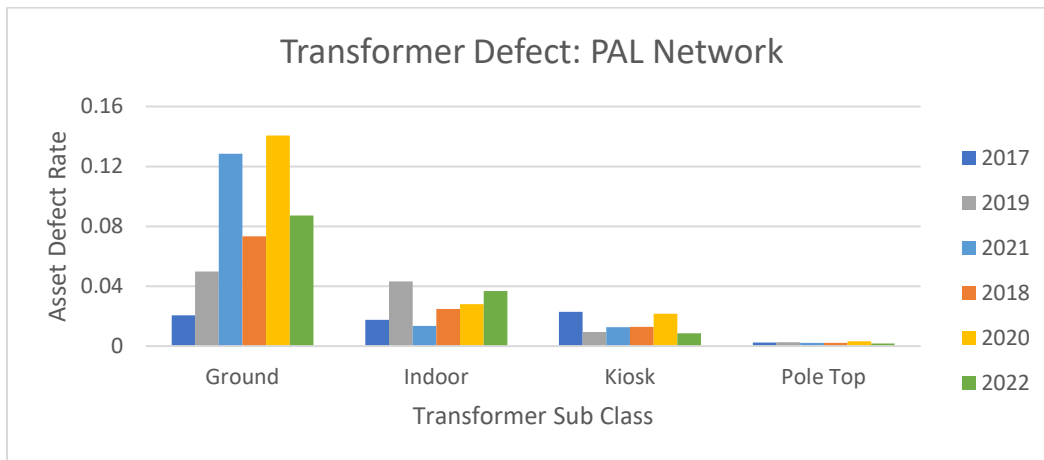
Transformer and Switchgear follow the same modeling process, differing only in how the dataset was separated for distinct analyses, as outlined in a later section.

To predict the asset defect rate for 2026 to 2031, it's crucial to assess the significance of time (year variable). Given the limited number of data points (range from 3 to 6) and the presence of only one explanatory variable (year), linear regression was chosen to model and understand the trend.

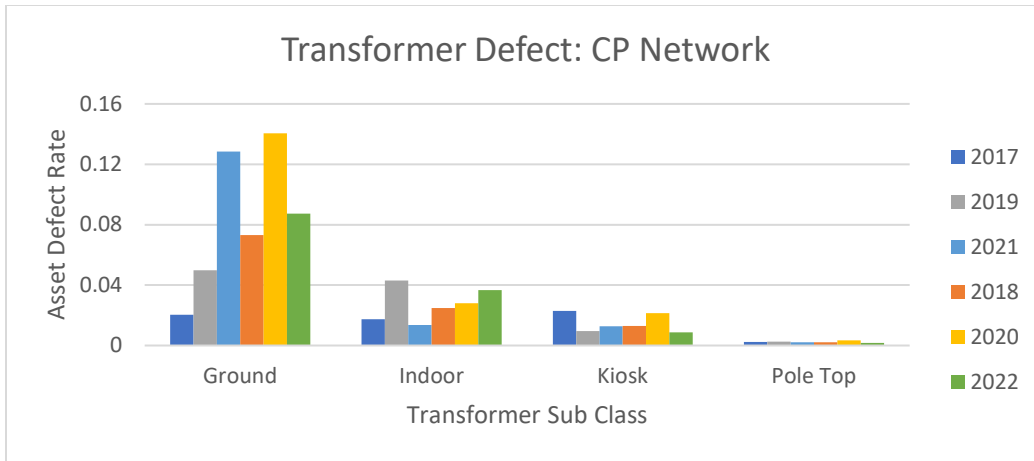
The efficacy of this model is evaluated based on the R-squared value it yields. A model is deemed satisfactory if the R-squared value exceeds 0.5 (refer to 2.6.2), indicating that over half of the variation is explicable by the model. In scenarios where the model doesn't meet this threshold, a different approach is adopted. Instead of relying on the regression model, we revert to using historical averages, proceeding to make predictions based on these averages. The final model performance was then evaluated using Root Mean Squared Error (RMSE).

#### *Transformer*

In analyzing the asset defect rate for transformers, the defect notifications dataset and the asset population dataset were used in combination to derive the asset defect rate per annum. The defect rate per subclass for each network is shown in below where it is apparent that the defect rate for each subclass behaves similarly where the ground type has the highest defect rate and pole top has the lowest defect rate comparatively to other subclasses for both networks regardless to the population count. The defect rate for each subclass is thus analysed independently, except for UE network where all defects were totaled.

