Developing a National Insurance Program to Mitigate Dam Failure Losses and Improve Relief Outcomes for Stakeholders in Tarrodan

[redacted]

ABSTRACT

Dams have been crucial for centuries, serving as vital infrastructure for managing water resources across various sectors. As populations grow, the demand for dams increases, but the potential consequences of dam failures are often overlooked. Many dams are not properly maintained, leading to significant risks to society. In the simulated country of Tarrodan, our case study aims to develop a comprehensive national insurance program that accounts for the financial losses of key stakeholders. Using statistical modeling techniques, including hypothesis testing, decision tree regression, the random forest, the frequency and severity model, and time series analysis, we assess key factors contributing to dam failures and determine the optimal government inspection strategies to reduce risks. We propose an insurance framework that strengthens financial resilience while also providing insights into risk management policies and government supervision strategies to prevent large-scale dam failures in Tarrodan.

INTRODUCTION

Dam failures have historically led to severe financial losses and environmental devastation, yet many nations, particularly developing economies, lack adequate risk management strategies. Without proper maintenance, inspections, and financial safety nets, dam failures can result in widespread economic and human losses. Despite the importance of dam infrastructure, existing insurance frameworks often fail to account for large-scale liabilities, leaving governments and stakeholders financially vulnerable.

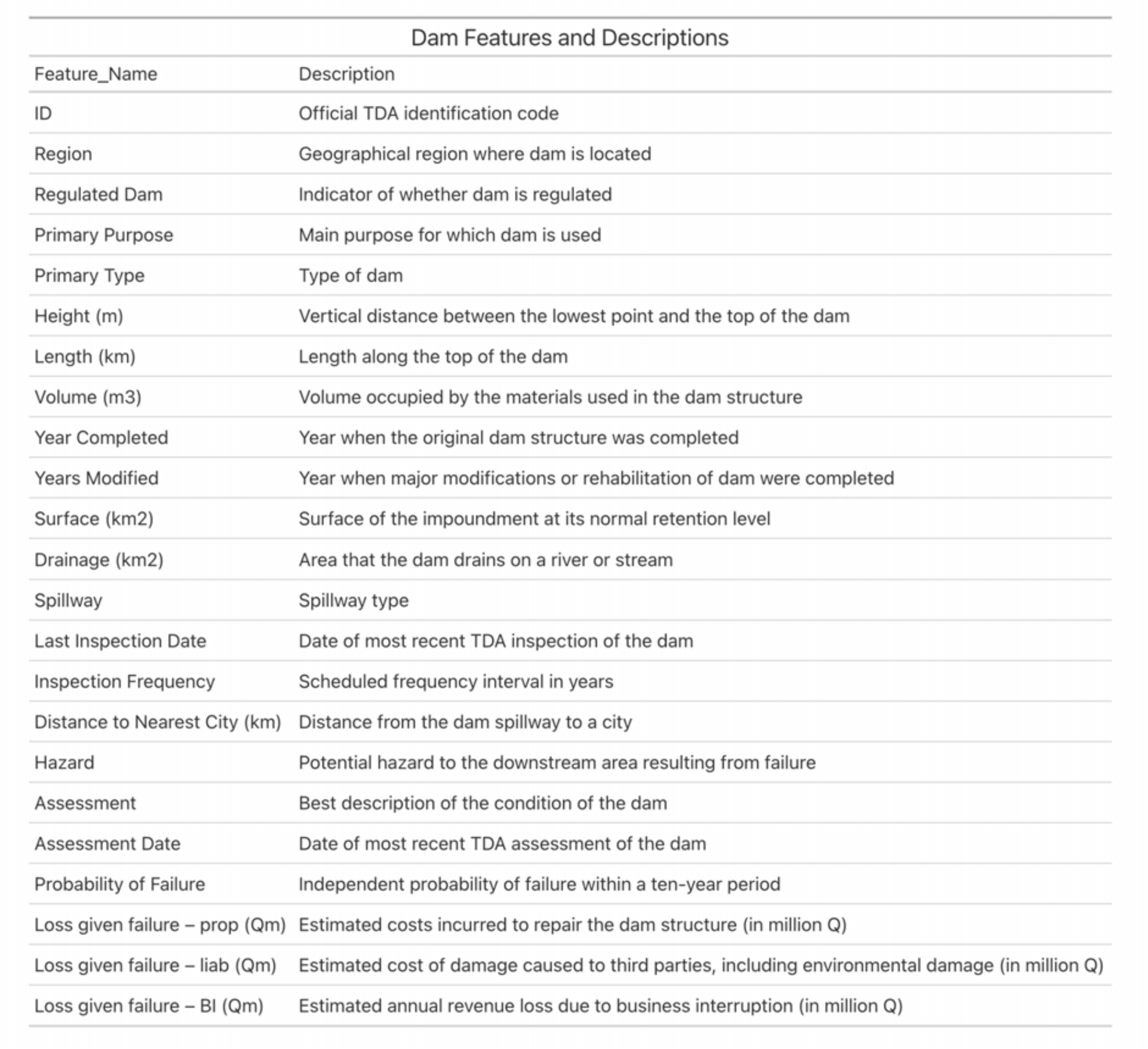
Tarrodan, a simulated nation spanning 2 million square kilometers with a population of 95 million, is heavily dependent on its 20,806 earthen dams for irrigation, hydroelectric power, and trade. The country consists of three main regions—Flumevale, Lyndrassia, and Navaldia—each with distinct geographical and economic characteristics. Flumevale, the agricultural center, relies on rivers for irrigation and transportation. Lyndrassia, with its mountainous terrain, utilizes dam systems for hydroelectric power. Meanwhile, Navaldia's coastal position makes it an economic hub for maritime trade.

Despite the crucial role of dams in Tarrodan's economy, inadequate maintenance and inspection protocols increase the risk of failure. Currently, the country lacks a comprehensive insurance program to mitigate financial damages from dam-related disasters. Stakeholders—including local communities, businesses, and the government—face potential losses in property, revenue, and infrastructure without any structured relief mechanism. To mitigate the risk, the insurance program seeks to cover expected costs to help dam owners to relieve from catastrophic losses.

DATA OVERVIEW

The data used in this analysis is sourced from the Society of Actuaries' official website. Two main datasets are used for the study: the Dam dataset and the historical inflation dataset. The Dam dataset provides detailed information on 20,806 individual dams across three different regions in Tarrodan—Flumevale, Lyndrasia, and Navaldia. *Table 1* summarizes the description of each explanatory variable in the Dam dataset.

To ensure statistical robustness, we conducted a power analysis using Cohen's *f* test with an estimated effect size of 0.11. Given the large sample size and an average of approximately 6,935 observations per region, the power was calculated to be greater than 0.9. This high level of statistical power reduces the risk of Type II errors and supports the reliability of our inferential findings.



Dam Features and Descriptions

Feature_Name	Description
ID	Official TDA identification code
Region	Geographical region where dam is located
Regulated Dam	Indicator of whether dam is regulated
Primary Purpose	Main purpose for which dam is used
Primary Type	Type of dam
Height (m)	Vertical distance between the lowest point and the top of the dam
Length (km)	Length along the top of the dam
Volume (m3)	Volume occupied by the materials used in the dam structure
Year Completed	Year when the original dam structure was completed
Years Modified	Year when major modifications or rehabilitation of dam were completed
Surface (km2)	Surface of the impoundment at its normal retention level
Drainage (km2)	Area that the dam drains on a river or stream
Spillway	Spillway type
Last Inspection Date	Date of most recent TDA inspection of the dam
Inspection Frequency	Scheduled frequency interval in years
Distance to Nearest City (km)	Distance from the dam spillway to a city
Hazard	Potential hazard to the downstream area resulting from failure
Assessment	Best description of the condition of the dam
Assessment Date	Date of most recent TDA assessment of the dam
Probability of Failure	Independent probability of failure within a ten-year period
Loss given failure – prop (Qm)	Estimated costs incurred to repair the dam structure (in million Q)
Loss given failure – liab (Qm)	Estimated cost of damage caused to third parties, including environmental damage (in million Q)
Loss given failure – BI (Qm)	Estimated annual revenue loss due to business interruption (in million Q)

Table 1: Explanatory Variable Descriptions in the Dam Dataset

Each dam in the dataset is uniquely identified and described by 23 variables, including physical characteristics, functional purposes, and risk indicators. Regionally, Navaldia, being a coastal area, has the most dams (8,878), followed by the mountainous region of Lyndrasia with 8,406, and the agricultural region of Flumevale with 3,522. The purpose of dams varies by region: most dams in Flumevale are used for irrigation and agricultural water supply, those in Navaldia serve mainly for flood control and recreational purposes, while dams in

Lyndrasia are primarily used as fish and wildlife ponds. About 94% of the dams are classified as Earth Type Dams, with the remainder including concrete, gravity, and stone types. Size-related attributes such as height, length, volume, surface area, drainage area, and spillway type are included, giving a comprehensive picture of each dam's structure and coverage.

In addition to structural details, the dataset contains historical information about dam completion, modification, inspection, and assessment, spanning from 1948 to 2023. This information can be used to explore how government supervision—such as inspection frequency and attention to older dams—relates to dam performance and risk. However, these variables suffer from inconsistencies and missing values, which limits the depth of analysis possible in this area.

Our main actuarial interest lies in evaluating the risk of dam failure and its financial impact. The dataset provides the 10-year probability of failure for each dam, along with estimated losses in three categories: repair and maintenance, third-party damage, and business interruption. For the purpose of premium pricing, we aggregate these three losses into a total estimated cost per failure event. Since we are interested in calculating annual insurance premiums, we convert the 10-year failure probabilities into annual failure probabilities using the assumption of independence and constant failure rate. The conversion formula is:

$$q_x = 1 - (1 - {}_{10}q_x)^{\frac{1}{10}}$$

which q_x represents the annual probability of failure and ${}_{10}q_x$ represents the 10-year period probability of failure. These annual probabilities, along with the total loss amounts, form the basis for assessing risk levels, pricing premiums, and calculating the net present value (NPV) of insurance benefits for dam failure events.

The dataset also includes government ratings and hazard assessments for each dam. These are categorized as:

- Hazard levels: Low, Significant, and High
- Assessment ratings: Poor to Satisfactory

These ratings, together with the physical and historical variables, will be used in the underwriting process to help determine eligibility for inclusion in a national insurance program. To support this, classification models such as random forests and decision trees will be employed. Average failure probabilities by region are summarized in *Table 2*, which will be referenced in our discussion on risk reduction strategies.

Region	Measures	Value
Flumevale	Average Annual Probability of Failure	0.009159
Navaldia	Average Annual Probability of Failure	0.010005
Lyndrassia	Average Annual Probability of Failure	0.009976

Table 2: Average Annual Probability of Failure for each Region

Finally, the historical inflation dataset contains annual inflation rates from 1962 to 2024. This data is essential for adjusting projected costs and insurance premiums, allowing us to reflect long-term economic trends and improve the financial accuracy of our models.

OBJECTIVES

This study aims to develop a national insurance framework that protects all stakeholders in Tarrodan—including the government, dam owners, and residents—from financial losses due to dam failures. Our goals focus on three main areas: risk prevention, risk mitigation and minimization, and financial feasibility. To achieve this, we address the following research questions:

1. What common characteristics of dams significantly impact the probability of failure?
2. What is the optimal national insurance policy to cover the total annual expected loss of dam failures in each region?
3. How effective is government supervision in reducing the annual probability of dam failure, and are there specific regions that require more attention or oversight?

To explore these questions, we analyze key perspectives from the dataset as in the following:

- Key dam characteristics (e.g., construction year, spillway type, hazard classification).
- Probability Distribution of Failure and Expected financial losses over one-year period in terms of third-party liability, maintenance, and business interruptions.
- The optimal frequency of government inspections to minimize failure risks.
- Socioeconomic factors (inflation) to assess financial feasibility.

In terms of risk prevention, we focus on identifying the dam characteristics that most strongly influence the likelihood of failure—such as older construction, location, or hazard level. Based on these findings, we propose guidelines and regulations that require dam owners to perform regular maintenance or request inspections if their dams fall within high-risk categories. Failure to comply with these standards may result in penalties. Furthermore, we examine how inspection frequency can be tailored to dam characteristics for a cost-effective approach. By keeping dam structures in good condition through timely inspections, we aim to reduce the probability of failure without relying heavily on financial compensation for losses.

For risk mitigation and minimization, we introduce an annual insurance model for eligible dam owners (detailed in Results Section). Using an underwriting process, we propose a pricing structure—either uniform or region-specific—based on the distribution of annual failure probabilities and total expected losses. We will fit a parametric model to estimate short-term premium payments and ensure they align with actual risk levels. Additionally, we consider long-term financial stability by analyzing historical inflation data and interest rate trends. This helps ensure that the insurance program remains sustainable and capable of covering increased costs due to economic fluctuations.

METHODS

Our primary goal is to identify the key factors that significantly impact the probability of dam failure within the next 10 years. From our initial analysis, we observed a positive relationship between hazard level and failure probability (*see Appendix A*). To dig deeper into this relationship, we applied a random forest regression model to determine which variables have the strongest influence on failure probability. Variable importance was measured using IncNodePurity (Mean Decrease in Gini). In building the random forest, we removed missing data and grew 500 trees to ensure accuracy and consistency in the results.

