Using decision trees to provide rapid estimates of earthquake loss

Bridget Ayling, Trevor Dhu, Ken Dale, Ole Nielsen

2007 AEES CONFERENCE

Summary

The goal of this study was to investigate the use of decision trees to predict financial loss associated with an earthquake, as an alternative to the Capacity Spectrum Method (CSM). Decision trees (rules) are built through regression analysis of a synthetic loss dataset (generated by EQRM and using the CSM). These rules can then be applied to other real or modelled earthquakes to approximate loss, which removes the iterative CSM step in the EQRM loss calculation, and increases its efficiency. This will allow more simulations to be conducted, hence capturing more of the uncertainty inherent in any impact modelling process.

Background

- The ability to provide rapid and accurate estimates of damage following an earthquake is central to effective disaster management.
- At Geoscience Australia, event impact analysis is achieved through an engineering-based approach, using an application known as EQRM [1,2]. EQRM generates a synthetic earthquake catalogue, models the associated ground motion and probability of occurrence, uses an attenuation model to describe the propagation of seismic waves to the locations of interest, incorporates a site-response model to account for effects of local regolith & geology, estimates the probability that a building portfolio will experience different levels of damage (using the CSM), and computes the direct financial loss using these damage probabilities and a financial loss model [1,3].
- The CSM is an iterative approach that compares the capacity of a structure (in the form of a pushover curve) with the demands on a structure (the earthquake response spectra) [4,5] and it is the most computationally-intensive step in the loss calculation.

Methods

Generating a synthetic loss dataset

The EQRM application was used to model losses at multiple sites for multiple earthquake scenarios. 200 sites were spaced evenly over 4 degrees of longitude (equal latitude), perpendicular to a vertically-dipping, N-S trending fault and each with the same building type. At each site, ground-motions were calculated at several periods within the earthquake response spectra (36 spectral periods from 0 to 3 seconds) for varying: regolith site classes (National site classes B, BC, C, CD, D, DE and E); earthquake magnitudes (4.5, 5.0, 5.5, 6.0, 6.5, 6.8, 7.0, 7.2, 7.5); and, ground motion models (Toro [6], Sadigh [7] and a preliminary Australian model (hereafter referred to as the Allen model; per comm. T. Allen)). Loss (structural damage) was calculated for each combination, resulting in 37800 loss calculations for a range of event scenarios on the same fault for the particular HAZUS building type. This process was repeated for all 56 of the HAZUS building types.

Building decision-trees using CART (Classification and Regression Tree Analysis)

CART is a software package that builds classification and regression trees for predicting categorical and continuous variables respectively. The rationale for using CART was: (a) to find which variables appear most important in determining structural damage (loss); and (b) to generate a decision tree (rules) using these variables and their values, to allow us to predict loss without the need for a full CSM analysis.

Loss (as a percentage of replacement cost) was selected as the target variable (the variable we hope to predict), and the following variables were selected as potential predictor variables: earthquake magnitude, ground motion model, regolith site class, Joyner-Boore distance, and ground motions at 36 spectral periods (between 0 – 3 seconds). A test dataset (30% of the input data) was selected at random for cross-validation of the tree.

CART analysis involves four steps: (1) tree building (2) end of tree building (3) tree 'pruning' (4) optimal tree selection. Criteria imposed on optimal tree selection were: a minimum of 5 data-points per terminal node, less than 100 terminal nodes in final decision tree, and absolute within-node variability ideally less than 5% (Figure 1). Total predictive accuracy of the decision trees was tested using an independent test dataset (generated using EQRM), and was consistently <0.5 RMS error for the whole tree.