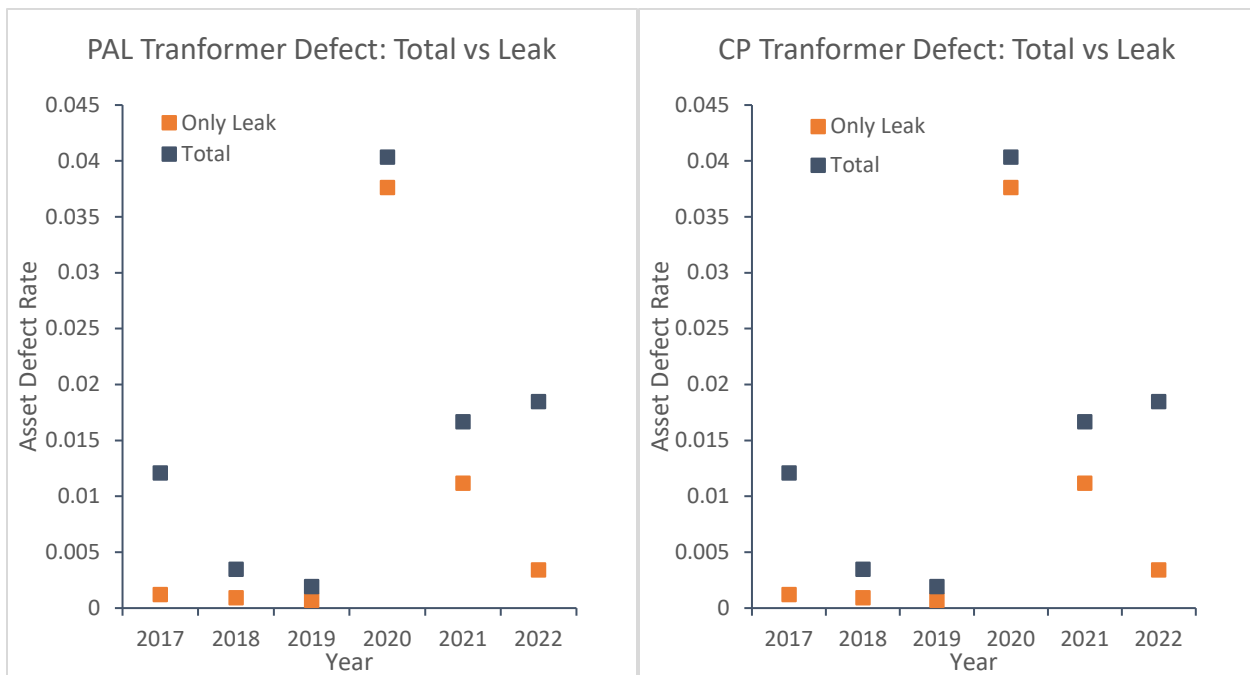






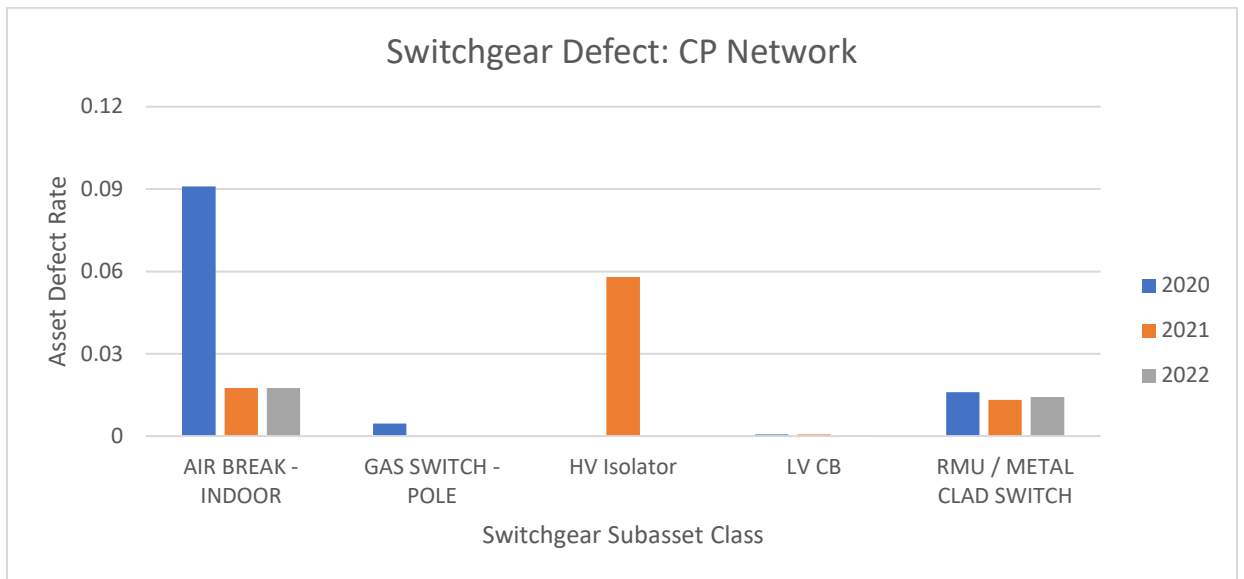
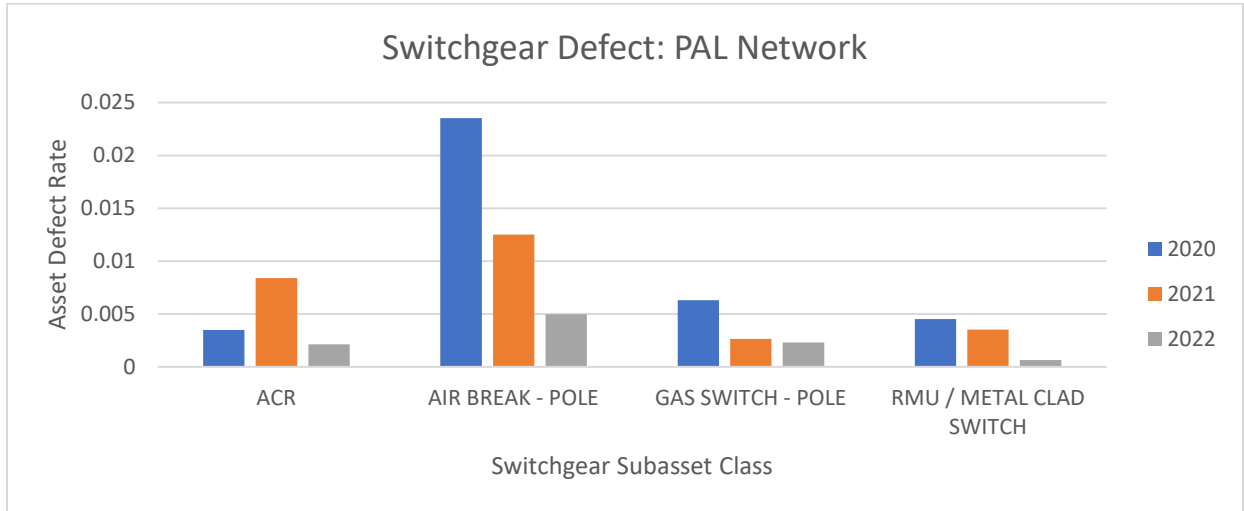
For subasset class that had no recorded defect notifications such as Ground Transformers in the CP network for 2018 and 2019. The defect rates for these subclasses were combined with other subclasses that also had no notifications (Ground + Pole Top) to obtain a combined defect rate.