After identifying important predictors, we used a decision tree model to visualize how these variables contribute to failure risk. Unlike the random forest, this model can handle missing values directly. We also applied cross-validation to avoid overfitting. The data was split into 70% for training and 30% for testing. Both pruned and unpruned trees were compared to determine the model with the best predictive performance.

Besides identifying key variables, we also analyzed the probability distribution of dam failure across regions and overall. With a large sample size in each region, we relied on the Central Limit Theorem and assumed a normal distribution for failure probabilities. To support this assumption, we ran diagnostic checks for normality. Using this underlying distribution, we performed hypothesis testing to evaluate whether the probability of failure differs across the three regions: Flumevale, Lyndrassia, and Navaldia. Specifically, we conducted ANOVA tests to assess differences in both the mean and variance across regions. This analysis helps us determine whether regional factors should influence premium pricing.

For the financial component, we turned to time series analysis to forecast key economic indicators like inflation. Since no seasonal trend was observed, we excluded seasonal components from our model. Instead, we built ARIMA models and selected the best one based on information criteria (like AIC/BIC). These forecasts allow us to estimate future cash flows and calculate the necessary reserve to cover potential dam failure losses.

Finally, to estimate the total annual expected loss from dam failures, we developed separate frequency and severity models. After identifying suitable distributions for both, we conducted a Monte Carlo Simulation to model a wide range of possible outcomes. This simulation also allowed us to estimate worst-case scenarios at the 95th percentile confidence level, giving us a more robust view of financial risk.

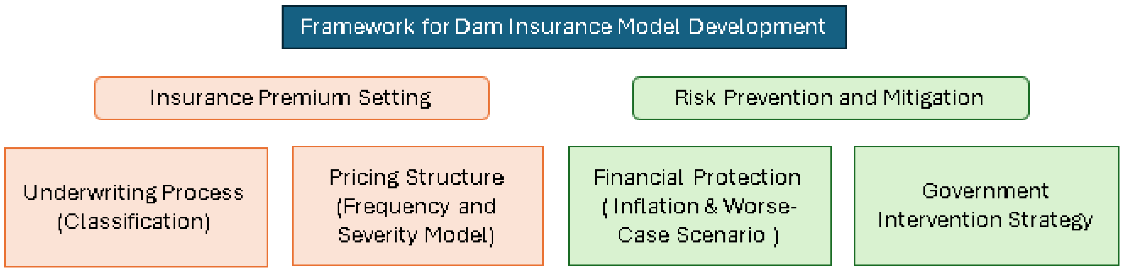

Framework for Dam Insurance Model Development
Insurance Premium Setting
Risk Prevention and Mitigation
Underwriting Process (Classification)
Pricing Structure (Frequency and Severity Model)
Financial Protection (Inflation & Worse-Case Scenario)
Government Intervention Strategy

Table 3: Overview of Methodological Approach

RESULTS

Decision Tree & Random Forest Regression

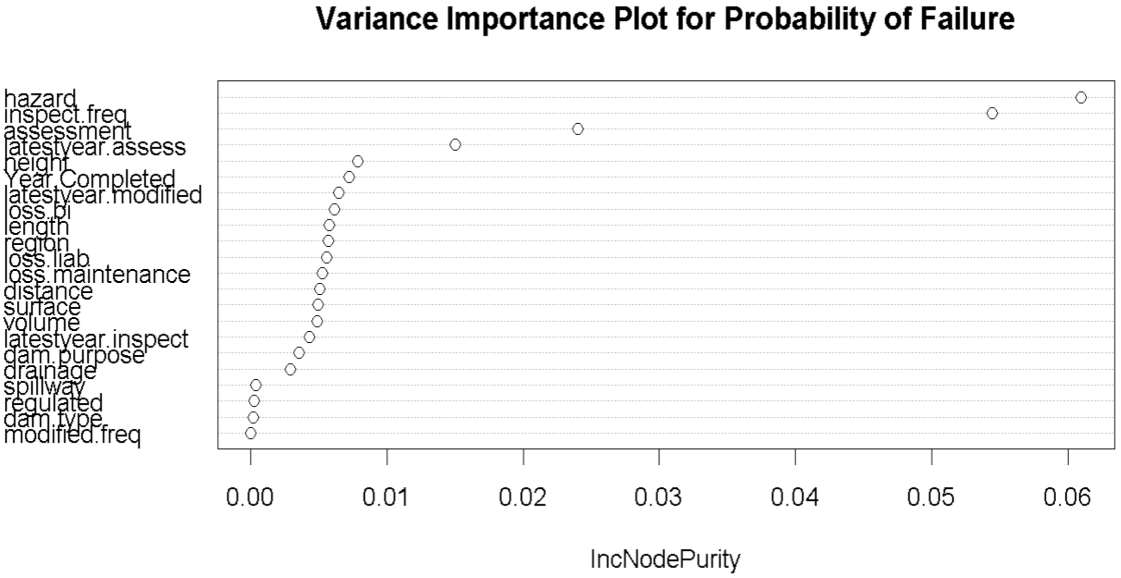


Figure 1: Random Forest Variance Importance Plot on Probability of Failure

The Random Forest Regression in *Figure 1* analysis indicates that hazard level, inspection frequency, and assessment rating are the three most significant factors influencing the annual probability of failure. To further support these findings, we examine the results from the decision tree model. We ran two decision tree models: one with pruning and one without pruning. As shown in the *(Appendix B)*, the pruned tree resulted in a higher Residual Mean Deviance and a higher test error rate. Therefore, we opted to use the unpruned tree model for our analysis.

The decision tree in *Figure 2* confirms that the results align with expectations. Specifically:

- Dams with low hazard levels and an assessment of "Satisfactory" have the lowest probability of failure among all classifications.
- Dams classified as significant hazard with an assessment of "Not Satisfactory" exhibit the highest probability of failure.
- High-hazard dams that are either "Not Rated" or assessed as "Unsatisfactory" have the second highest probability of failure.
- Dams with lower inspection frequency generally have a higher probability of failure than those inspected more frequently.

Overall, the decision tree results align with our initial classification assumptions, reinforcing the validity of our model's findings. We will incorporate this classification into our program design and pricing structure to ensure accurate risk assessment and financial planning.

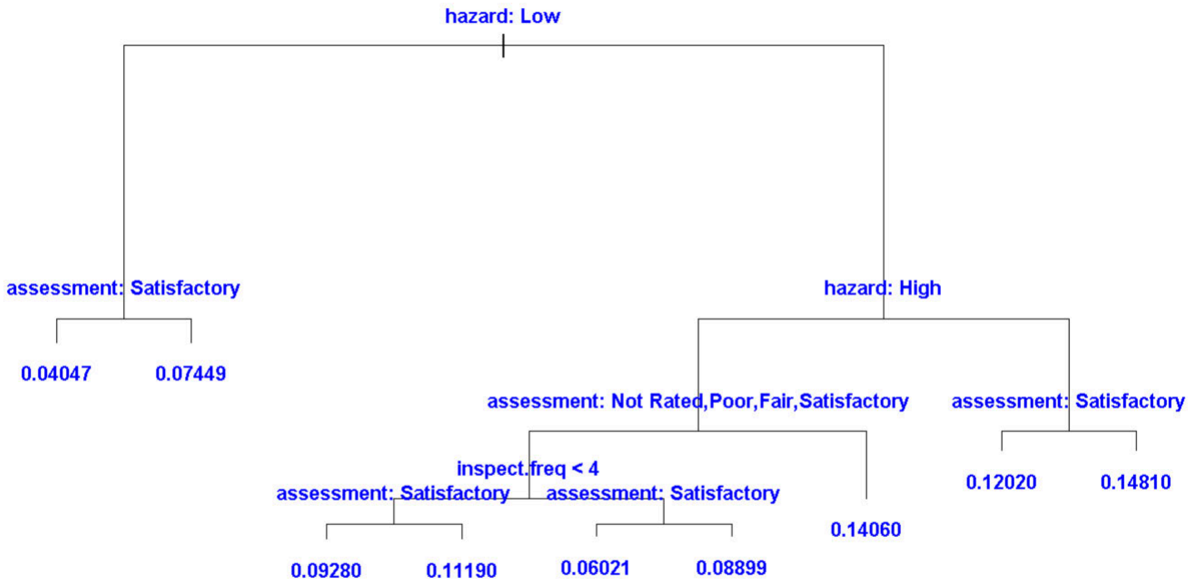


Figure 2: Decision Tree Plot (Unpruned)

The goal is to include all dams in the program while minimizing risk as much as possible. Although the hazard level of a dam cannot be changed, the assessment rating can be improved—therefore, dams with higher hazard levels will be required to meet stronger assessment standards. To meet underwriting criteria, we filtered the data based on specific requirements—for instance, only including dams with a Low Hazard with an Assessment of Not Rated, Fair, or Satisfactory, High hazard level with an Assessment rating of Fair or Satisfactory, and Significant hazard level with an Assessment rating of Satisfactory. This criteria decreases the probability of failure as shown in *Table 3*. According to *Table 3*, both Flumevale and Navaldia exhibit high acceptance rates. In contrast, Lyndrassia shows a significantly lower acceptance rate, indicating that dams in this region may require additional inspections to meet the required standards.

Region	Average Annual Probability of Failure	Eligible Dams	Non-Eligible Dams	Enrollment Rate
Flumevale	0.00852	2,958	564	84%
Lyndrassia	0.00925	4,437	3,969	53%
Navaldia	0.00940	7,082	1,796	80%

Table 4: Summary of the Filter Criterion

Hypothesis Testing

We begin by analyzing the probability distribution of dam failure across different regions and hazard levels. According to the *Figure 3*, there's clearly a different shape in distribution across all regions. While the distributions generally approximate a normal distribution, some exhibit slight skewness due to the limited sample size in some categories. Nevertheless, after plotting the overall normal distribution for the probability of failure, we find that it aligns reasonably well with expectations, due to the Central Limit Theorem *(Appendix I & C2).*

Additionally, we plot Q-Q plots for the annual probability of failure across each region, and the results indicate that all distributions follow a normal pattern *(Appendix C1)*. These visualizations further support the assumption that the failure probabilities are normally distributed across the different regions.

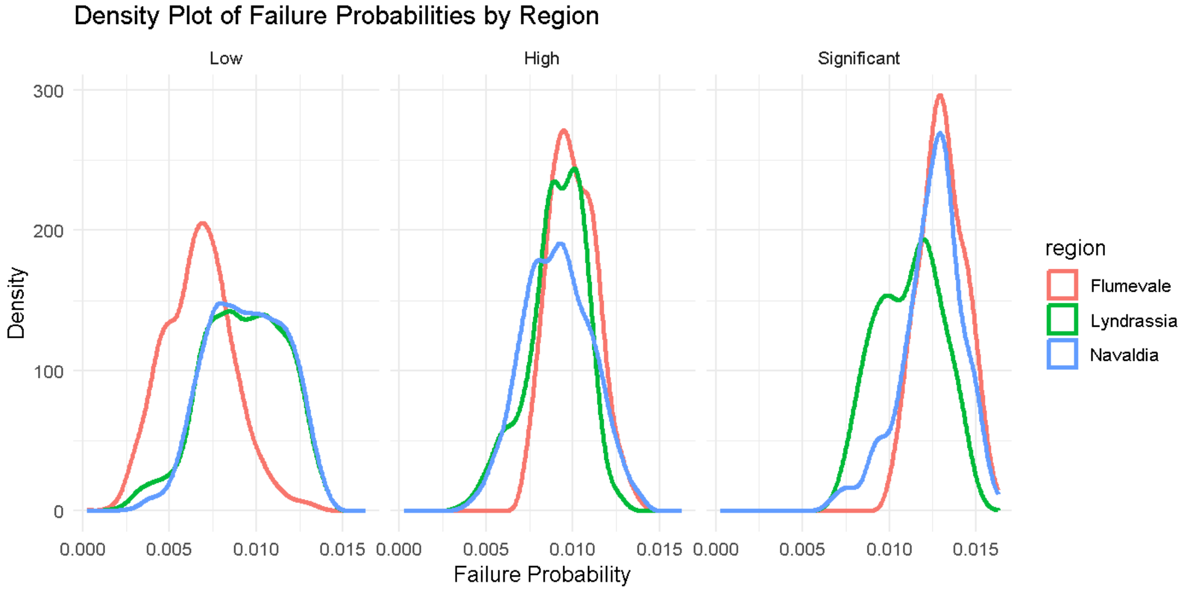


Figure 3: Density Plot of Annual Failure Probabilities by Region

To provide further evidence of regional differences, we conduct Hypothesis Testing to determine if there are significant variations in the underlying distributions across the regions. Specifically, we use the Tukey Honest Significant Difference (HSD) Test to do a pairwise comparison of the means across the three regions.

Region Comparison	Difference in Mean	p-value
Lyndrassia - Flumevale	0.00082	0.00000
Navaldia - Flumevale	0.00084	0.00000
Navaldia - Lyndrassia	0.00003	0.82220

Table 5: Tukey HSD Test Comparisons of Means Across Region

From *Table 4*, we see that the p-value for the comparison between Navaldia and Lyndrassia is greater than 0.05 (p-value = 0.82220), which suggests that there is no significant difference in the means between these two regions. Therefore, we fail to reject the null hypothesis (Ho), indicating that the means for these two regions are statistically similar. However, for the other region comparisons (Lyndrassia - Flumevale and Navaldia - Flumevale), the p-values are less than 0.05, suggesting significant differences in their means.

