Development of a Seismic Loss Prediction Model for Residential Buildings - Christchurch, New Zealand

Samuel Roeslin[1] (s.roeslin@auckland.ac.nz), Quincy Ma[1], Pavan Chigullapally[1], Joerg Wicker[1], Liam Wotherspoon[1]
[1]Department of Civil and Environmental Engineering, University of Auckland, New Zealand

OBJECTIVES

- **To develop a machine learning using EQC's insurance building claim data from the 2010-2011 Canterbury Earthquake Sequence.**
- **To find critical features that influenced building loss during the 2010-2011 Canterbury Earthquake Sequence.**

INTRODUCTION

In 2010-2011, New Zealand experienced the most damaging earthquakes in its history, known as the Canterbury Earthquake sequence (CES). The CES led to extensive damage to Christchurch buildings, infrastructure and its surroundings, affecting commercial and residential buildings. The total economic losses of more than NZ$40 billion[1] accounted for 20% of New Zealand's GDP[2].Owing to New Zealand's particular insurance structure, the insurance sector contributed to over 80% of losses for a total of more than NZ$31 billion[1-3]. Losses from residential building accounted for 50% of the total economic losses[2]. The residential building losses were covered either partially or entirely from the NZ government backed Earthquake Commission (EQC) cover insurance scheme. Following the CES, EQC collected detailed financial loss data and building characteristics for more than 500,000 claims[2].

EQC DATA

- EQC claim data set is a wide dataset with 62 features.
- For some of the attributes (e.g. construction year, primary construction material, number of stories), more than 80% of the data points was not collected as it was not necessary for settlement purposes.

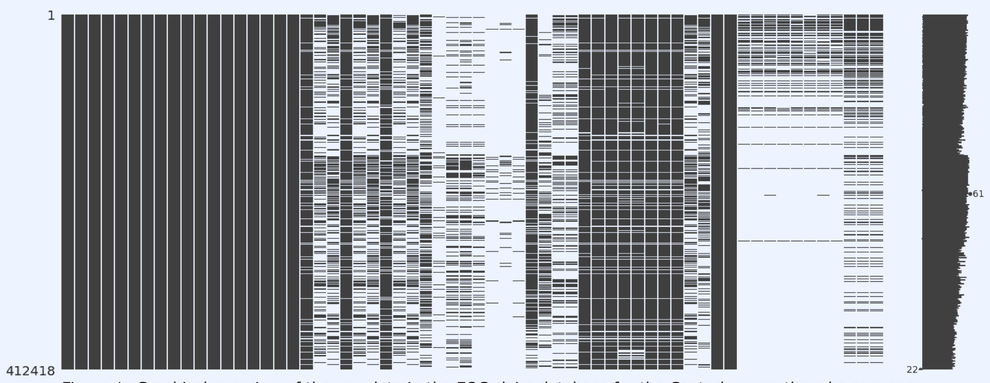

Figure 1: Graphical overview of the raw data in the EQC claim database for the Canterbury earthquake sequence. Each column is an attribute and each row is an instance.

- The first step of data cleasning consisted of selecting claims where the payment is complete.
- After the claim status selection, four earthquake events remain with enough instances to apply machine learning:
  - 4 September 2010 event
  - 22 February 2011 event
  - 13 June 2011 event
  - 23 December 2011 event
- Even if the claims are related to one event, the amounts paid or repaired may represent damage from multiple events (due to the short time between events resulting in ambiguity about which event caused the damage).

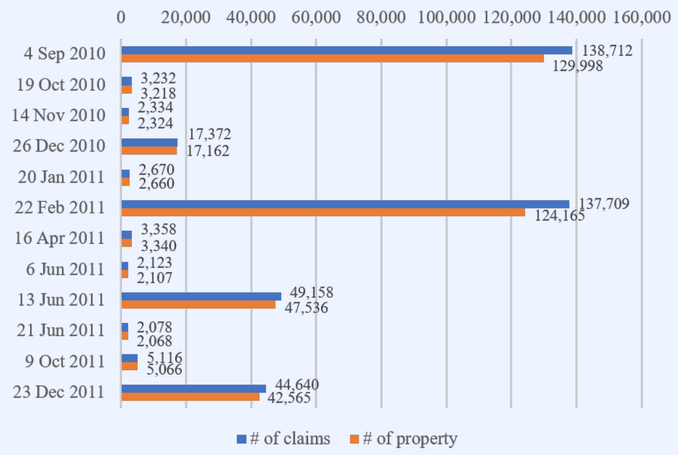

Figure 2: Number of claims and property for events in the CES with more than 1,000 instances and ClaimStatus=ClaimPaymentsComplete