Since oil leakage can have a significant implication for transformers, the defect notifications were further divided into two categories: defects caused by oil leaks and the overall total defect count. The charts below suggest the defect rate caused by oil leaks had a smaller variation when compared accounting for the total defect counts, except for 2020, when the defect rate was much higher for both networks. Thus, defects due to oil leaks is analysed independently.



### Switchgear

Similar to transformers, each subclass was analyzed independently. Due to business context only the historical defect rate from 2020 to 2022 was used to forecast the defect rate. Subasset classes with no recorded defect notification for at least one year were combined with other subclasses to obtain a combined defect rate.



### 3.3.2. Empirical Results: Asset Defect Rate

#### Transformer

The table below outlines the specific models utilized for each subclass for defect due to leakage and total defect, reflecting the modelling process to determine the most suitable model for predictive analyses. All subclasses of transformers in both networks indicate that a linear relationship is not present in the defect rate across 2017 to 2022. The yellow highlighted cells are subasset class with combined defect rate.

Sub Class	PAL - Total	CP - Total	PAL - Leak	CP - Leak	UE - Total
Indoor	AVG	AVG	AVG	AVG	AVG
Kiosk	AVG	AVG	AVG	AVG	AVG

Pole Top	AVG	AVG	AVG	AVG
Ground	AVG	AVG	AVG	AVG

The following tables show the asset defect rate estimated.

Sub Class	PAL - Total	CP - Total	PAL - Leak	CP - Leak	UE - Total
Indoor	0.0273	0.0095	0.0137	0.0044	0.0172
Kiosk	0.0147	0.0113	0.0076	0.0026	
Pole Top	0.0024	0.0023	0.0007	0.0026	
Ground	0.0833	0.0023	0.0345	0.0026	

Table 4: Transformer Defect Rate Estimated using Average Historical

The errors for these estimates derived in RMSE are as follows:

Sub Class	PAL - Total	CP - Total	PAL - Leak	CP - Leak	UE - Total
Indoor	0.01026	0.00742	0.00569	0.00343	0.00344
Kiosk	0.00555	0.00681	0.00483	0.0033	
Pole Top	0.00050	0.0018	0.00016	0.0033	
Ground	0.04186	0.0018	0.02047	0.0033	

Table 5: Transformer Defect Rate Error Measurement (RMSE)

### Switchgear

For switchgear, the asset defect rate for all subclasses does not appear to have a time trend in the historical data. Therefore, the average of the historical data was computed and will be used as the defect rate. The following table shows the estimated asset defect rate.

Sub Class	PAL	CP	UE
ACR	0.00466	0.004583	0.01623
Air Break- Indoor	0	0.041999	
Air Break- Pole	0.01367	0.004583	
Gas Switch- Pole	0.00375	0.004583	
HV Isolator	0	0.004583	
LV CB	0	0.000485	
RMU	0.00290	0.01445	

Table 6: Switchgear Defect Rate Estimated using Average Historical

The error measurements derived in RMSE for the predictions is as follows:

Sub Class	PAL	CP	UE
ACR	0.00269	0.002353	0.003255
AIR BREAK - INDOOR	0	0.034585	
Air Break- Pole	0.00762	0.002353	
GAS SWITCH - POLE	0.00181	0.002353	
HV Isolator	0	0.002353	
LV CB	0	0.000144	
RMU / METAL CLAD SWITCH	0.00164	0.00118	

Table 7: Switchgear Defect Rate Error Measurement (RMSE)

### 3.3.3. Modelling Process: Asset Population

Asset population data was obtained from in-service equipment in the RIN data. Specific data processing can be referred to in Data Cleaning section 2.3.2.

To predict the eventual defect volume for transformers and switchgears, future asset populations were forecasted using linear regression, chosen for its compatibility with the observed historical linear patterns. This relationship was then applied to project future asset populations, assuming ongoing consistent growth.

For transformers, this modelling process applies for all CP, PAL, and UE networks. For switchgears, some exceptions apply. The modeling process is only applicable to CP and PAL, as the UE network switchgear population is assumed to be constant. Another exception is for Air Break – Pole switchgears in the CP and PAL networks. These assets are being phased out by 2032 and gradually replaced with Gas Switch – Pole equipment. A constant decay factor is applied over time, and this decay is added to the gas switch – pole population in the forecasted data.

### 3.3.4. Empirical Results: Asset Population

#### Transformer

The tables below display forecasted populations for each network, with consistent growth evident in most asset subclasses, boasting an R-squared value exceeding 0.9. The exceptions include indoor transformers in PAL and ground transformers in CP, which remain constant.

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction	R-squared
Ground	361	347	333	319	304	290	0.973729
Indoor	524	524	524	524	524	524	0.000366
Kiosk	6300	6534	6769	7004	7238	7473	0.991374
Pole Top	83464	83989	84514	85039	85564	86090	0.937229

Table 8: PAL Transformer Asset Population Forecast

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction	R-squared
Ground	25	25	25	25	25	25	-
Indoor	3808	3836	3865	3893	3922	3950	0.971689
Kiosk	478	484	491	497	503	509	0.967218
Pole Top	852	857	861	865	870	874	0.896511

Table 9: CP Transformer Asset Population Forecast

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction	R-squared
UE Total	14558	14646	14734	14822	14910	14998	0.963826

Table 10: UE Transformer Asset Population Forecast

#### Switchgear

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction	R-squared
ACR	3720	4021	4322	4623	4924	5225	0.9437

Air Break-Indoor	41	41	41	41	41	41	-
Air Break-Pole	1450	1209	967	725	483	241	Linear Decay
Gas Switch - Pole	3514	3796	4077	4359	4640	4922	0.9884
HV Isolator	106	109	111	114	117	119	0.8773
LV CB	4505	4581	4657	4733	4809	4885	0.9918
RMU / METAL CLAD SWITCH	5700	5986	6272	6558	6843	7129	0.9888

