Background

Earthquakes present a unique challenge in disaster prediction. Unlike cyclones or tsunamis that can be forecast with some warning, earthquakes strike with ruthless immediacy. This is where the development of Earthquake Early Warning Systems (EEWS) becomes more important. New Zealand (NZ), situated along the volatile edge where the Pacific and Australian tectonic plates meet, experiences frequent earthquakes. Although the country has seismic monitoring in place to gather earthquake data, it currently lacks a comprehensive EEWS.

Recent advancements in machine learning research worldwide offer promising approaches for intensity estimation, utilizing large datasets and sophisticated algorithms to enhance predictive accuracy. However, these models often face challenges related to station variability, affecting their generalizability and reliability. This research addresses these issues by developing a generalized earthquake intensity estimation model for New Zealand.

Objectives

1. Determine the best factors for accurately estimating S-wave intensity.
2. Explore the impact of various parameters (P-wave parameters, station characteristics, earthquake-id, year index) on S-wave intensity.
3. Develop a generalized earthquake intensity estimation model that can be used to predict upcoming S-wave intensity at any station in NZ.

Methodology

The historical earthquake dataset for NZ includes multiple records of the same event from different stations, leading to correlated data. Classical regression models are unsuitable due to data non-independence and the influence of station parameters like geographical location. Ignoring non-independence can result in inflated Type I error rates, biased standard errors, and unreliable inferences about predictor effects. Single-station data models are impractical due to insufficient data points for some stations and the complexity of building and fine-tuning models for over 290 stations. Instead, a multilevel model (MLM) is appropriate. MLM uses all available data, accounting for correlations between records from different stations, and incorporates parameters such as station, earthquake event, and time of occurrence as random effects. MLM effectively handles unbalanced data by incorporating random effects, ensuring better model fit, accurate estimations, accurate standard errors, and reliable statistical inferences.

In this research, P-wave parameters (Pa, Pv, Pd) and S-wave parameters (PGA, PGV, PGD) were established using computed seismic wave data. To identify suitable parameter for measuring S-wave intensity and select the best predictor variable among P-wave parameters, a correlation analysis was conducted between S-wave parameters and P-wave parameters, evaluating their R-squared values, RMSE, and MAPE. The necessity of MLM was assessed by checking the Intraclass Correlation Coefficient (ICC) and design effect values. The best model was selected using maximum likelihood estimation, comparing AIC, BIC, log-likelihood values, and checking significance using p-values. Finally, predictions were made using the multilevel model with prediction intervals.

MLM model expressed by,

$$y_{ij} = \alpha + \beta x_{ij} + a(\text{subject}_i) + b(\text{subject}_i)\, x_{ij} + \varepsilon_{ij}$$

where,

- y_{ij} is the response variable for the j-th observation within the i-th subject.
- α is the fixed intercept, representing the average intercept across all subjects.
- β is the fixed slope, representing the average effect of x on y across all subjects.
- a is the random intercept for the i-th subject, which is assumed to follow a normal distribution: a $\sim N(0, \sigma_\alpha^2)$.
- b is the random slope for the i-th subject, which is assumed to follow a normal distribution: b $\sim N(0, \sigma_\beta^2)$.
- x_{ij} is the predictor variable for the j-th observation within the i-th subject.
- ε_{ij} is the residual error term, assumed to follow a normal distribution: $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Results and Discussion

The initial regression analysis identified log(PGV) as the best parameter for measuring S-wave intensity, and log(Pv) as the best predictor for estimating S-wave intensity. The log(PGV) ~ log(Pv) model indicated the highest R-squared (0.7538) and the lowest AIC, BIC as well as the lowest RMSE, and MAPE values on the test set. The classical linear regression model showed statistical significance (F-statistic: 7.126e+04, p-value < 2.2e-16) and explained 75.4% of the variance in log(PGV). However, it failed to account for data non-independence and nested data structure. To address this, MLM was employed. The unconditional mean model of MLM revealed significant variance between stations (ICC = 0.459), earthquakes (ICC = 0.610), and years (ICC = 0.162), highlighting the importance of random effects and indicating data clustering.

The model with random intercepts and slopes for stations and random intercepts for earthquake-id and year-index significantly improved fit (p-value<2.2e-16), achieving higher R-squared values (conditional: 0.826, marginal: 0.718). This model was chosen as the best for the analysis, with the lowest AIC (50814.5) and BIC (50879.0). Comparing the classical regression model to the chosen MLM

showed an improvement in the R-squared value by 7%. Cross-validating the models with the test set demonstrated that the MLM significantly reduced RMSE and MAPE values compared to the ordinary regression model.

Overall, the chosen MLM provides a more robust and accurate approach for estimating log(PGV) by accounting for correlated errors and station-specific variations. By incorporating random effects for station, earthquake-id, and year-index, the model captures variations specific to each station and earthquake event, leading to more accurate estimations and credible intervals. This improved accuracy helps understand baseline variations in S-wave intensity associated with different earthquakes, stations, and years.