DATA MERGING

- The primary database from the Earthquake Commission (EQC) contains property information and insurance claim data for residential buildings.
- This project merged additionnal information from private and open-source databases on top of EQC's claim database:
  - Building characteristics from RiskScape
  - Liquefaction occurrence from the New Zealand Geotechnical database
  - Peak Ground Acceleration (PGA) from GeoNet
  - Soil conditions Land Resource Information Systems (LRIS)
- The data integration was challenging due to the non-presence of a common feature between EQC and RiskScape. The merging was performed using the building location. Nevertheless, the merging process entailed limitations which led to the loss of instances[4].

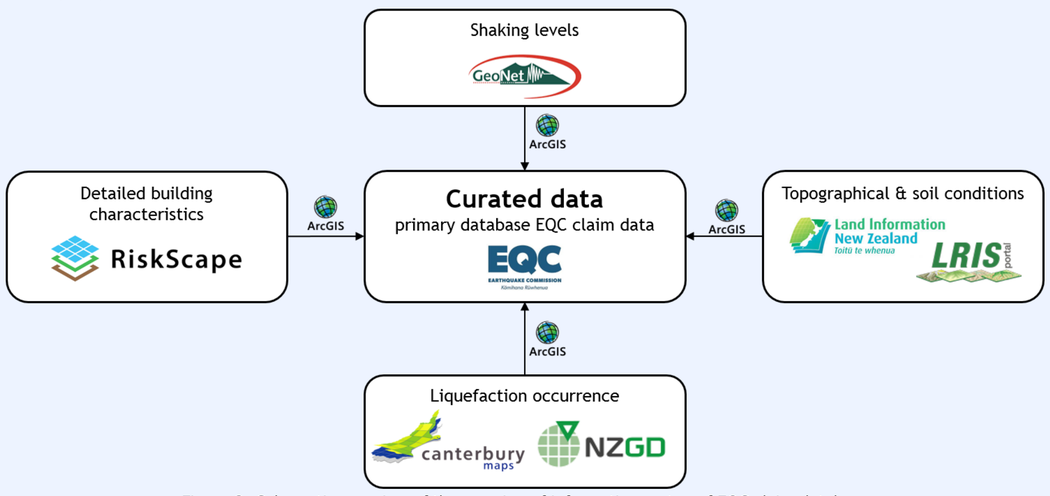

Figure 3: Schematic overview of the merging of information on top of EQC claim database

DATA PRE-PROCESSING

- In the original EQC claim data set 'Building Paid' is a numerical attribute.
- Prediction for a regression machine learning model using 'BuildingPaid' as a numerical target variable did not deliverd satisfactory outputs.
- Data pre-processing included the transformation of 'Building Paid' from a numerical attribute to categories.
- Thresholds for the cut-off were chosen according to the EQC definitions related to the cash settlement of the claim, the Canterbury Home Repair Programme, and the maximum coverage provided[5].

Figure 4: Schematic overview of the thresholds for the transformation of Building Paid from a categorical to a numerical attribute

TARGET ATTRIBUTE

- 4 September 2010:
  - 59.3% of the claims were low
  - 35.7% of the claims were medium
  - 5.0% of the claims reached the maximum cap
- 22 February 2011:
  - 44.5% of the claims were medium
  - 30.3% of the claims were low
  - 25.2% of the claims reached the maximum cap