Table 11: PAL Switchgear Asset Population Forecast

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction	R-squared
ACR	48	49	50	51	52	54	0.713208
AIR Break - Indoor	59	59	60	61	61	62	0.84
Air Break- Pole	9	7	6	4	3	1	Linear Decay
Gas Switch - Pole	377	404	431	459	486	513	0.980084
HV Isolator	80	82	85	88	90	93	0.889967
LV CB	3837	3913	3989	4065	4141	4216	0.988011
RMU / METAL CLAD SWITCH	2553	2639	2726	2812	2899	2986	0.984913

Table 12: CP Switchgear Asset Population Forecast

### 3.4. Plant Assets: Underground Asset

For the underground assets, where asset population and equipment features were unavailable, the only dataset accessible was the defect notifications. Given that most associated assets in this class are underground, defect notifications are triggered upon issues. Consequently, we employed time series analysis and forecasting to estimate defect volumes. Each associated asset class in each network was analyzed separately.

#### 3.4.1. Modelling Process: Defect Volume

The notification data was aggregated to monthly data spanning from January 2017 to December 2022, with a total of 72 months. In instances where no notifications were recorded, the notification count was considered as 0. The monthly aggregated data did not show clear trend and seasonality, thus to select the most reliable forecasting methodology, three models were tested: simple exponential smoothing (SES), linear regression, and historical average. (refer to 2.5.3 for model specifications)

To ensure a robust model evaluation, we performed a time-series variant of test-train split (refer to 2.5.1) and compared the results from the three models. The training data covered the period from January 2017 to September 2021. The trained model was subsequently applied to forecast data from October 2021 to December 2021. To evaluate the appropriateness of forecasting for 2026 to 2031, the Root Mean Square Error (RMSE) was computed using the predicted values and actual values from the testing set.

The priority was to identify the model with the smallest Root Mean Square Error (RMSE). As well as ensuring the yielded results that were reasonably close to those of the other models. Additionally, when SES algorithm,

the hyperparameter, smoothing factor ( $\alpha$ ), was selected by iteratively testing alpha from 0.2 to 1.0 to identify the lowest RMSE most suitable and balanced value for  $\alpha$  to apply. Finally, to forecast defect volumes for 2026 to 2031, each model was applied to the complete historical dataset, with the SES model being updated using the selected  $\alpha$ . The forecast output of time this analysis is in monthly data, which is subsequently aggregated into annual figures.

### 3.4.2. Empirical Results: Defect Volume

The table below shows the forecasted annual replacement volume for each of the associated classes in the underground asset group along with the RMSE values in parentheses. Additionally, the table includes the most suitable smoothing factor alpha for Simple Exponential Smoothing (SES), and the lowest RMSE values are highlighted in yellow.

Both the Simple Exponential Smoothing (SES) and average yield a point forecast which results in constant forecasts for each year. Linear regression also provides a stable forecast for the entire prediction period because of the stable defect notification volumes when distributed into months, leading to a relatively flat slope in the linear regression model. The yellow highlighted values are those with the lowest RMSE.

UG Associated Asset	Models	PAL	CP	UE
Pits	SES	122 (35.92) $\alpha = 0.20$	17 (6.77) $\alpha = 0.20$	12 (2.14) $\alpha = 0.20$
	LR	66 (21.21)	12 (5.47)	Negative (3.87)
	AVG	65 (27.53)	11 (5.80)	10 (2.50)
Pillars	SES	117 (24.03) $\alpha = 0.20$	-	12 (1.21) $\alpha = 0.45$
	LR	44 (15.85)	-	1 (1.19)
	AVG	44 (20.27)	-	1 (1.18)
Subtransmission Cable	SES	1 (1.54) $\alpha = 0.20$	5 (3.00) $\alpha = 0.20$	-
	LR	0 (1.48)	1 (2.87)	-
	AVG	0 (1.44)	1 (2.89)	-
HV Cable	SES	14 (3.63) $\alpha = 0.20$	49 (11.57) $\alpha = 0.20$	12 (2.96) $\alpha = 0.45$
	LR	8 (2.38)	11 (9.32)	13 (3.42)
	AVG	8 (2.66)	11 (10.85)	13 (3.92)
LV Cable	SES	7 (2.05) $\alpha = 0.20$	4 (2.04) $\alpha = 0.20$	41 (5.35) $\alpha = 0.20$
	LR	3 (1.68)	10 (6.63)	23 (5.60)
	AVG	3 (1.79)	10 (2.93)	23 (8.87)

This shows the best model with the lowest RMSE. Linear regression yields a slightly smaller error than that of an average value in most cases suggesting there accounting for the time-dependent variability, the model performs only slightly better than average. For instances where simple exponential smoothing is chosen as the best performing algorithm, the alpha value is low of 0.20 – 0.45 indicates less emphasis on recent observations.