Region Comparison	Estimated Variance Ratio	95% Confidence Interval	p-value
Lyndrassia - Flumevale	1.436	(1.359,1.519)	0.000
Navaldia - Flumevale	1.335	(1.264,1.411)	0.000

Navaldia - Lyndrassia	0.930	(0.891,0.970)	0.001

Table 6: F Test Comparisons of Variance Across Region

To examine whether there are differences in variance across the regions, we apply both the Levene's Test and the F-test. Levene's Test, shown in the *Appendix J)*, indicates that there is at least one group with a significantly different variance, with a p-value less than 0.05. Further, the F-test results presented in *Table 5* confirm that the variances across the regions differ significantly, all p-values are less than 0.05. Thus, we can reject the null hypothesis (Ho) that the variances are equal across the regions.

Based on the results of the Tukey HSD test and F-test, we conclude that there are significant differences in both the mean and variance of dam failure probabilities across the regions. This suggests that the probability distribution varies by region, and we should consider these differences when setting premiums or designing region-specific interventions.

Time Series Analysis - Inflation

The goal of this project is to establish an equitable premium structure across all regions while maintaining sufficient reserves for unforeseen future expenses. The pricing of premiums for the national dam insurance program is designed to balance affordability and financial sustainability. However, the inflation is crucial when determining the premium for each year. Thus, we created the ARIMA model to help us forecast inflation for the future. We first assessed whether the time series data is stationary by performing the Dickey-Fuller test. The results showed that taking one or two differences made the data stationary, with a p-value less than 0.05. To determine the optimal time series model, we examined the autocorrelation (ACF) and partial autocorrelation (PACF) plots. The ACF cut off at lag 8, and the PACF cut off at lag 2, which indicated the appropriate model structure. Based on these plots, we selected the model with the lowest AIC. We also investigated whether taking two differences improved model performance, and found that the ARIMA(2,2,1) model yielded the lowest AIC. After analyzing the ACF and PACF plots, we identified the top two candidate models for further evaluation.

Model	AIC	BIC
ARIMA(2,2,1)	-5.232754	-5.078320
ARIMA(8,1,2)	-5.121289	-4.700644

Table 7: Inflation Forecast ARIMA(2,2,1)

Based on this *Table 6*, ARIMA(2,2,1) is a better model since it has a lower AIC and BIC. (Even though, the ARIMA(8,1,2) model looks more pleasing according to *Figure 5*, but according to the Goodness-of-Fit, we will use ARIMA(2,2,1)). According to *Figure 4*, we can see that the forecasted inflation is around 4% each year.

$$x_t^n = \phi_1 x_{t-1}^{\mathrm{n}} + \phi_2 2x_{t-2}^{\mathrm{n}} + \theta_1 x_{t-1}^{\mathrm{n}}$$

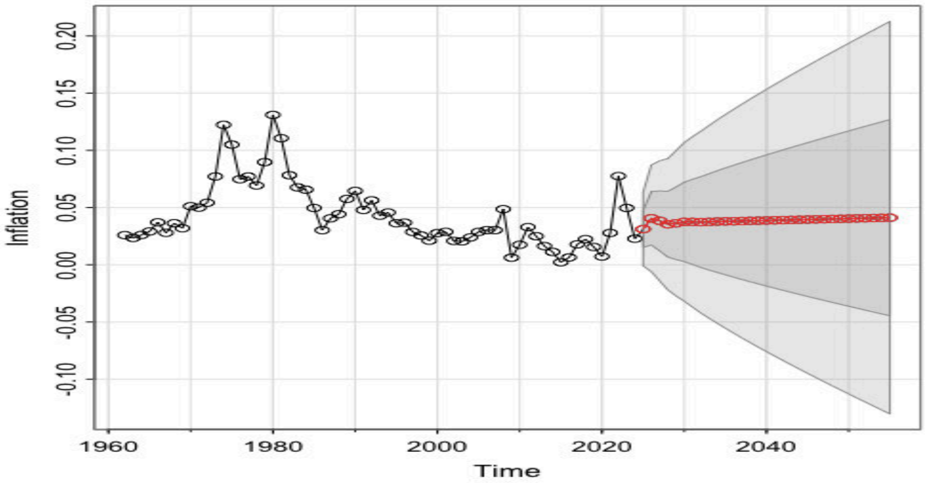


Figure 4: Inflation Forecast ARIMA(2,2,1)

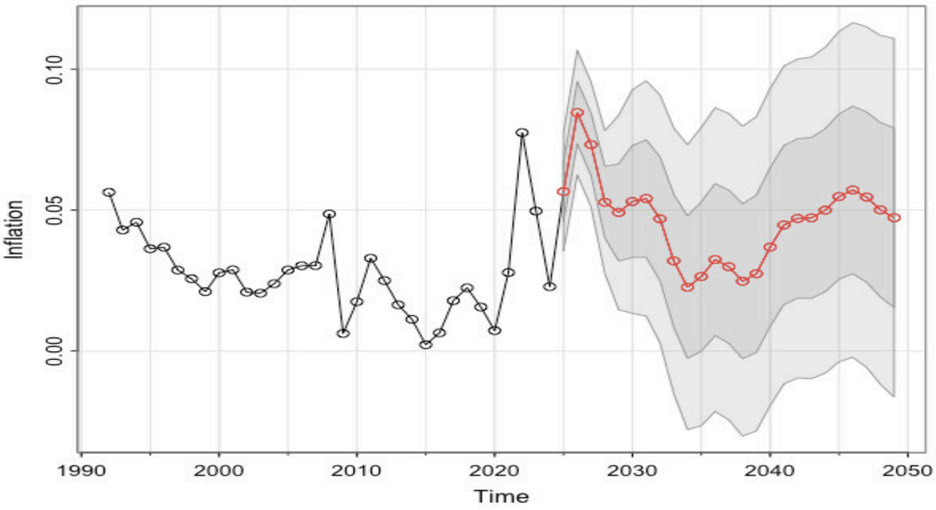


Figure 5: Inflation Forecast ARIMA(8,1,2)

Based on *Figure 4*, the projected inflation rate remains consistently between 4% and 5%, suggesting that costs are expected to rise by approximately 4% to 5% each year. Consequently, the premium structure must be designed to withstand and adapt to this steady inflation over time. In contrast, *Figure 5* shows that projected inflation is much more volatile, ranging from approximately 2% to 10%. Initially, inflation is expected to rise, then gradually converge toward the 4% to 5% range. However, as shown in *Table 7*, the ARIMA(8,1,2) model performs worse than the ARIMA(2,2,1) model due to the volatility in its projections and the added complexity of the model. Therefore, a simpler and more interpretable model, such as the parsimonious ARIMA(2,2,1), is preferred for projecting inflation. Thus, ARIMA(2,2,1) model will be used to estimate the future cost (*Figure 7*).

Expected Loss Calculation (Premium) - Frequency and Severity Model

In an actuarial setting, it is common to combine frequency and severity to calculate the expected aggregate loss. Aggregate loss is found by multiplying the number of claims (frequency) by the cost per claim (severity). For example, if a claim costs $500 and occurs 3 times, the aggregate loss would be $1,500. For simplicity, we assume that frequency and severity are independent—meaning the number of claims does not influence the cost of each claim. Under this assumption, the total expected aggregate loss can be expressed as the product of the expected frequency and expected severity. Let N represent the number of claims (frequency), X the loss given a claim (severity), and S the total aggregate loss.

$$E[S] = E[NX] = E[N]E[X]$$

To ensure sufficient reserves are available to cover claims each year, we estimate the expected annual total loss and extreme cases using Frequency and Severity Modeling. This approach accounts for both typical losses and worst-case scenarios. The expected total loss from dam failures is calculated using frequency and severity modeling, where the total loss is determined by multiplying the number of claims per year (frequency) by the loss amount per claim (severity). After classifying the hazards based on their assessment ratings, we model the distributions of frequency and severity as follows.

Given the probability of dam failure over a 10-year period, the annual failure probability is calculated, which represents the annual probability of failure. Note, the annual probability was calculated under the assumption that each year is independent of the others. To determine the best-fitting distribution for annual failure probabilities, we compared different statistical models using AIC, BIC, and Log-likelihood. The results are summarized in *Table 7*.

Model	AIC	BIC	Log-likelihood
Normal	-133,735.4	-133,720.2	66,869.7
Log-Normal	-131,226.4	-131,211.2	65,615.2
Beta	-132,545.5	-132,530.4	66,274.8

Table 7: Frequency Model Goodness-of-Fit Evaluation

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The result suggested the annual failure probabilities follow a normal distribution since it has the lowest AIC and BIC, and it is supported by the Q-Q plot in *(Appendix C2*).

We are provided with the severity data for each dam, which shows a right-skewed distribution, as shown in *Appendix H*. Therefore, we consider three right-skewed distributions for the severity: Log-Normal, Gamma, and Weibull. The performance of these models is summarized in *Table 8*.

Model	AIC	BIC	Log-likelihood
Log-Normal	284,269.7	284,632.2	-142,312.8
Gamma	282,971.3	282,973.8	-141,483.6
Weibull	282,945.9	282,948.4	-141,471.0

Table 8: Severity Model Goodness-of-Fit Evaluation

Therefore, the most appropriate model for severity is the Weibull distribution. We will utilize the frequency and severity model to calculate the 95% confidence interval for the annual expected loss using Monte Carlo Simulation.

$$g(x; \alpha, \beta) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}} ; x \geq 0$$

Monte-Carlo Simulation Methods

The optimal model for the annual probability of failure follows a Normal distribution, while the loss distribution is best represented by a Weibull distribution. In a dataset, there is only one total expected loss, but we are interested in the range of total expected loss. Therefore, we will be performing Monte Carlo simulation to estimate the 95% confidence of total expected loss. The procedure of the simulation is as follow:

1. Calculate the maximum likelihood estimates (MLE) of each parameter. For the annual probability of failure, estimate the parameters μ and σ assuming a Normal Distribution. For the severity (loss given failure), estimate the parameters α and β assuming a Weibull distribution.
2. Generate N observations of the annual probability of failure from the Normal distribution using the estimated parameters, and generate corresponding loss given failure from the Weibull distribution using its estimated parameters.
3. Sum the aggregate loss for each simulation, and repeat the process for S iterations.

$$N \sim N(\mu,\ \sigma)$$

$$X \sim Wei(\alpha, \beta)$$

The Monte Carlo simulation for the frequency and severity individually are shown in *Appendix G*. As shown in *Figure 6* with S = 100,000 iterations, we are 95% confident that the annual total expected loss from the insurance coverage will fall between 39,406 million Q and 40,406 million Q. The 97.5th percentile gives us an idea on how well does premium structure tolerate the extreme case scenario.

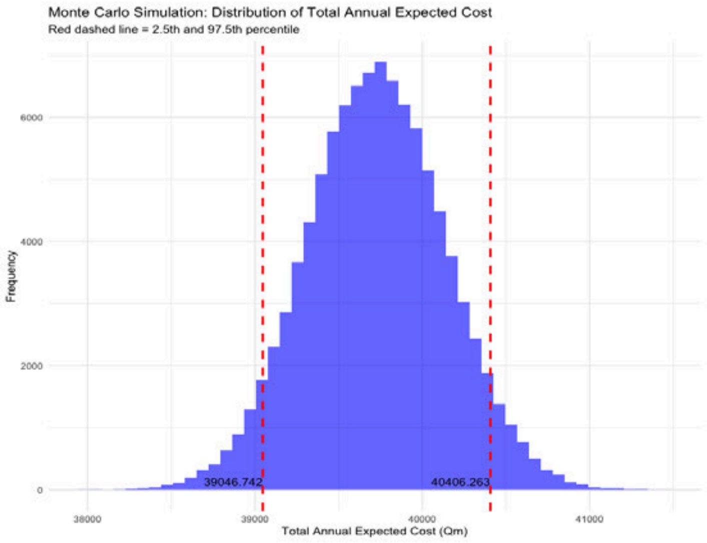


Figure 6: Monte Carlo Simulation of Total Annual Expected Cost

Financial Results

We incorporated inflation to visualize the approximate cost for each year by region and hazard level. This allows us to identify which hazards are associated with the highest expected costs and what region requires extra supervision (Research Question 2). Based on *Figure 7*, we observe that Flumevale and Navaldia follow a similar trend, where high-hazard dams contribute the most to expected costs—especially since most dams in Tarrodan are classified as high hazard. Notably, Flumevale shows a higher expected cost for significant hazards, which makes sense given its proximity to a city, where dam failure could result in catastrophic losses. As a result, we recommend extra government supervision on the dam in the Flumevale region, and it's important to reduce the Flumevale's dam probability of failure as low as possible. We also see that Navaldia has the highest overall cost since Navaldia has the number of observations in the data. Lyndrassia, on the other hand, displays a unique trend where low-hazard dams contribute the highest cost. This is because Lyndrassia is a rural area, and the majority of its dams are classified as low hazard. Consequently, we expect Lyndrassia to contribute the smallest proportion of total costs over the years.