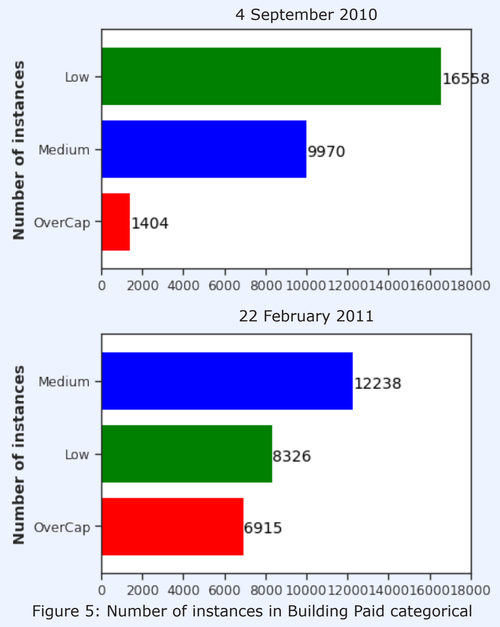

Figure 5: Number of instances in Building Paid categorical

DATA PREPARATION

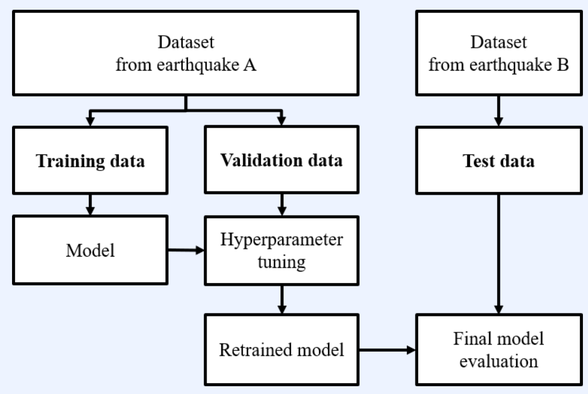

Figure 6: Overview of the use of the training, validation, and test sets

- Before starting training a machine learning model, it is necessary to split the data into distinct sets known as the training, validation (or development), and test set.
- Unlike a 'traditional approach' where the test set is held out from the same data as the training and validation set, the test set here employed comes from another earthquake.
- Testing the model using data from another earthquake in the CES (pre-processed in the same way as the training and validation set) enables to evaluate the model capacity to generalise to other events and find the model which works the best for the entire CES.

MACHINE LEARNING

- The availability of the target and observation makes this project a supervised learning problem for classification
- Four algorithms were trialled: logistic regression, decision trees, SVM and random forest
- Random forest is the best performing algorithm

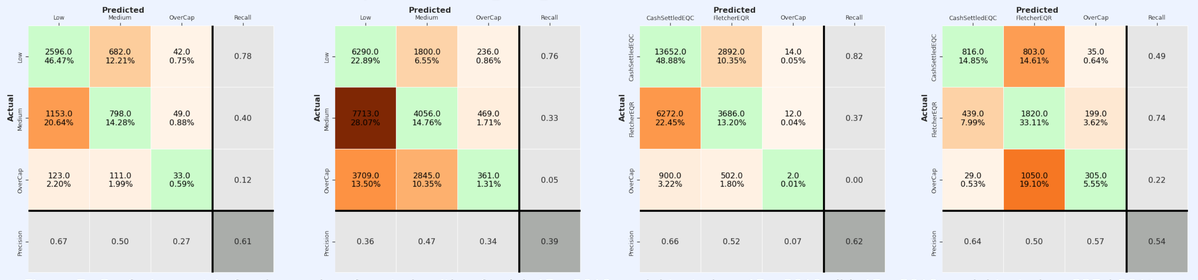

Figure 7: Confusion matrix for the random forest algorithm for (a) 4Sep2010 model tested on 4Sep2010, (b) 4Sep2010 model tested on 22Feb2011, (c) 22Feb2011 model tested on 4Sep2010, (d) 22Feb2011 model tested on 22Feb2011

INSIGHTS

- The SHapley Additive exPlanations (SHAP) post-hoc method was applied on the Random Forest models.
- PGA stands out as being the most important feature for all events.
- The liquefaction occurrence is second for 22 February 2011 model and fourth for 4 September 2010 model confirming the influence of liquefaction on the building damage/loss.
- The construction year and the floor area of the building appear in the top five most important features, however at a different position depending on the event.