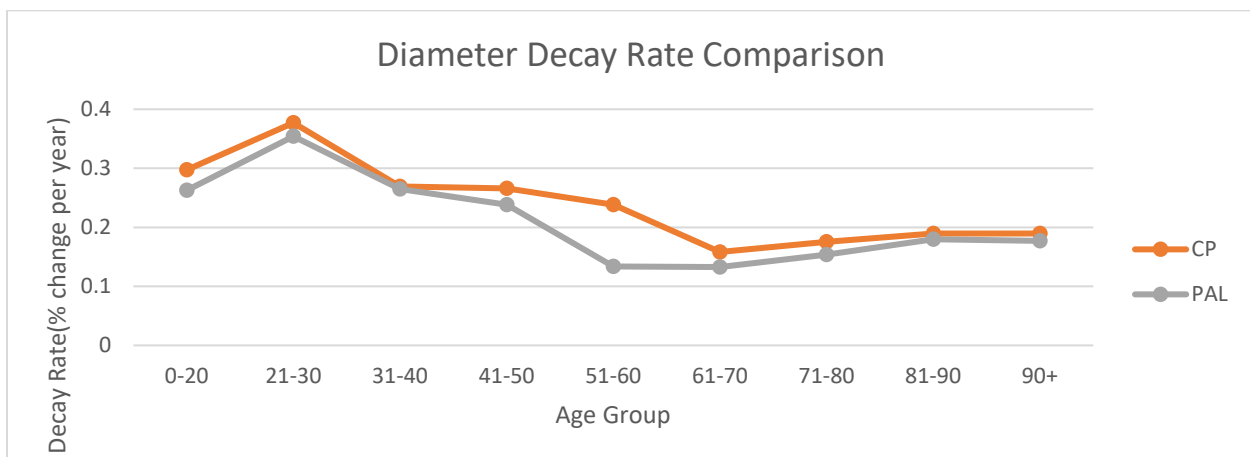
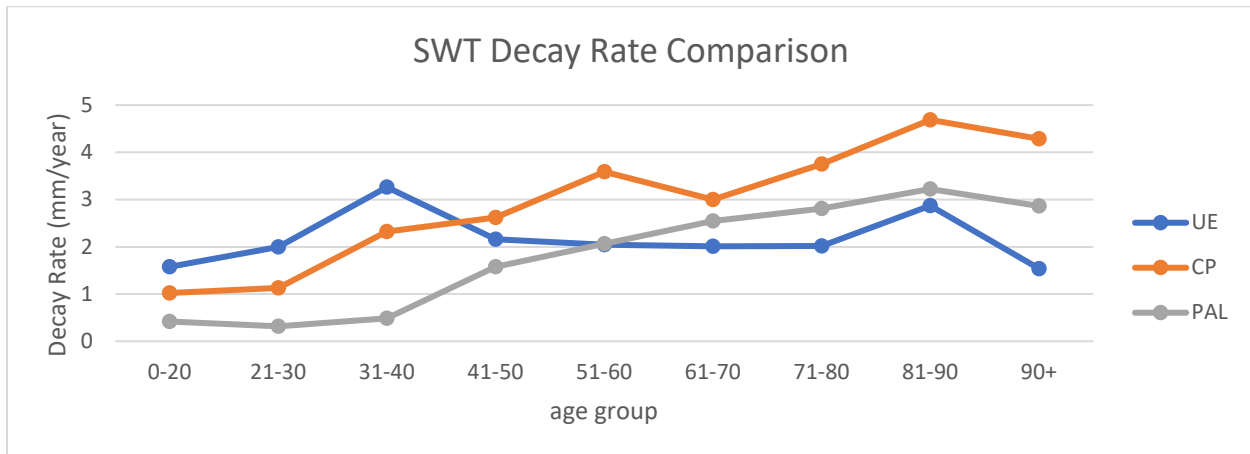
UG Associated Asset	PAL	CP	UE
Pits	LR	LR	SES (0.45)

<b>Pillars</b>	LR	-	AVG
<b>Subtrans Cable</b>	AVG	LR	-
<b>HV Cable</b>	LR	LR	SES (0.45)
<b>LV Cable</b>	LR	SES (0.20)	SES (0.20)

## 4. Key Findings and Analysis

### 4.1. Line Assets: Decay Rate Summary

The following are the visualizations for comparing the wood pole measurement decay rates:

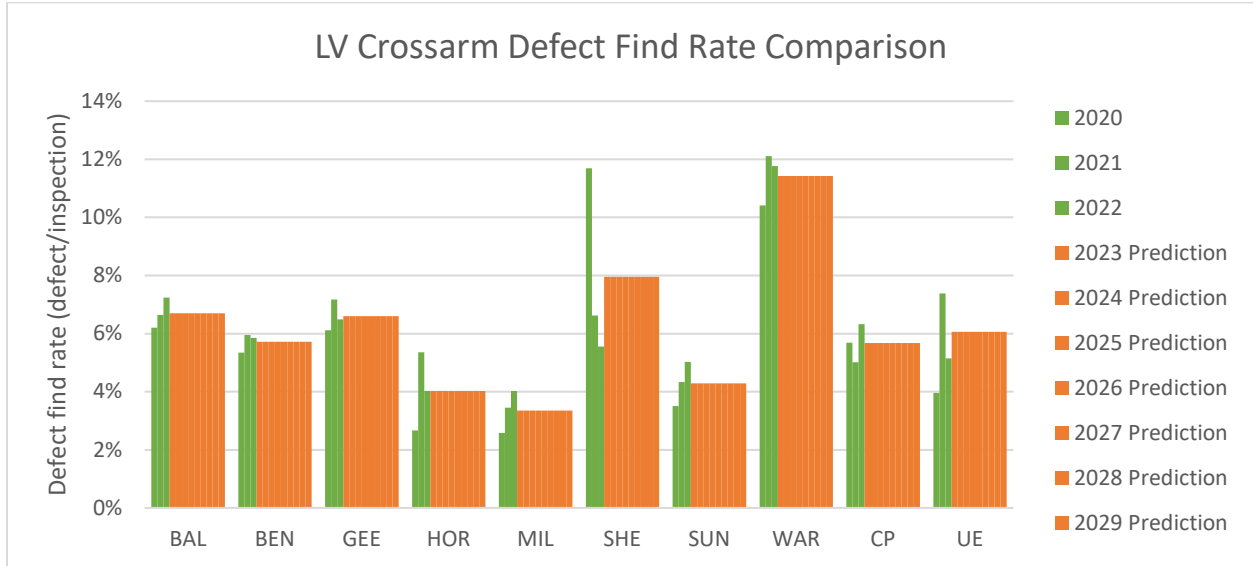
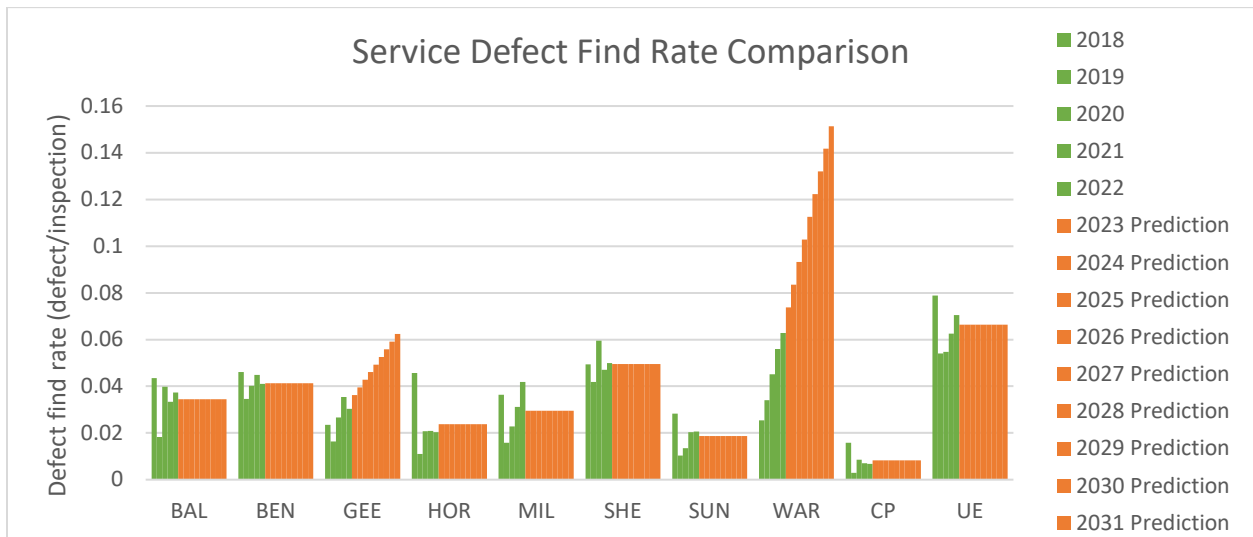