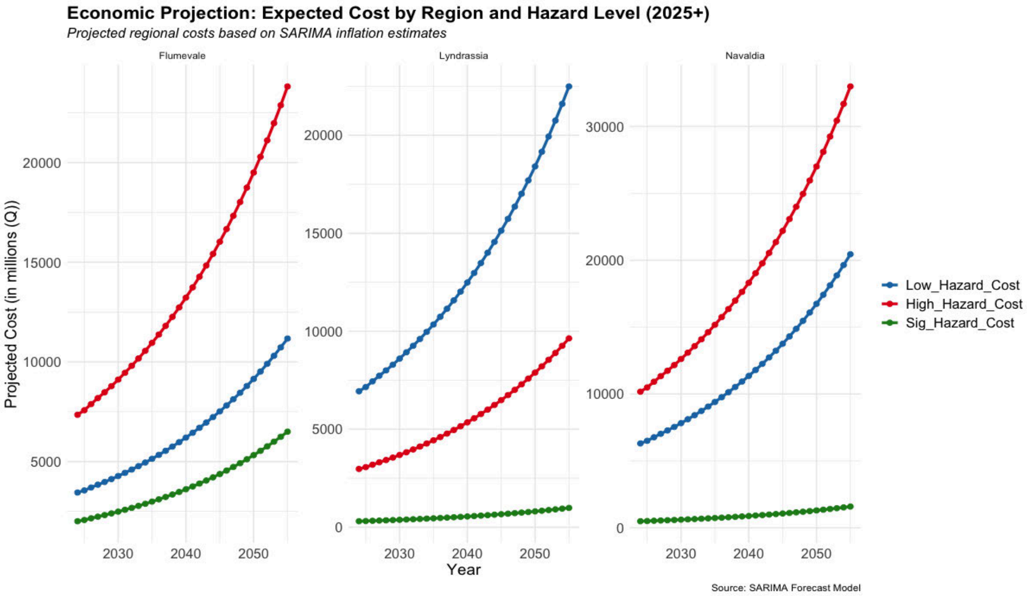


Figure 7: Financial Projection

Aggregate Statistical Model (Premium Prediction Model)

In this section, we'll explore what model is the best at predicting the expected total cost. In the claim data, the response variables are often right skewed as there are a lot of claims with no claims or a low cost. In the insurance field, Tweedie distribution is often used to model the claim to account for the data that has heavy weight at around 0. According to *Figure 8*, we can see that there's a lot of expected cost located around 0. In this section, we'll explore the generalized linear model with a link function of Tweedie distribution.

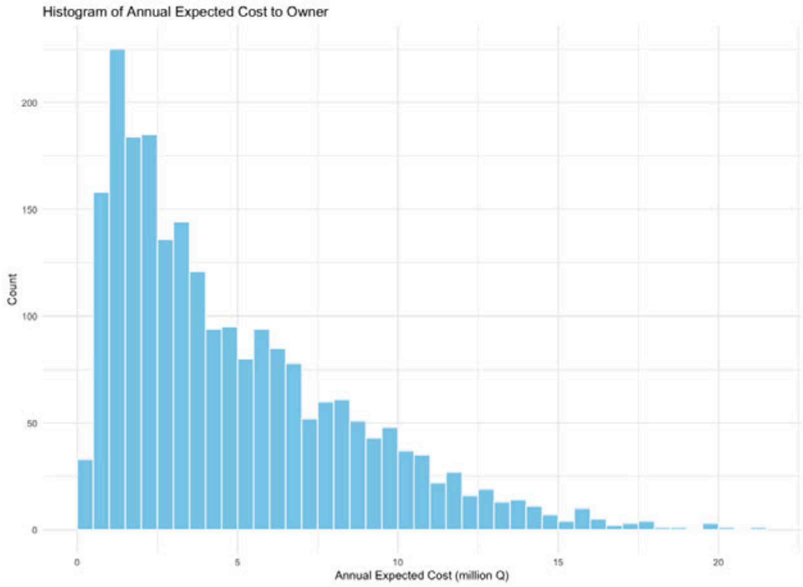


Figure 8: Histogram of Annual Expected Cost

The Tweedie distribution is a special case of an exponential distribution. It is useful when the data has a cluster at zero. The Tweedie distribution has the following characteristics:

$$E[Y] \; = \; \mu$$

$$Var[Y] \; = \; \phi\mu^{p}$$

The p is an additional shape parameter for the distribution and ϕ is the dispersion parameter. The value of p determine the shape of the distribution, and the details are provided in *Table 9*.

$p = 0$	Normal Distribution
$p = 1$	Poisson Distribution
$1 < p < 2$	Compound Poisson/Gamma Distribution
$p = 2$	Gamma Distribution
$2 < p < 3$	Positive Stable Distribution
$p = 3$	Inverse Gaussian Distribution / Wald

	Distribution
$p > 3$	Positive Stable Distributions
$p = \infty$	Extreme Stable Distributions

Table 9: Description of p value for Tweedie Distribution

First, we estimated the optimal value of the power parameter p to determine the shape of the distribution. Using maximum likelihood estimation in R, we found the optimal p to be approximately 1.5, indicating that the data follows a Compound Poisson-Gamma distribution. Based on this, we retained the significant predictors in the model, and the results are presented in *Table 10*.

	Estimate	P-Value
(Intercept)	-10.8512	0.0120
RegionLyndrassia	-0.1115	0.0018
RegionNavaldia	-0.4760	0.0000
Regulated.DamYes	-0.3652	0.0001
Height..m.	0.0082	0.0000
Length..km.	0.2609	0.0000
Year.Completed	-0.0021	0.0020
Last.Inspection.Date	0.0088	0.0103
Distance.to.Nearest.City..km.	-0.0058	0.0000
HazardLow	-0.8098	0.0000
HazardSignificant	0.2165	0.0002
AssessmentNot Rated	0.1515	0.0412
AssessmentPoor	-0.0228	0.8374
AssessmentSatisfactory	-0.1374	0.0023
AssessmentUnsatisfactory	0.4056	0.0000

Table 10: Generalized Linear Model with Tweedie Distribution Output

Table 10 summarizes the results of a Tweedie generalized linear model analyzing factors associated with the annual expected cost. Several predictors show statistical significance, including region, dam regulation status, structural characteristics such as height and length, year completed, and hazard and assessment categories. Positive coefficients suggest that factors like greater dam length, unsatisfactory assessment ratings, and regulated status are associated with an increase in the annual expected cost. Negative coefficients indicate that being located farther from a city, lower hazard classifications, and certain region indicators are associated with a decrease. Some variables, such as the "poor" assessment rating, do not appear to have a significant effect. Overall, the model suggests that both physical characteristics and evaluation metrics of dams play a role in influencing the annual expected cost. For the model diagnostics, refer to *Appendix L* for further details.

Next, we perform a diagnostic on generalized linear regression with Tweedie distribution. From *Figure 9*, we can see that the Tweedie distribution residual plot is clustered around the middle and the residuals are fairly uniformly distributed, which implies homoscedastic variance. This means that the model is fairly good at predicting high and low values.

Lastly, we compared the model's performance with ordinary linear regression and a generalized linear model using a Gamma distribution. Since linear regression assumes a normally distributed response, a log transformation was applied. However, as shown in Table 11, the Tweedie model demonstrated the best predictive accuracy. Linear regression performed the worst, struggling with the many observations that had very low expected costs, even after transformation. The Gamma model performed better due to its ability to handle skewed data and some degree of zero inflation. The Tweedie model further improved on this by introducing a power parameter, making it particularly effective for modeling data that is both highly skewed and zero-inflated. As a result, we selected the Tweedie distribution to determine policyholder premiums. Note that this prediction result is only for the year 2025 only, so it's important to adjust the model to the new data to incorporate the increase in cost.

Models	MSE
Linear Regression (log transformed)	73.9790
Gamma	30.2537
Tweedie	11.3894

Table 11: Models Comparison

DISCUSSION

Based on our results, the underwriting process—specifically filtering dam eligibility based on hazard level and assessment rating—led to a notable reduction in the average annual probability of failure by region, decreasing by approximately 5–8%, as shown in Table 2 and Table 4. Enrollment rates were above 80% in most regions, except for Lyndrassia, which had only 53% enrollment. This lower rate is mainly due to a large number of dams lacking assessment ratings. To improve participation and ensure broader insurance coverage, we recommend that government agencies or related authorities prioritize inspection efforts in Lyndrassia. Increasing assessment coverage would expand eligibility for the program, thereby reducing public financial vulnerability in the event of a hazard.

In our underwriting approach, we focused on two key classification variables: hazard level and assessment rating. These were identified as important predictors through a random forest model, which ranked them among the top variables based on reduced IncNodePurity. To support and validate this selection, we also fitted a generalized linear model (GLM). The GLM confirmed the statistical significance of both hazard level and assessment rating in predicting dam failure (see *Appendix K*), reinforcing our decision to use them as primary criteria. In this way, the GLM served to support the results from the random forest model, ensuring our underwriting choices were both data-driven and statistically sound.

While other variables—such as dam height, year completed, and length—also showed high importance in the random forest model, we excluded them due to a large proportion of missing values that could compromise model accuracy. Similarly, although the GLM identified additional significant variables like region, dam purpose, surface area, inspection frequency, and assessment date, many of these had data quality issues or posed challenges for standardized pricing. For example, we chose not to include dam purpose because the dataset only listed the primary purpose, and many dams serve multiple functions. Creating an equitable pricing structure for such cases would require deeper domain-specific research beyond our scope.

From *Appendix E*, we observed that the average annual probability of failure decreases as inspection frequency increases, particularly when the frequency reaches three or more inspections. This insight underscores the importance of regular inspections in reducing dam failure risk, and it can be used to inform risk mitigation strategies, even though inspection frequency was not included in the underwriting criteria due to missing data.

Hypothesis testing revealed that dam failure probabilities differ significantly across regions, prompting us to implement a region-specific premium structure. Under this structure, dams in different regions—even with the same classification—would pay different premiums. For instance, a dam in Lyndrassia would not pay the same premium as one in Navaldia, even if both have similar risk profiles. We built a frequency-severity model, treating the total estimated loss per event as the severity component and the annual probability of failure as the frequency. This allowed us to estimate the net present value of expected aggregate losses, which forms the basis for premium pricing. With a large applicant pool, the insurer can diversify risk and ensure liquidity over time.

Because our insurance plan targets both short- and long-term coverage (e.g., 30-year plans), we required a predictive model that accounts for future changes in cost and risk. For this, we implemented a GLM with a Tweedie distribution, which outperformed both linear regression and a GLM with a Gamma distribution. This model is well-suited for skewed, zero-inflated data like ours and provided a robust framework for pricing future expected losses. It also enables dynamic pricing by incorporating inflation predictions.

To support the long-term stability of the insurance program, it is essential for the government to maintain adequate reserve capital to cover projected losses, including under adverse scenarios, as illustrated in Figure 6. Diversifying investments of collected premiums can also help grow the fund and reduce the risk of financial shortfalls over time.

Our use of a generalized linear model (GLM) with a Tweedie distribution provided a strong foundation for pricing, especially given the skewed and zero-inflated nature of the loss data. However, the model currently yields a mean squared error (MSE) of approximately 11 million Q, indicating that further refinement is needed.

We also aim to evaluate the robustness of our premium structure under various risk scenarios. In addition, we are considering strategies such as reinsurance, contingency reserves, and policy adjustments to strengthen the program's ability to manage future uncertainty and maintain financial resilience. By aligning our underwriting criteria with both machine learning insights and statistical validation, we've built a framework that supports regionally tailored, risk-based pricing and contributes to a more sustainable and data-informed insurance program.

DATA LIMITATIONS AND FUTURE CONSIDERATIONS

This dataset contains a significant amount of missing data, especially for time-related variables such as year of completion, inspection dates, and assessment years. This makes it challenging to fully assess the reliability of our results, particularly for evaluating the impact of government intervention and building a strong financial model. For instance, our time series analysis on the year of dam completion proved unreliable due to sparse data before 1900. This gap could affect our underwriting process, especially since we rely on assessment ratings that are time-sensitive.

A specific concern is the lack of recent assessment dates. For example, if a dam's last assessment was in 1970 and it was rated as satisfactory at that time, that rating may no longer hold true in 2023 due to potential deterioration. Ignoring the timing of assessments introduces bias in the underwriting process, which is why we recommend that assessments be updated at least every five years. To address this data gap, our program design includes mandatory short-term inspections for all dam owners. This requirement not only improves the accuracy of the underwriting process but also ensures that decisions are based on current data rather than outdated assumptions.

Another limitation is the dataset's limited geographic detail. While it includes general regional information and the distance from the nearest city, it lacks precise spatial data or hazard mapping. This restricts our ability to design more tailored insurance models. For example, in the region of Lyndrassia, a dam located in a mountainous area may face higher pressure and risk than one located on flatter ground, and a dam situated in a densely populated city would likely cause more costly damage in the event of a failure. With more detailed spatial data, we could design better region-specific insurance plans. This is a key area for future research.