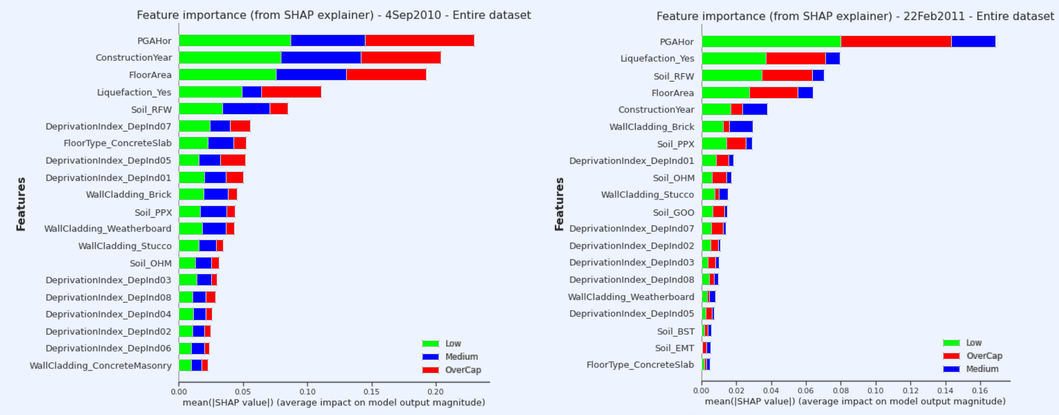

Figure 8: SHAP feature importance for (a) 4Sep2010 random forest model (b) 22Feb2011 random forest model

CURRENT CHALLENGES/FUTURE WORK

- Machine learning requires complete and clean data.
- Key building characteristics are missing in the initial EQC claim database.
- Need for a unique building identifier to facilitate the merging of information.
- Once developped, a machine learning pipeline can easily be retrained. This facilitates future studies employing different combinations of building parameters.
- Taking into account apportionment[6] would provide a more accurate allocation of loss to each event and enable to capture more details about over cap instances.
- For each event, segregating the data by geographical area where the majority of damage occured might lead to a "cleaner" train set and thus might led to more accurate predicitions.
- The introduction of additionnal parameters related to properties and social factors might deliver an improved model accuracy as well as new insights.

CONCLUSION

This poster presented the development of a seismic loss prediction model using insurance claims data from EQC collected following the 2010-2011 Canterbury Earthquake Sequence. The lack of structural building information in the original EQC dataset led to the need for the merging of building characteristics from RiskScape. Before the application of machine learning algorithms, the data was pre-processed and the target variable transformed into categories. Supervised models were trained. Random forest delivered the best accuracy. The SHAP post-hoc methodology highlighted the importance of PGA, liquefaction, year of construction, floor area of the building, and soil conditions, thus delivering key insights related to building damage.

Acknowledgement
We acknowledge the Earthquake Commission (EQC), especially the Risk Modelling Team for the help with data interpretation.

References
1. Insurance Council of New Zealand (ICNZ). (2019). Cantebury Earthquakes. https://www.icnz.org.nz/natural-disasters/canterbury-earthquakes/ [accessed 30-Nov-2020]
2. King, A., Middleton, D., Brown, C., Johnston, D., & Johal, S. (2014). Insurance: Its Role in Recovery from the 2010–2011 Canterbury Earthquake Sequence. Earthquake Spectra, 30(1), 475–491. https://doi.org/10.1193/022813EQS058M
3. Bevere, L., & Balz, G. (2012). Lessons from recent major earthquakes. Swiss Reinsurance Company Ltd, 16 p
4. Roeslin, S., Ma, Q., Chigullapally, P., Wicker, J., & Wotherspoon, L. (2020). Feature Engineering for a Seismic Loss Prediction Model Using Machine Learning, Christchurch Experience. Proceeding of the 17th World Conference on Earthquake Engineering, 17WCEE.
5. Earthquake Commission (EQC). (2019). Briefing to the Public Inquiry into the Earthquake Commission: Canterbury Home Repair Programme (Issue 24 June 2019). https://www.eqc.govt.nz/sites/public_files/documents/Inquiry/7. Canterbury Home Repair Programme Briefing rs.pdf
6. Earthquake Commission (EQC). (2018). Apportionment Factsheet. https://www.eqc.govt.nz/canterbury-earthquakes/claims-assessment/apportionment/apportionment-factsheet