### 4.2. Line Assets: Find Rate Comparisons

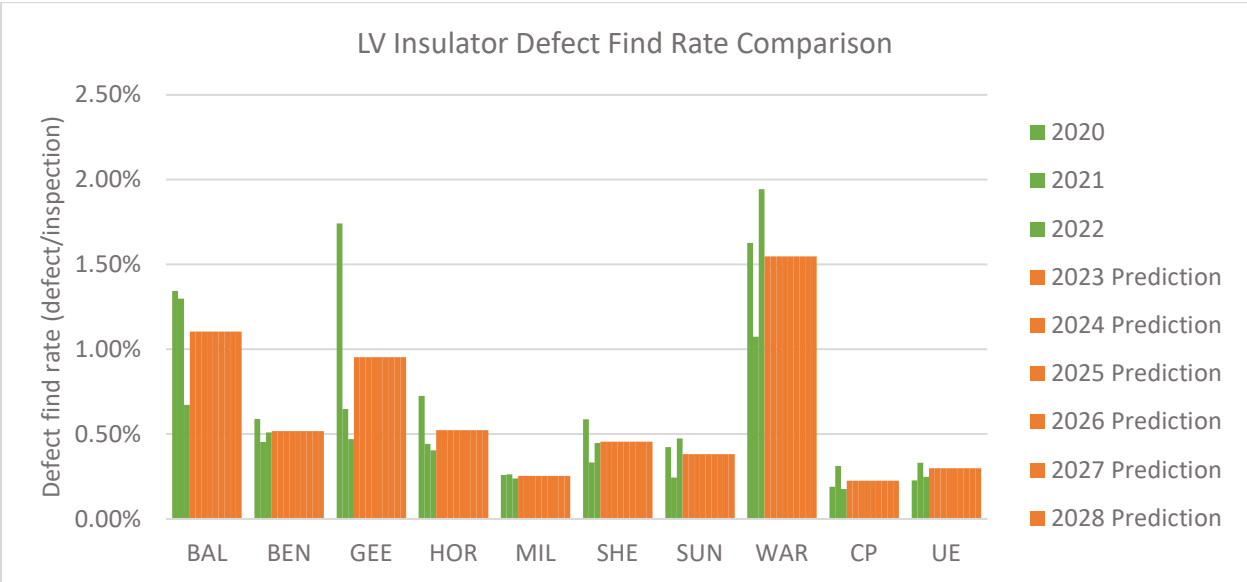
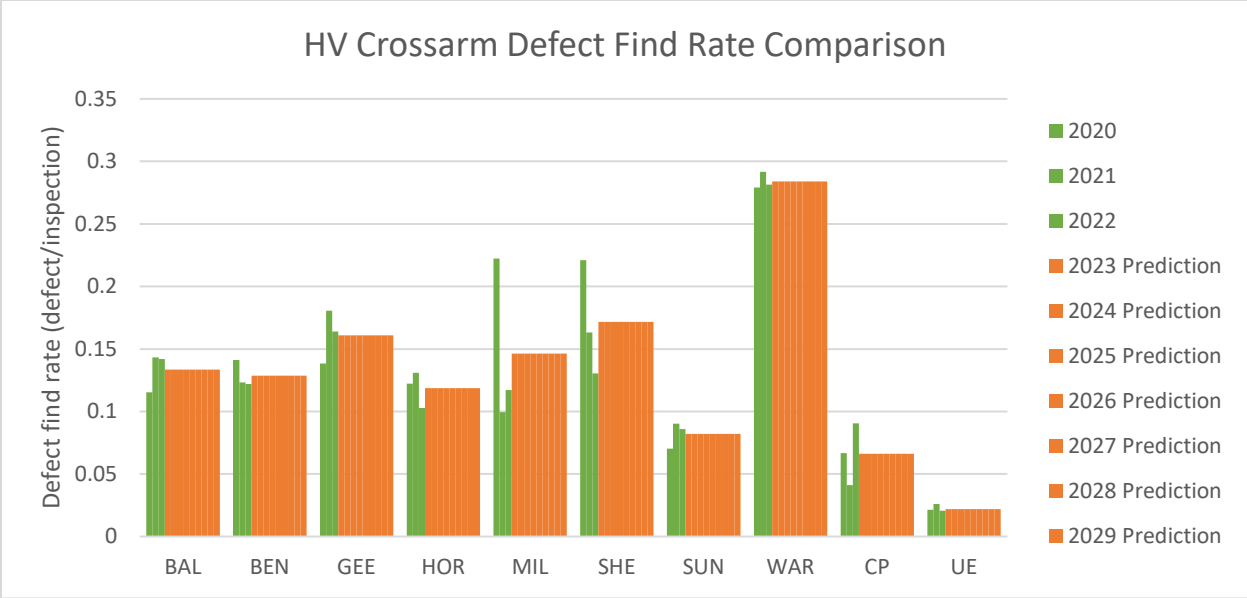
Visualizations of the defect find rate across different locations and owner groups for each asset category have revealed notable disparities and trends. A significant observation is the pronounced variation in defect find rates between different asset owners and locations, both in magnitude and directional trend.