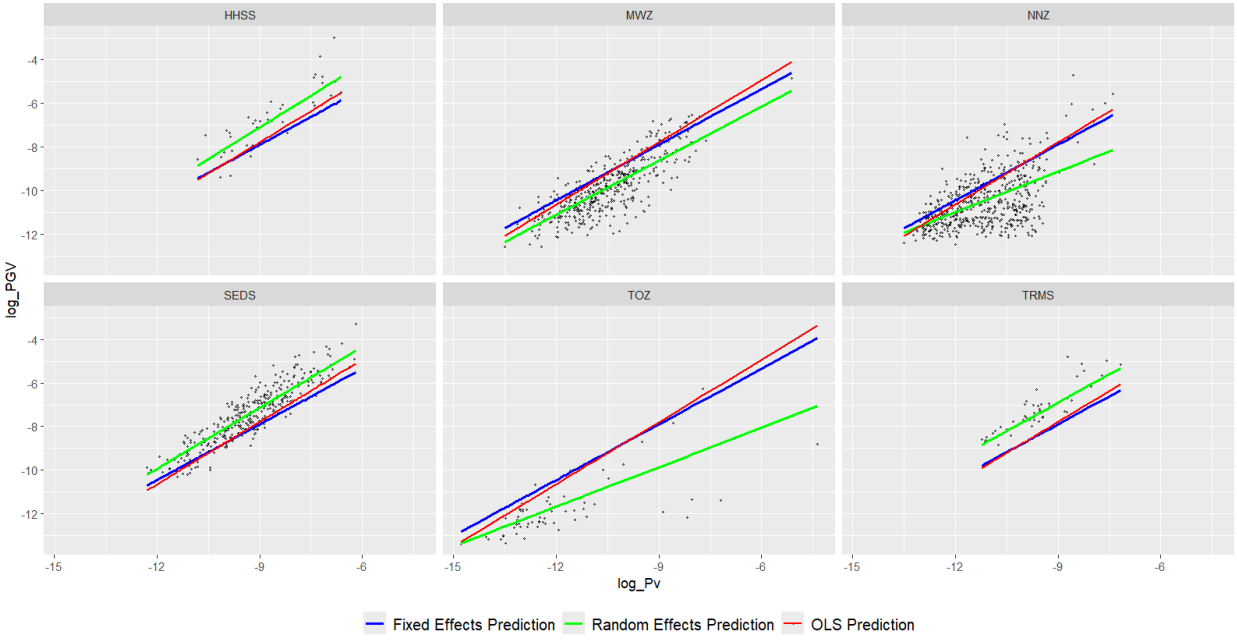

Figure 1: Comparison of log(PGV) predictions: Fixed Effects, Random Effects and OLS Prediction

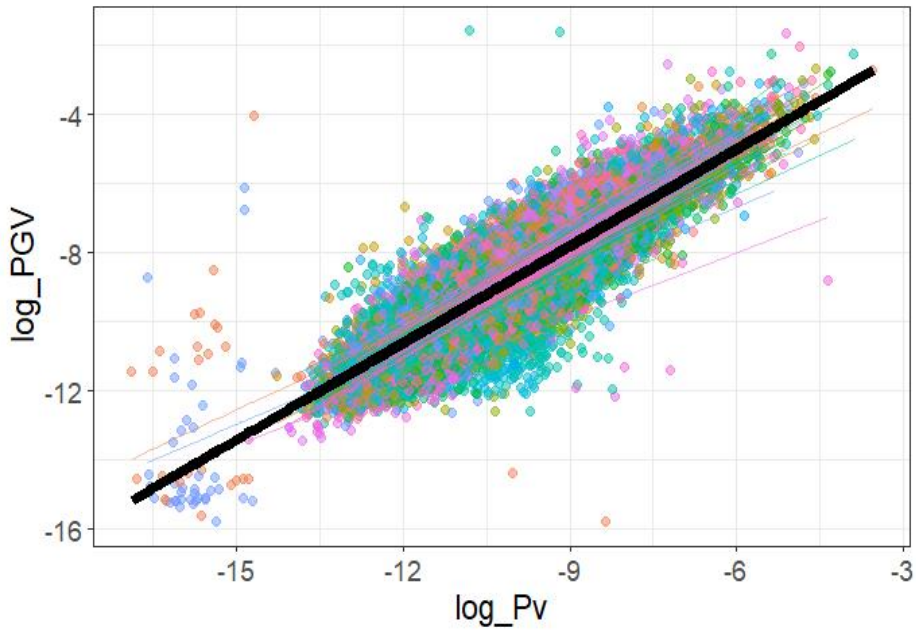

Figure 2: Scatter Plot with Independent Linear Multilevel Model Fits for all stations: Fixed and Random Effects by Station. Fixed effect (heavy black line) and random effect (thin lines).

Conclusion

This study presents a comprehensive framework for predicting S-wave intensity across all regions of New Zealand, accounting for overall trends and variations due to stations, earthquake events, and time effects. It highlights the limitations of conventional regression models in handling non-independent nested data structures and underscores the importance of geological elements in seismic assessments. The results indicate that multilevel models provide a better fit by considering these variances. This model's station-level application enhance the accuracy of ground-shaking intensity predictions, potentially improving the reliability of warning systems and public safety.