We also lack environmental and climate data, such as rainfall patterns or natural disaster frequency. These factors are important because regions with heavier rainfall, for example, have a higher likelihood of flooding, which should be reflected in premium pricing. Without this information, we are unable to incorporate these potentially significant covariates into our model, which could impact the accuracy of both failure probability estimates and total loss projections.

Furthermore, the dataset provides dam failure probabilities on a 10-year basis, but we apply them as if the annual probability is constant and independent from year to year. In reality, dam failure events are not independent—if a dam fails or shows signs of weakening, the risk of future issues is likely to increase. This means our assumption of consistent annual failure probability may not be realistic. To improve our model, we would need historical event data to better estimate annual probabilities and understand how risks evolve over time.

Finally, our assumption that total losses remain constant each year is another simplification that may not hold true in practice. Urban development, population growth, or government relocation policies could all significantly change the financial impact of a dam failure over time. For example, an area that becomes more densely populated would lead to higher third-party losses, while a region undergoing relocation efforts might see reduced exposure. These dynamic factors should be considered to enhance the model's realism and accuracy.

In summary, while this dataset provides a foundation for our analysis, it lacks several key dimensions that would improve the model's accuracy, including recent inspection data, precise geographic and climate information, and dynamic risk factors. These limitations should be addressed in future research to build a more robust, reliable, and actionable national insurance framework.

REFERENCES

[1] Association of State Dam Safety Officials (ASDSO). (n.d.). *The National Dam Safety Program Roadmap*. Retrieved from https://damsafety.org/Roadmap

[2] Bensinger, D., Beck, A.-J., Guilbert, S., & Englert, J. (2025, February 12). *America's aging dams and other infrastructure is an urgent insurance coverage issue*. Pro Policyholder. https://www.propolicyholder.com/2025/02/americas-aging-dams-and-other-infrastructure-is-an-urgent-insurance-coverage-issue/

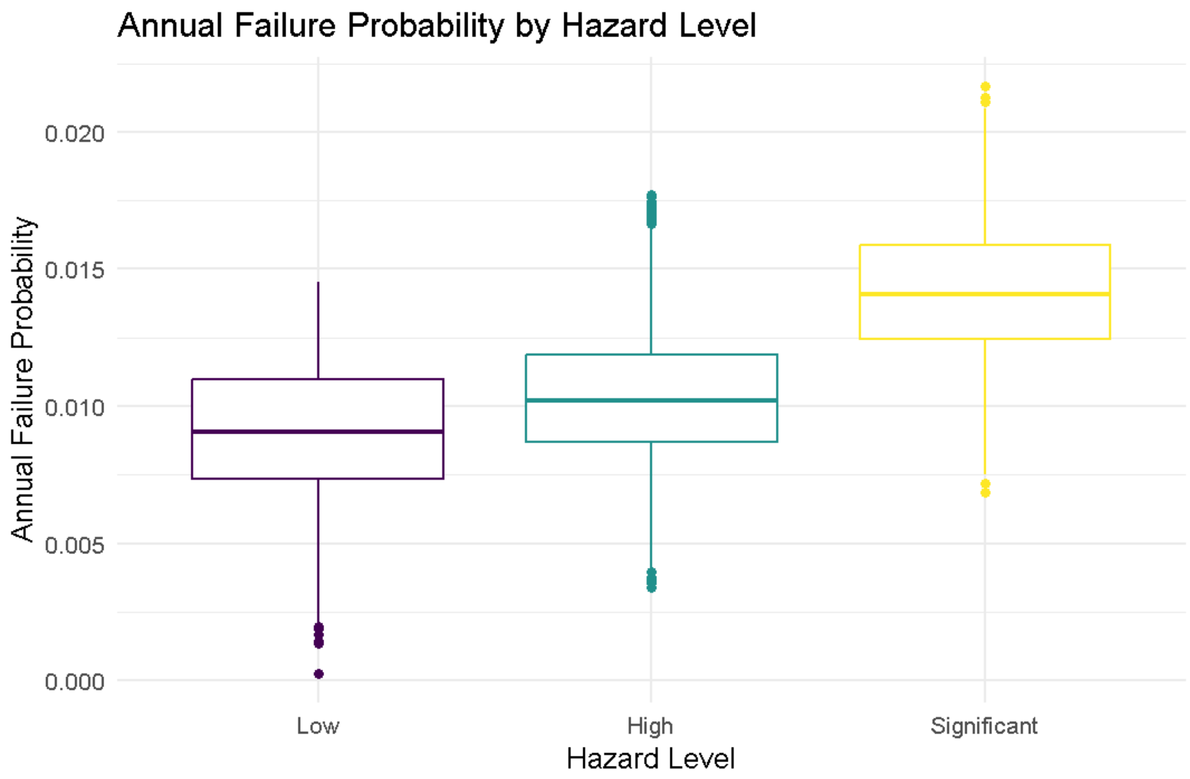
[3] Bharti, M. K., Sharma, M., & Islam, N. (2020). Study on the dam & reservoir, and analysis of dam failures: A database approach. *International Research Journal of Engineering and Technology (IRJET)*, *7*(5), 1661. https://www.irjet.net/

[4] Federal Emergency Management Agency (FEMA). (n.d.). *Dam Safety and Emergency Management*. Retrieved from https://www.fema.gov/emergency-managers/risk-management/dam-safety

[5] Society of Actuaries (SOA). (n.d.). *2025 Student Research Case Study Challenge*. Retrieved from https://www.soa.org/research/opportunities/2025-student-research-case-study-challenge/

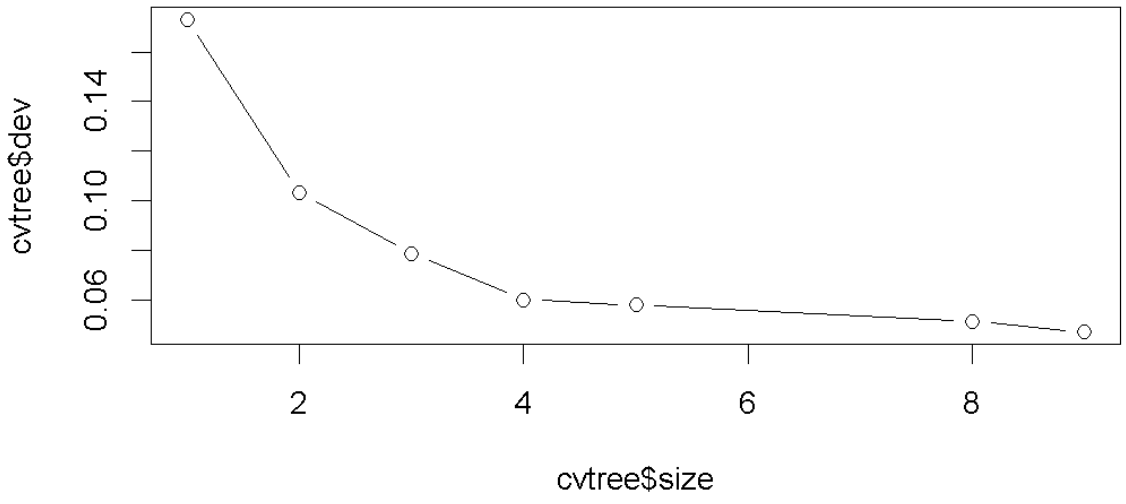
APPENDICES

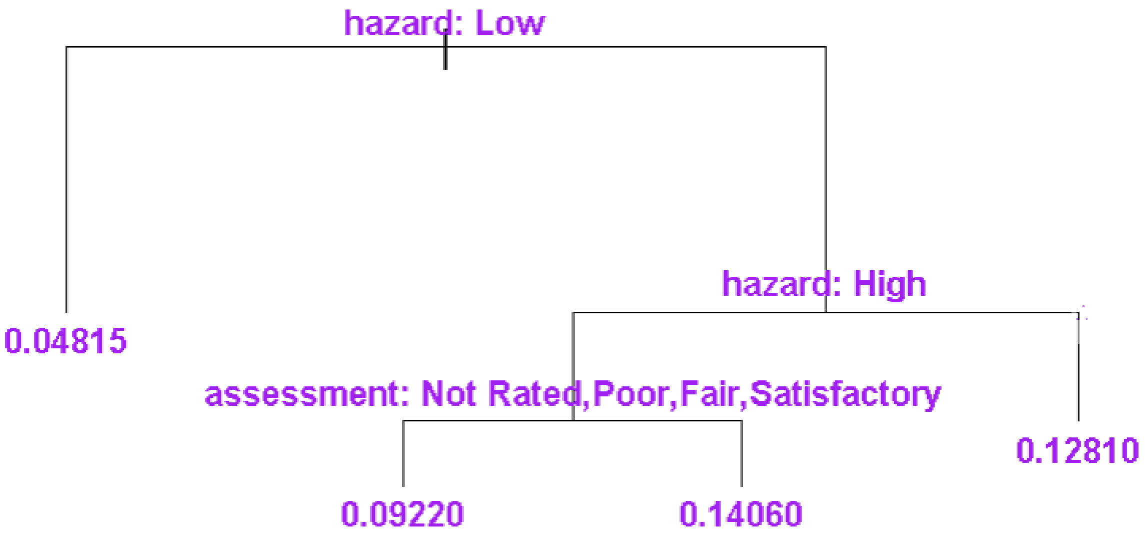
Appendix A: Annual Probability of Failure VS Hazard



Appendix B: Decision Tree (Pruned VS Unpruned)

Appendix B1: Pruned Tree





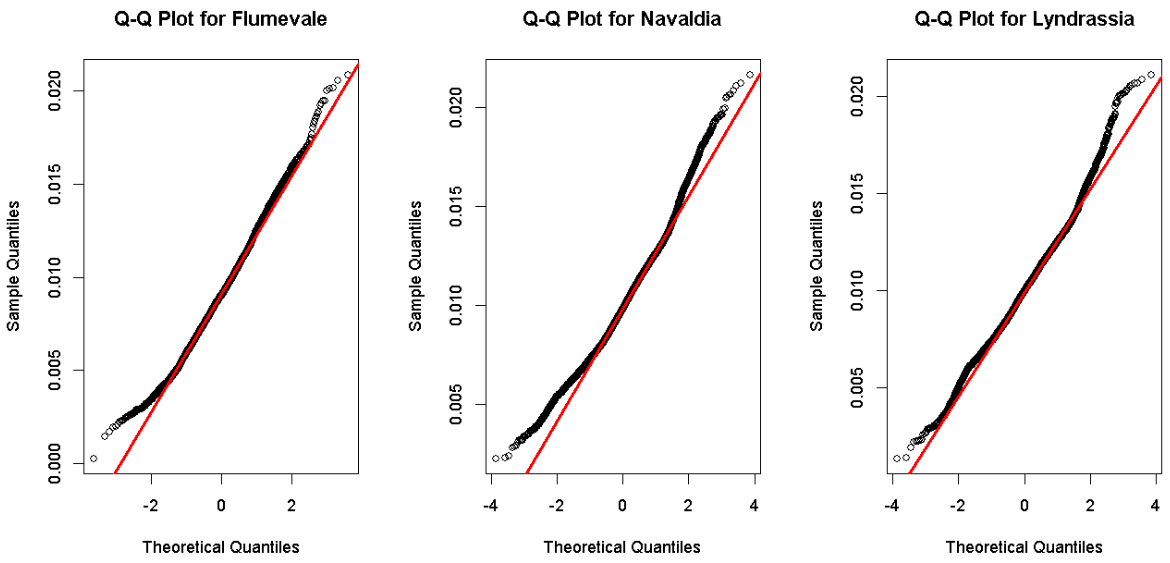
Appendix B1: Pruned Tree with "Best=4"

Appendix B2: Cross-Validation Between Pruned Tree and Unpruned Tree

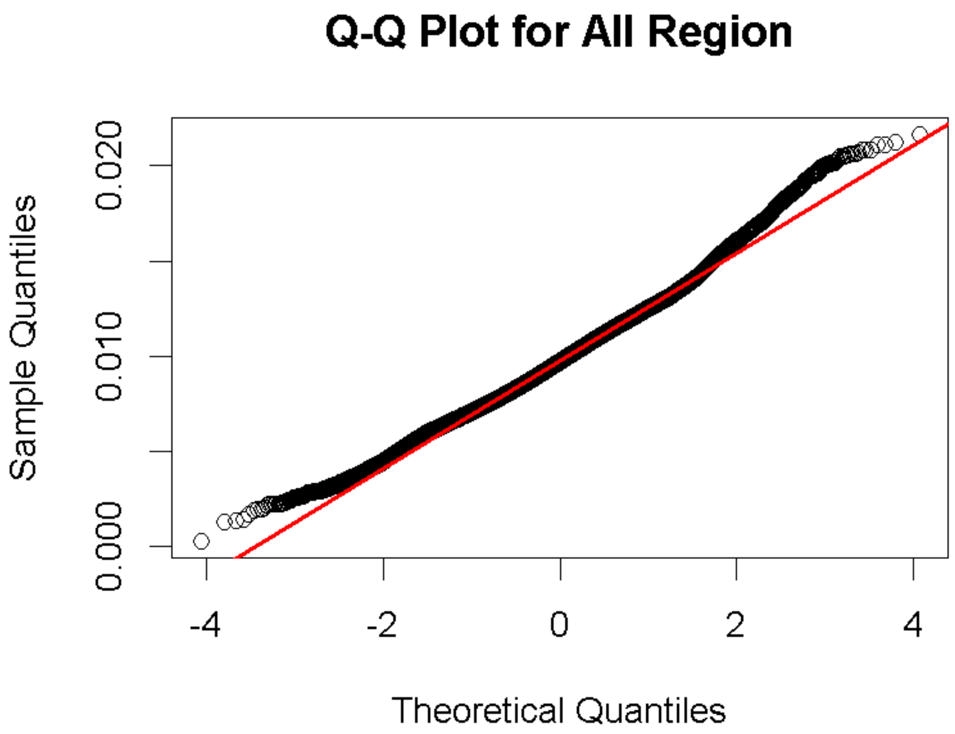
Decision Tree	Residual Mean Deviance	MAE
Unpruned	0.0003961	0.0314
Pruned	0.0007389	0.0321