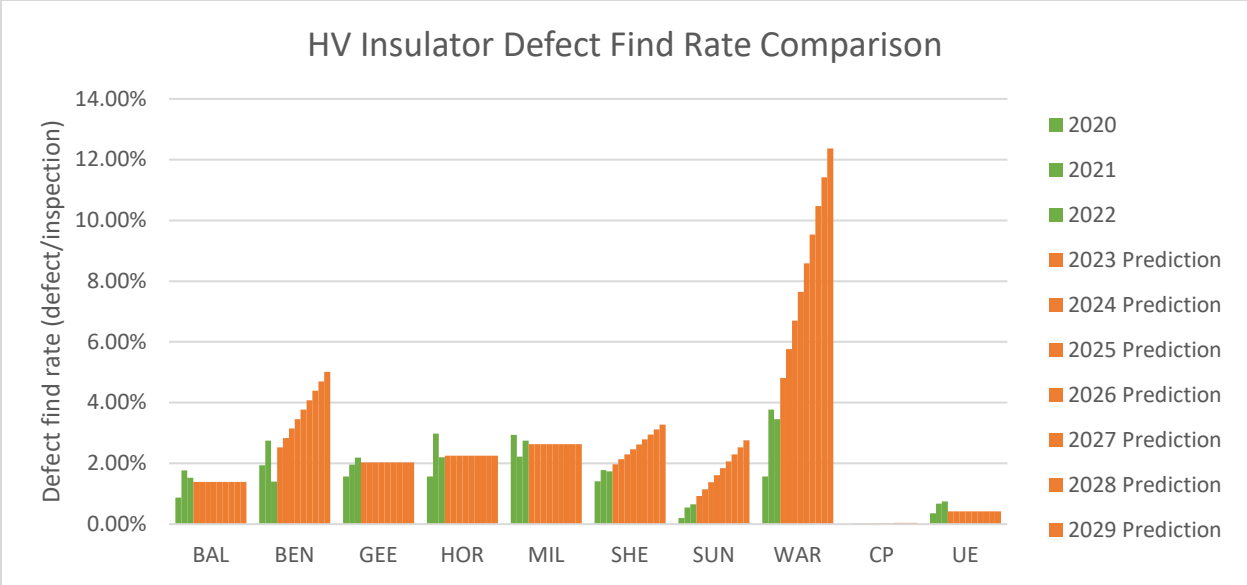
For instance, considering the LV crossarm, the Shepparton ('SHE') location tagged 'SHE' exhibits a discernible downward trend in defect rates, contrasting with other locations that either remain trendless or display an upward trajectory. Additionally, the Warnambool ('WAR') consistently registers the highest defect find rates across most assets.

These observations underscore the necessity of conducting nuanced analyses and forecasts at a granular, location-specific level. Relying on a consolidated average across all locations can introduce biases, especially considering the diverse historical patterns and asset sizes seen in different locations. Tailoring the analysis to reflect location-specific trends and magnitudes ensures a more accurate and representative understanding of the defect find rates, thereby enhancing the precision and relevance of the resulting forecasts. Refer to Appendix 1 for location codes and their corresponding regions.









### 4.3. Plant Assets: Asset Defect Rate

#### 4.3.1. Transformer Replacement Volume

Combining the prediction values of defect rate and asset population, the replacement volume is derived. The lower bound number indicates the defect volume due to oil leaks and the upper bound includes the entire defect notifications. The following shows the replacement volume within this range for PAL and CP network.

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction
Ground	(12, 31)	(12, 30)	(12, 29)	(11, 28)	(11, 26)	(10, 25)
Indoor	(7, 14)	(7, 14)	(7, 14)	(7, 14)	(7, 14)	(7, 14)
Kiosk	(48, 93)	(49, 96)	(51, 99)	(53, 103)	(55, 106)	(56, 110)
Pole Top	(61, 198)	(61, 199)	(62, 200)	(62, 202)	(62, 203)	(63, 204)

Table 13: PAL Transformer Replacement Volume

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction
Ground	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
Indoor	(17, 36)	(17, 36)	(17, 37)	(17, 37)	(17, 37)	(17, 37)
Kiosk	(1, 5)	(1, 5)	(1, 6)	(1, 6)	(1, 6)	(1, 6)
Pole Top	(2, 2)	(2, 2)	(2, 2)	(2, 2)	(2, 2)	(2, 2)

Table 14: CP Transformer Replacement Volume

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction
UE Volume	241	241	241	241	241	241

Table 15: UE Transformer Replacement Volume

#### 4.3.2. Switchgear Replacement Volume

Combining the prediction values of defect rate and asset population, the replacement volume is derived. The asset class where defect rate is 0 will be left out in the result output.