Appendix C: QQPlot for the Failure Probability Distribution

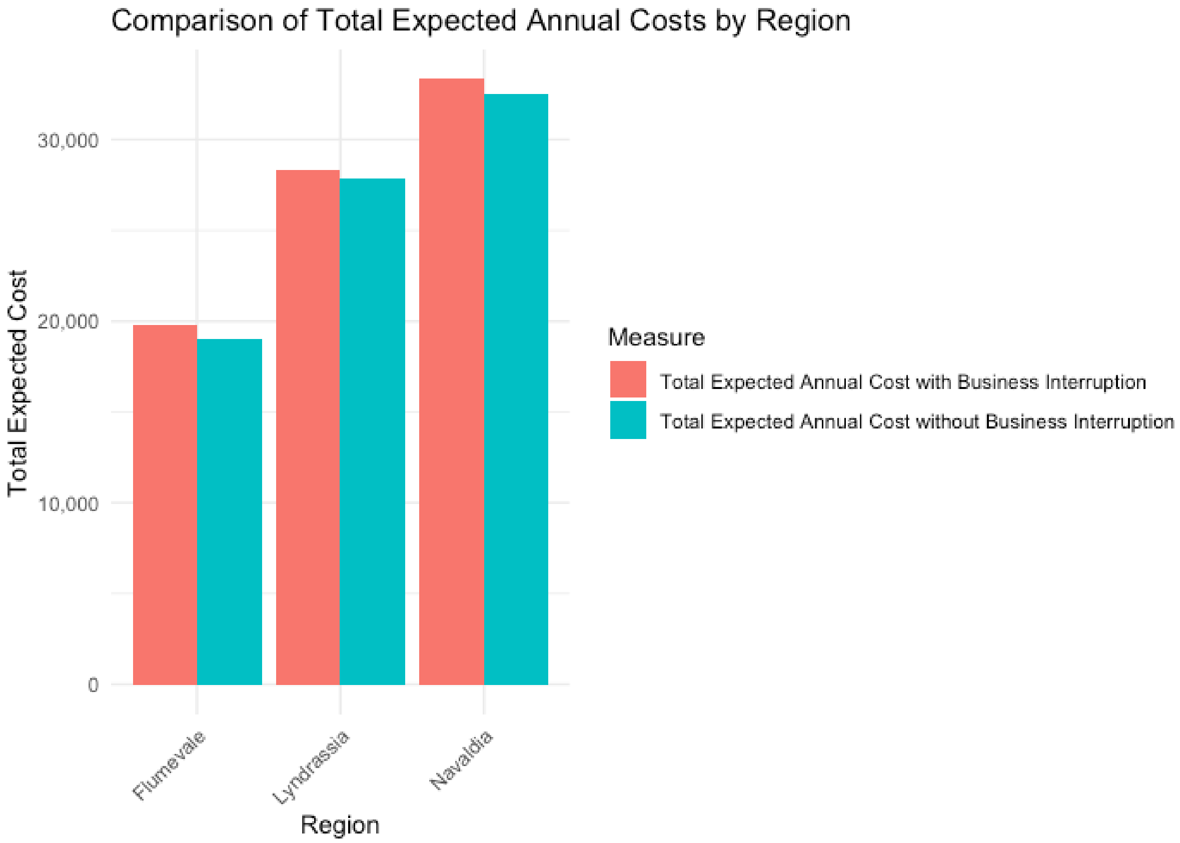
Appendix C1: Q-Q Plot for Failure Probability Across Each Region



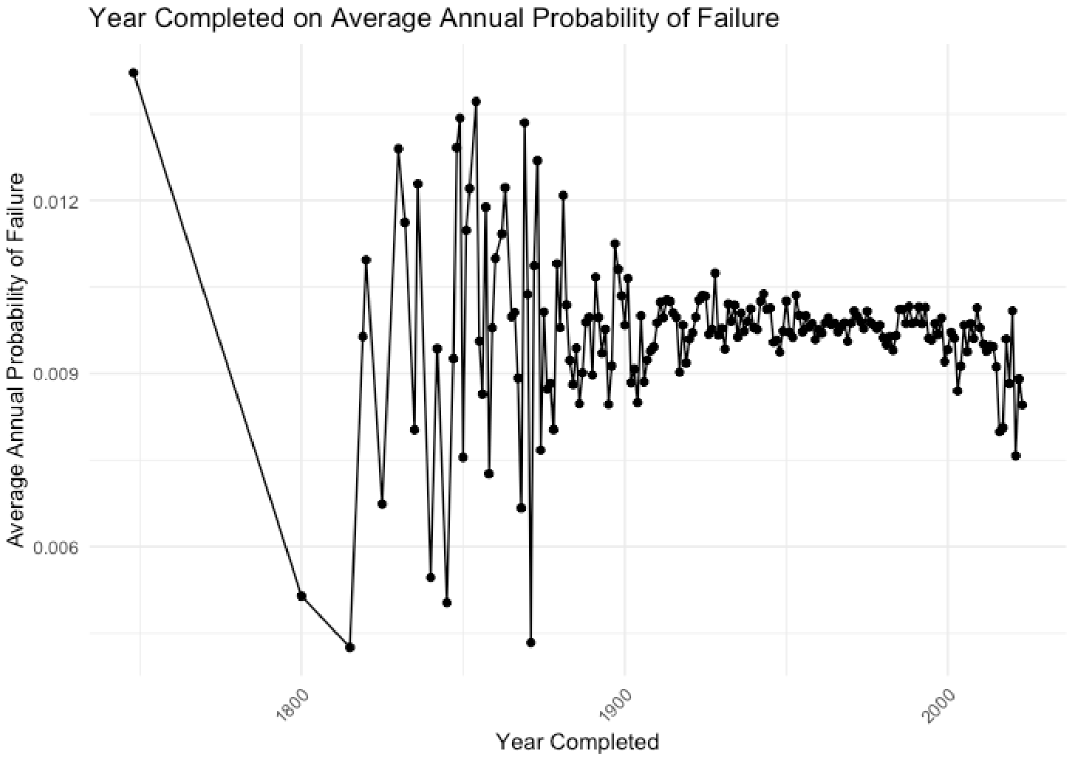
Appendix C2: Q-Q Plot for Failure Probability in All Region

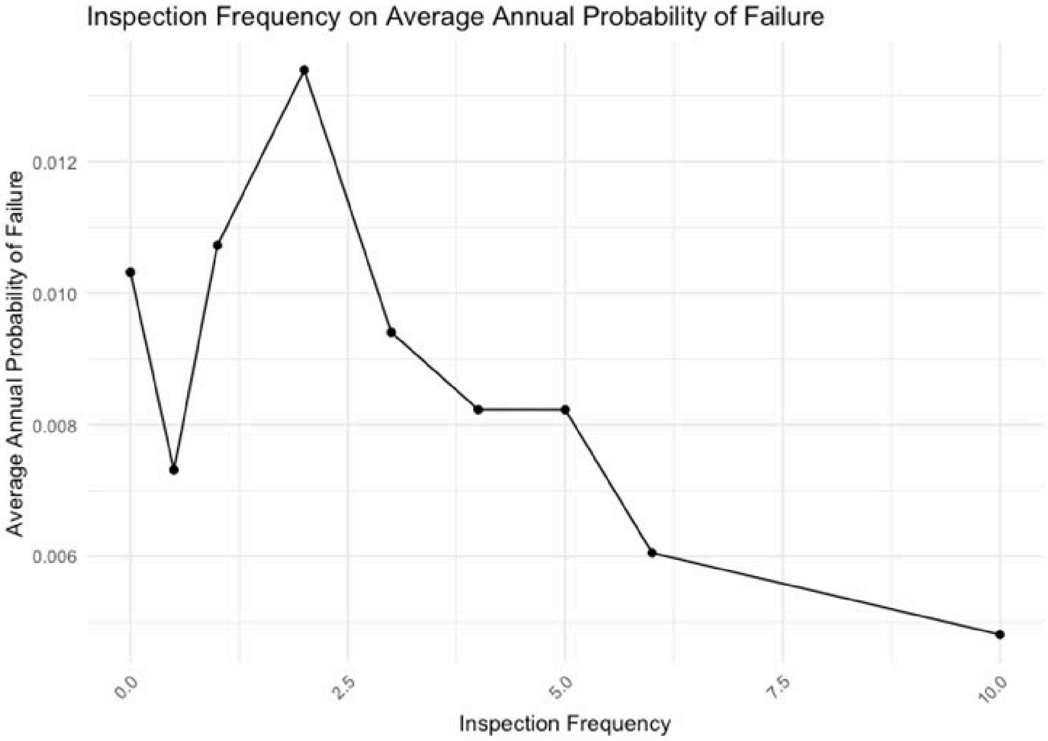


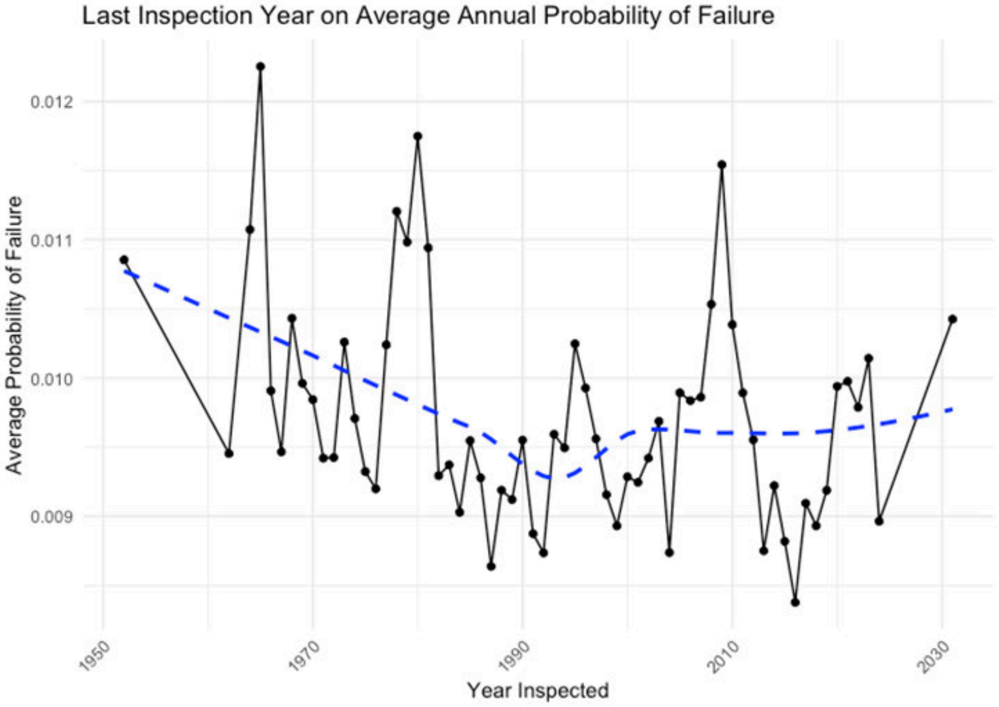
Appendix D: Hazard Level vs. Total Cost



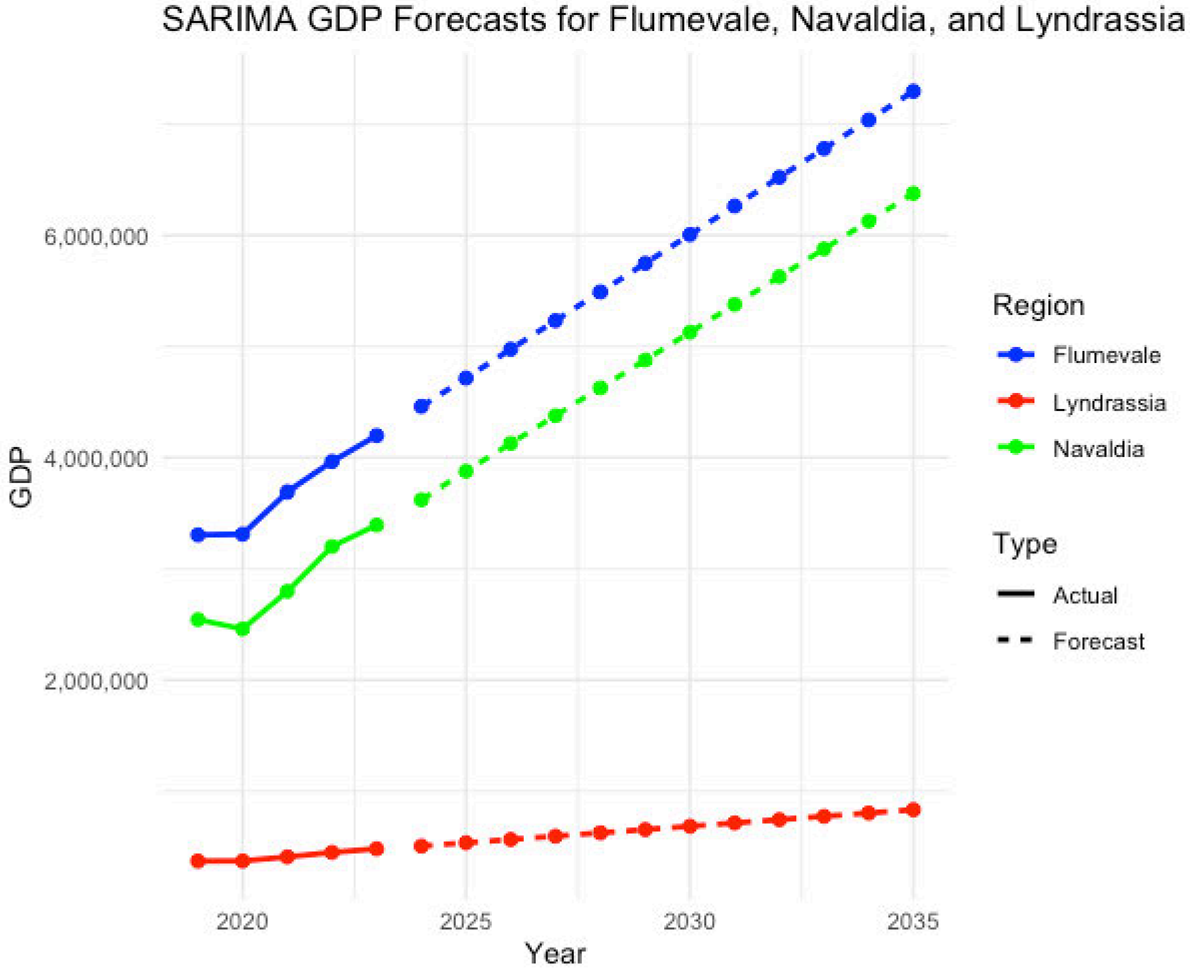
Appendix E: Year Completed, Inspection Frequency, Last Inspection Year Against Annual Probability of Failure

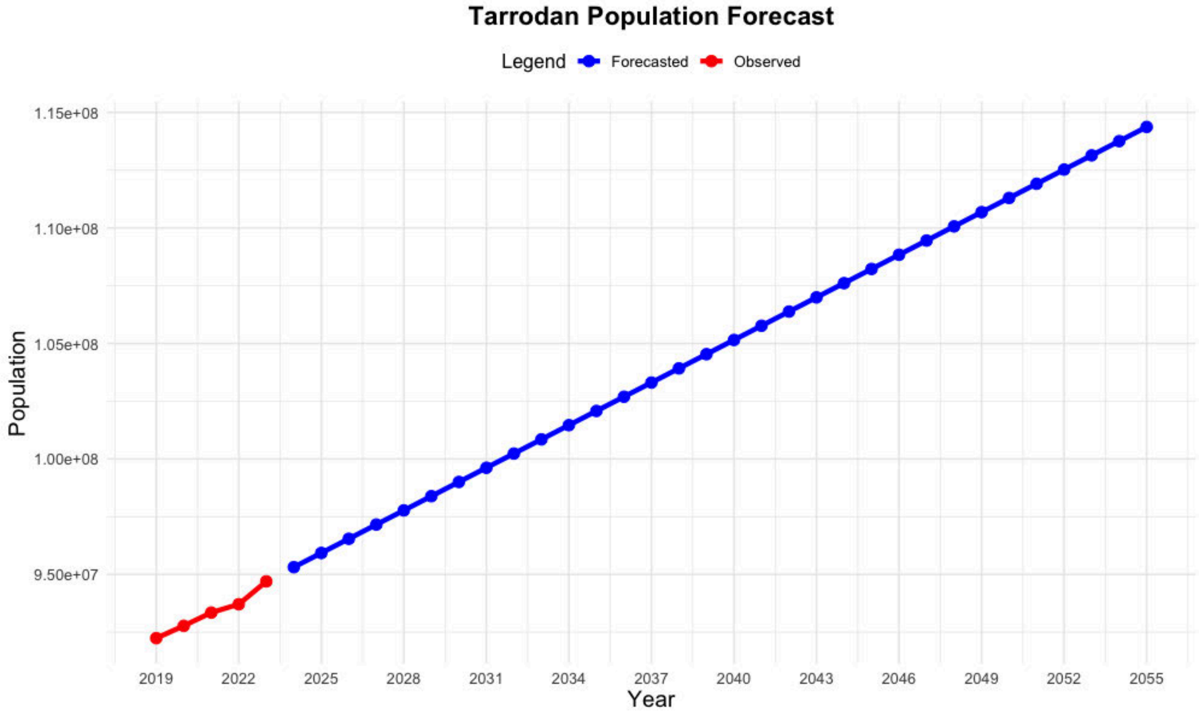


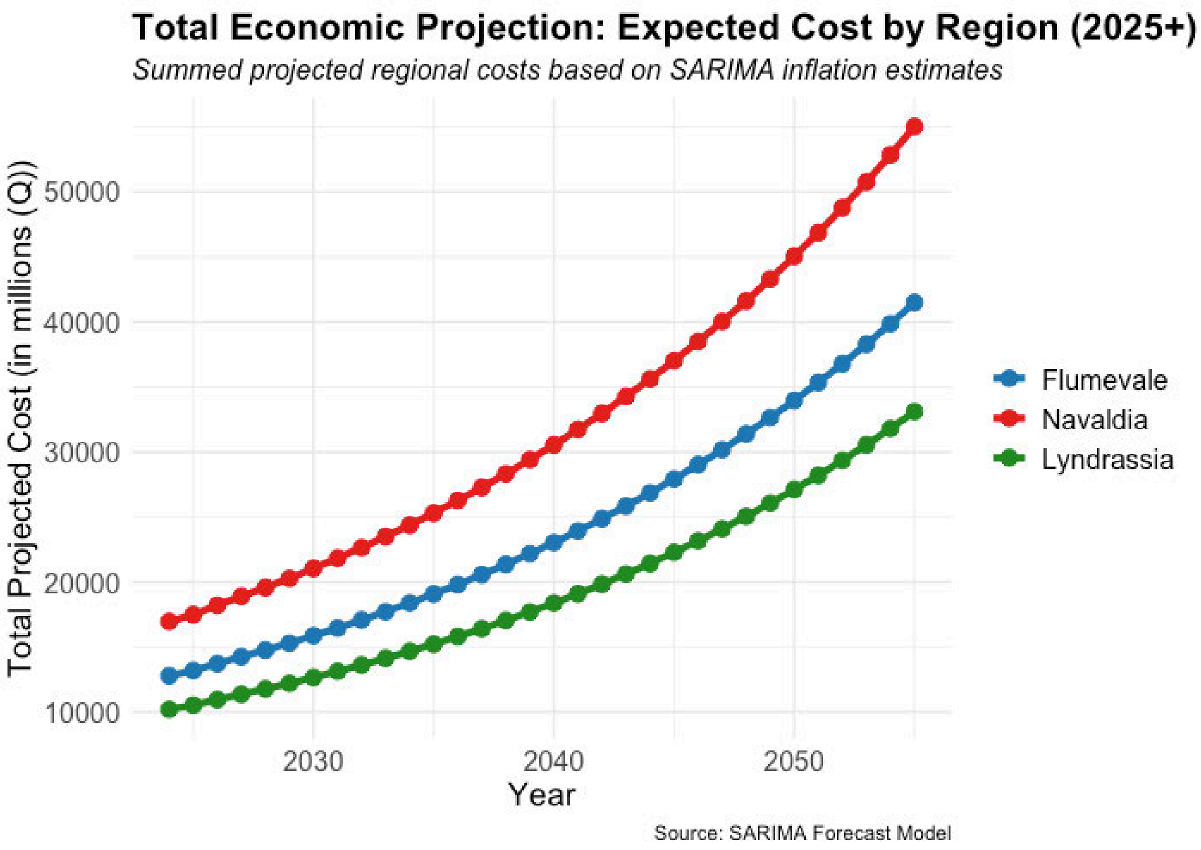

Inspection Frequency on Average Annual Probability of Failure
Average Annual Probability of Failure
0.012
0.010
0.008
0.006
0.0
2.5
5.0
7.5
10.0
Inspection Frequency


Last Inspection Year on Average Annual Probability of Failure
Average Probability of Failure
0.012
0.011
0.010
0.009
1950
1970
1990
2010
2030
Year Inspected

Appendix F: GDP, Population, Total Cost, and Bond Projection



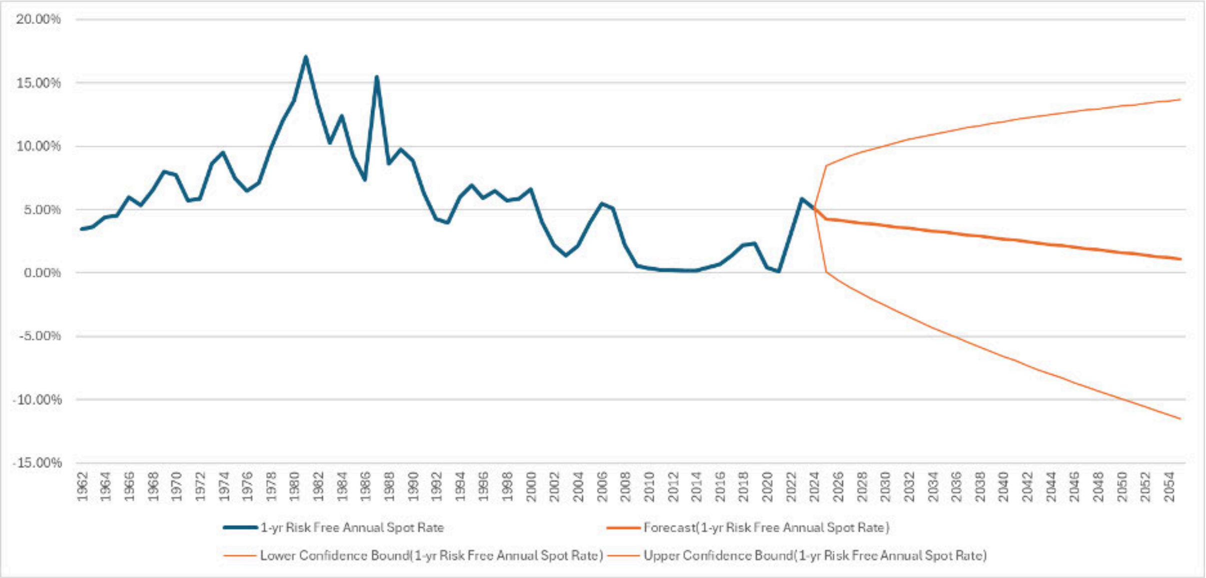




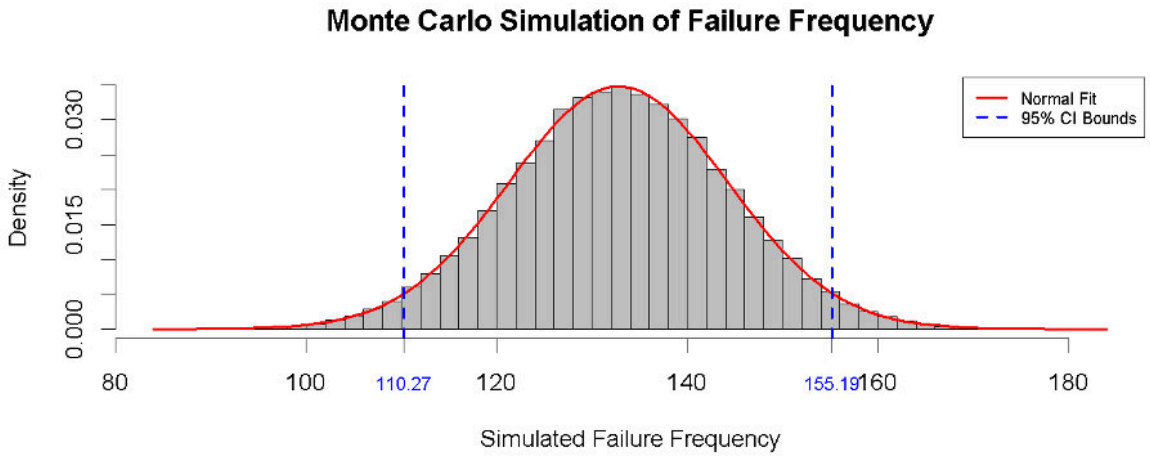
Total Economic Projection: Expected Cost by Region (2025+)