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction
ACR	13	15	16	17	19	20
Air Break- Pole	30	26	23	20	17	13
Gas Switch- Pole	10	11	12	13	14	15
RMU	14	15	16	17	17	18

Table 16: PAL Switchgear Replacement Volume

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction
AIR BREAK - INDOOR	2	2	2	2	2	3
GAS SWITCH - POLE	1	1	2	2	2	2
LV CB	2	2	2	2	2	2
RMU / METAL CLAD SWITCH	33	34	36	37	38	39

Table 17: CP Switchgear Replacement Volume

Sub Class	2026 Prediction	2027 Prediction	2028 Prediction	2029 Prediction	2030 Prediction	2031 Prediction
<b>UE Volume</b>	241	241	241	241	241	241

Table 18: UE Switchgear Replacement Volume

#### 4.4. Plant Assets: Defect Volume

The table below shows the forecasted annual replacement volume for each of the associated classes in the underground asset group.

UG Associated Asset	PAL	CP	UE
<b>Pits</b>	66 (21.21)	12 (5.46)	12 (2.14)
<b>Pillars</b>	44 (15.85)	-	1 (1.18)
<b>Subtrans Cable</b>	0	1 (2.87)	-
<b>HV Cable</b>	8 (2.38)	11 (9.33)	12 (2.96)
<b>LV Cable</b>	3 (1.67)	4 (2.04)	41 (5.35)

Table 19: UG Assets Forecasted Replacement volume with Error Measurement

## 5. Recommendations

Accurately forecasting the replacement volume for the asset is imperative for the network to ensure the seamless delivery of energy to consumers while maintaining safety. Overestimating this volume results in resource wastage and underutilization of the asset's operational lifespan. Conversely, underestimating the volume leads to insufficient resources for adequate asset maintenance, potentially falling short of expectations and requirements.

## Plant Distribution

For the plant distribution side of the business, we employ two key variables to determine replacement volume: asset defect rate and volumetric forecasting. Estimating replacement volume through asset defect rate can cause some limitations. This approach necessitates forecasting multiple variables, specifically the defect rate and asset population. The challenge here lies in the joint error associated with these variables, which affects the accuracy of the final value.

## Transformers and Switchgears

Transformers' defects can be influenced by several factors, such as equipment specifications, environmental conditions, and loading conditions. Each of these variables can significantly impact the operational life of these devices. Unfortunately, the absence of these crucial data elements limits the predictive capabilities. Similarly, switchgear exhibits distinctive characteristics that might influence operational life, such as distance to the zone substation or proximity to the coast. However, as defect notifications are predominantly raised against the nearest pole, connecting the timing and cause of these notifications to the specific asset posed as a limitation for both asset types.

To attain a more comprehensive view of asset replacement patterns, the availability and combination of historical failure data, structural data, and environmental data were used. Historical failure data encompasses records of previous failures or inspections. Structural data includes detailed equipment specifications, and environmental data encompasses information about environmental conditions and loading. When combined, this holistic dataset equips stakeholders with the insights needed to mitigate current risks by prioritizing the replacement of the riskiest assets. This holistic approach enhances the understanding of the operational assets and their expected service life over the next 5 to 10 years. Additionally, it reduces corrective maintenance costs, improves reliability, and supports well-informed investment decisions aimed at minimizing customer interruptions.

## Underground Cables

For underground cables, their replacement cost depends on the cable's length and location, not the occurrence of defect notifications. However, the analysis present primarily relies on the notifications, which can complicate cost predictions. To improve the forecasting capability, features such as the cable type, installation year, and the cable's length each time a replacement is made. This information provides insights into the cable's condition at various locations and the proportion of cables posing a risk. Additionally, it's operationally convenient to replace entire cable sections in one go. Therefore, clustering cables can further enable a more realistic predictions about the lengths of replacement needed, leading to a better estimate of the replacement costs for the next 5 to 10 years.

## Line Assets Comparison

For Line assets, the availability of more comprehensive and structured data enables the exploration and application of a broader spectrum of forecasting approaches, enhancing predictive accuracy. The constraints faced in the current approaches predominantly stem from limitations in data availability and quality.

## Compliance with AER Requirements & Recommendations for Improvement

The replacement volume is categorized into specific asset groupings. However, the raw data available doesn't align directly with these categories. As a result, we had to make multiple imputations, particularly for the switchgear and underground asset groupings (accounting for 40% of the imputations), which served as the basis for our replacement volume analysis. To enhance this process, we recommend including these specific information fields as a requirement during data collection and processing, ensuring a more streamlined and accurate analysis.

In conclusion, refining the forecasting methodologies through enhanced data utilization and methodological adjustments will bolster the accuracy of asset replacement volume predictions, supporting informed decision-making and optimize operational strategies.

## 6. References

Galit Shmueli and Lichtendahl, K.C. (2018). *Practical Time Series Forecasting with R*. Axelrod Schnall Publishers.

## 7. Appendices

### Appendix 1:

<b>Location Code</b>	<b>Location</b>
BAL	BALLARAT
BEN	BENDIGO
GEE	GEELONG
HOR	HORSHAM
MIL	MILDURA
SHE	SHEPPARTON
SUN	SUNSHINE
WAR	WARNAMBOOL
BAL	BALLARAT

### Appendix 2

Data Cleaning Rules for Line Asset:

Wood Pole Decay Rate Model:

- Construction Year between 1800 and 2022
- Decay in measurements needs to be positive
- Time difference between inspections for the same equipment needs to be at least one year
- Measurement date needs to be later than construction year
- Records with more than 10 years in time duration but no measurement changes are removed
- Sound wood thickness needs to be less than 200
- Unknown wood pole class is replaced by Class 3
- If there is inconsistency between measurements of the same equipment (i.e., when the measurement increases in one of its inspections), this record is removed