Summed projected regional costs based on SARIMA inflation estimates

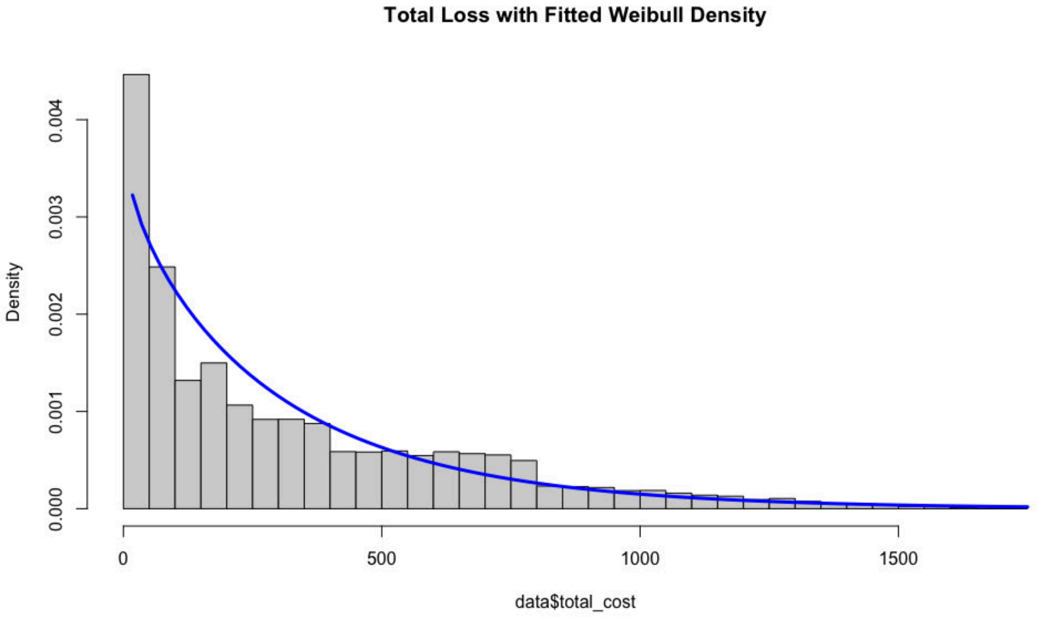
Source: SARIMA Forecast Model



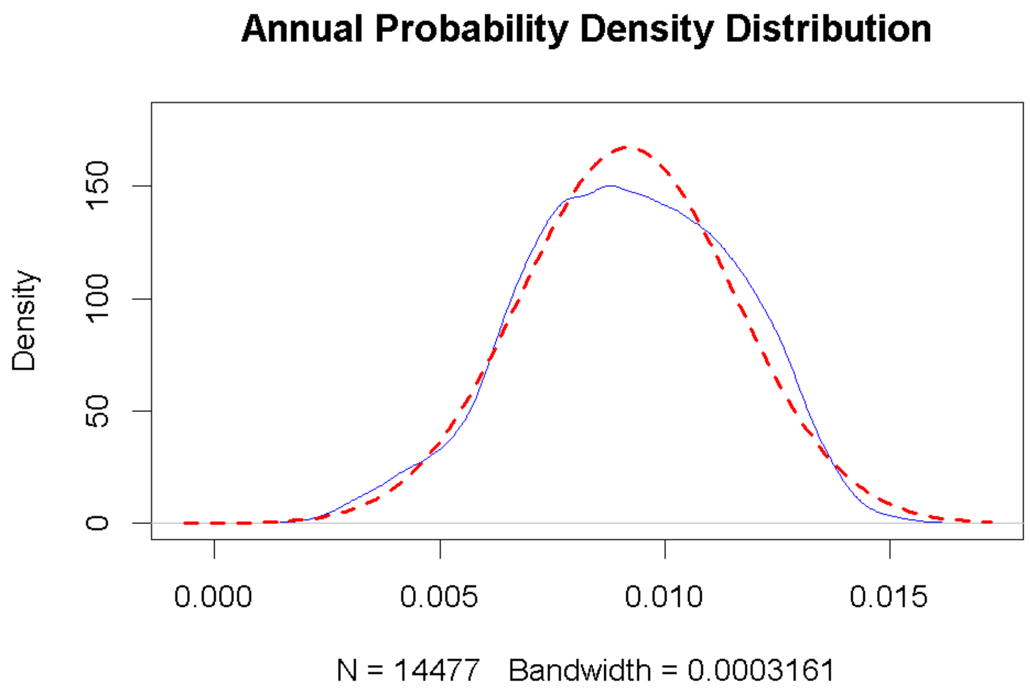
Appendix G: Monte Carlo Simulation (Frequency)



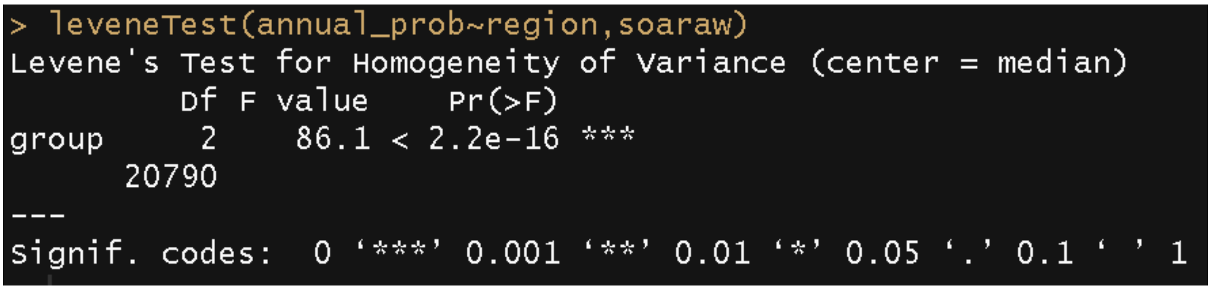
Appendix H: Fitted Distribution (Frequency and Severity)



Appendix I: Density Distribution of Annual Probability in All Regions

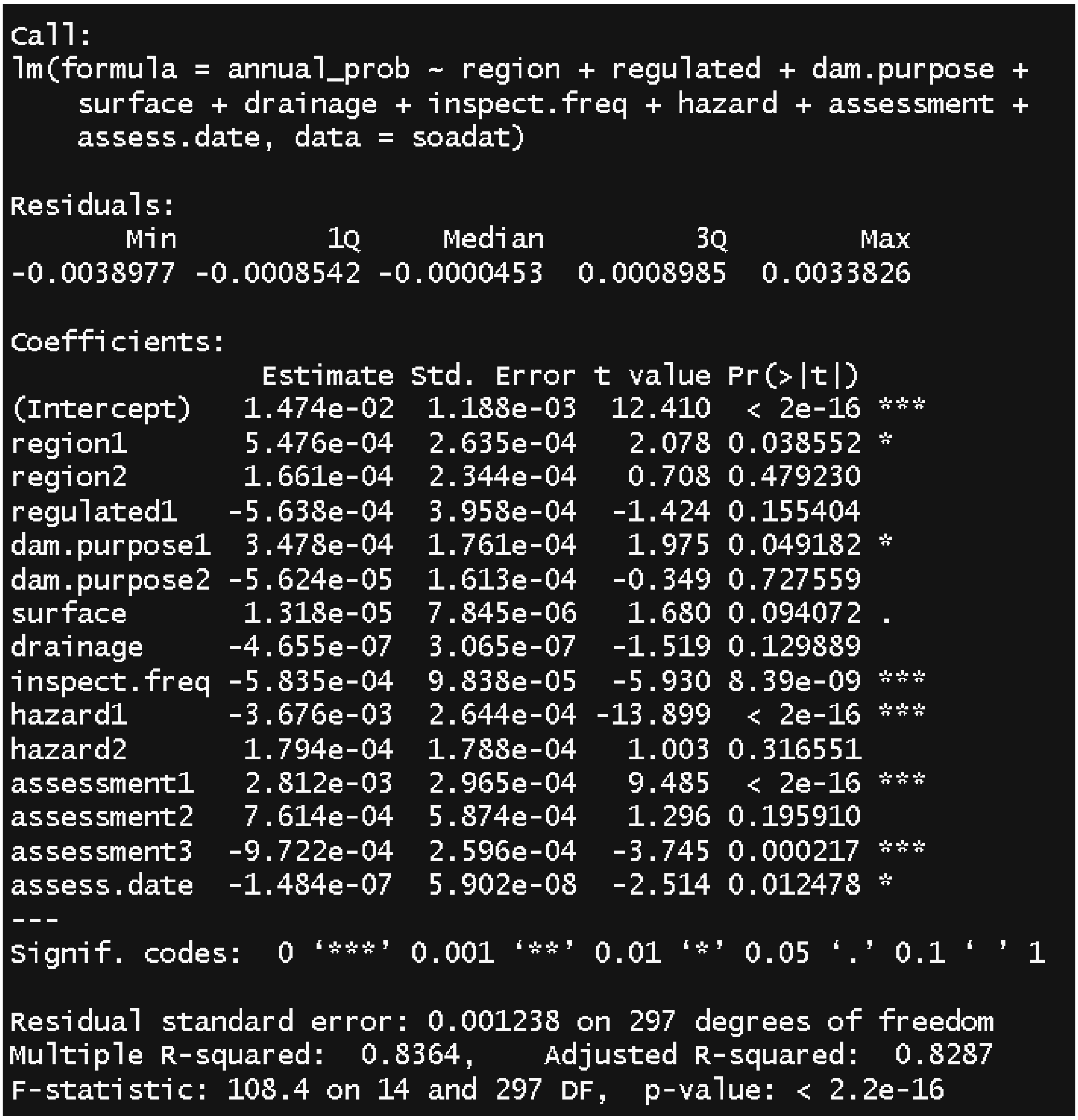


Appendix J: Levene’s Test Result



```
> leveneTest(annual_prob~region,soaraw)
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)    
group     2    86.1 < 2.2e-16 ***
      20790                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix K: Supporting Underwriting Decision from GLM model



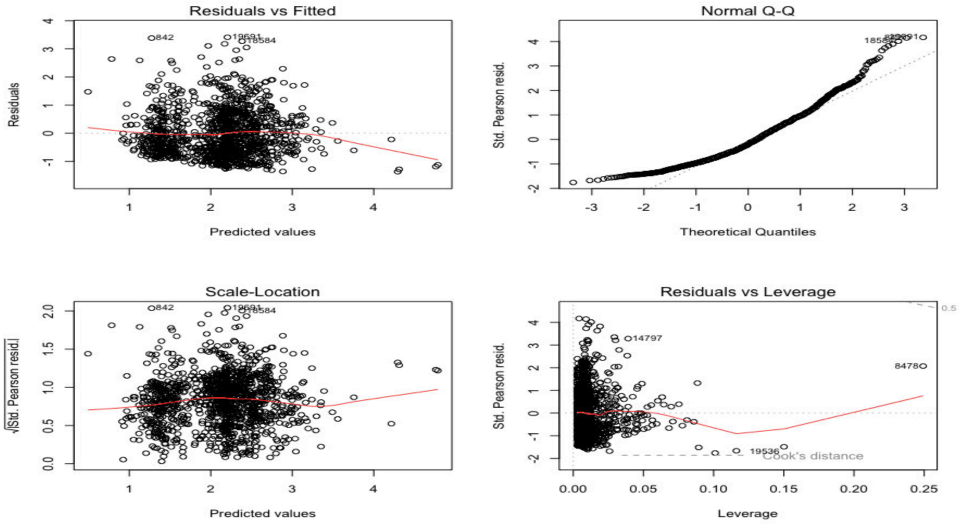
```
Call:
lm(formula = annual_prob ~ region + regulated + dam.purpose +
    surface + drainage + inspect.freq + hazard + assessment +
    assess.date, data = soadat)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0038977 -0.0008542 -0.0000453  0.0008985  0.0033826

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.474e-02  1.188e-03  12.410  < 2e-16 ***
region1       5.476e-04  2.635e-04   2.078 0.038552 *
region2       1.661e-04  2.344e-04   0.708 0.479230
regulated1   -5.638e-04  3.958e-04  -1.424 0.155404
dam.purpose1  3.478e-04  1.761e-04   1.975 0.049182 *
dam.purpose2 -5.624e-05  1.613e-04  -0.349 0.727559
surface       1.318e-05  7.845e-06   1.680 0.094072 .
drainage     -4.655e-07  3.065e-07  -1.519 0.129889
inspect.freq -5.835e-04  9.838e-05  -5.930 8.39e-09 ***
hazard1      -3.676e-03  2.644e-04 -13.899  < 2e-16 ***
hazard2       1.794e-04  1.788e-04   1.003 0.316551
assessment1   2.812e-03  2.965e-04   9.485  < 2e-16 ***
assessment2   7.614e-04  5.874e-04   1.296 0.195910
assessment3  -9.722e-04  2.596e-04  -3.745 0.000217 ***
assess.date  -1.484e-07  5.902e-08  -2.514 0.012478 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001238 on 297 degrees of freedom
Multiple R-squared:  0.8364,    Adjusted R-squared:  0.8287
F-statistic: 108.4 on 14 and 297 DF,  p-value: < 2.2e-16
```

Appendix L: Tweedie Model Diagnostics Plot



Residuals vs Fitted
Residuals
Predicted values
842
19691
18584
Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
Scale-Location
√|Std. Pearson resid.|
Predicted values
842
19691
18584
Residuals vs Leverage
Std. Pearson resid.
Leverage
14797
8478
19536
Cook's distance
0.5