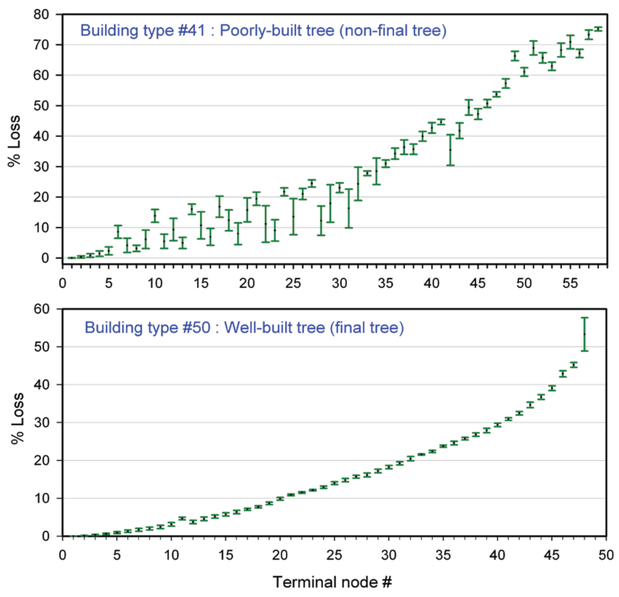


Figure 1: Distribution of % loss prediction in terminal nodes for a poorly-built decision tree (top), which displays large within-node variability and multiple nodes overlapping in target variable prediction, and (bottom): a well-built decision tree.

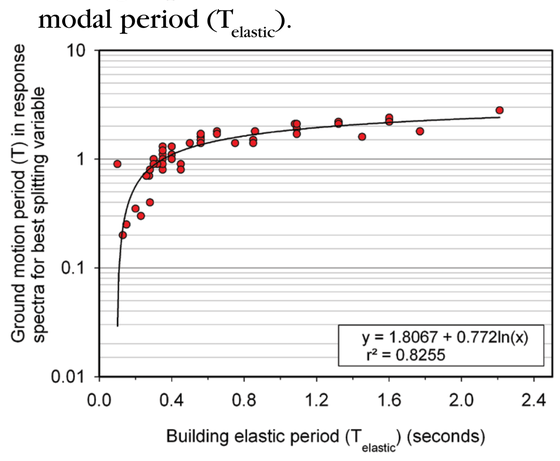
Testing decision-tree performance vs. Capacity Spectrum Method results

Another synthetic loss dataset was created that incorporated every HAZUS building type, with 140 sites per building type. The 140 site locations for each building type were randomly sampled from within a 4° radius extending from the epicentre (thus sites were different for each building type), to incorporate aleatory uncertainty. These were run under different earthquake event scenarios and ground motion models (Toro, Allen and Sadigh). This produced a dataset with more than 100,000 individual loss scenarios. These sites were also run using the Atkinson and Boore ground motion model [8] to enable the CART rules to be tested on an independent dataset. Using the site-database, the relevant rule was accessed for each building type, with each rule accessing the appropriate periods of ground-motion (predictor variables) needed for that building type to estimate loss.

Results and Discussion

Synthetic loss datasets: primary predictor variables for each building type

- Periods of ground motion were always found to be the best predictor variables (Table 1), and these are positively and logarithmically correlated with the elastic periods for each building type ($T_{elastic}$) (Figure 2). This conforms to the belief that buildings should experience the most damage when subjected to ground shaking at periods near their natural/modal period ($T_{elastic}$).



◀ **Figure 2:** Logarithmic relationship between building modal period and best predictor variable.

Building type	Top splitter*	Building type	Top splitter*	Building type	Top splitter*
1	0.35	20	1.6	39	0.8
2	1.1	21	1.9	40	0.9
3	1.4	22	1.1	41	0.7
4	2.1	23	1.5	42	0.8
5	2.8	24	1.9	43	0.9
6	1.3	25	0.8	44	0.8
7	1.8	26	1.2	45	1.4
8	1.8	27	1.7	46	1.5
9	1.1	28	2.1	47	1.4
10	1.3	29	1.0	48	2.3
11	1.8	30	1.4	49	2.4
12	2.2	31	1.0	50	2.2
13	1.3	32	1.4	51	0.25
14	1.7	33	1.7	52	0.25
15	2.1	34	0.9	53	0.2
16	1	35	1.0	54	0.4
17	1.4	36	0.9	55	0.4
18	1.6	37	0.7	56	0.3
19	1.1	38	0.9		

**Period (T) in earthquake response spectra*

▲ **Table 1:** Highest ranked predictor variable for each Hazus building type (established through regression tree analysis in CART).

Capacity Spectrum Method vs. decision-tree approach

- For higher earthquake magnitudes (eg: Mag 7.5), results obtained using the decision-tree approach closely match those generated using the CSM (Figure 3). For lower earthquake magnitudes (eg: Mag 5.5), there is greater discrepancy between the two approaches (Figure 4). This suggests that the rules do not adequately characterise the response of some building types for low levels of ground motion. This may reflect a need for different-sized decision trees (larger), or that more complicated relationships exist between predicted loss and the top splitting variable for some building types.

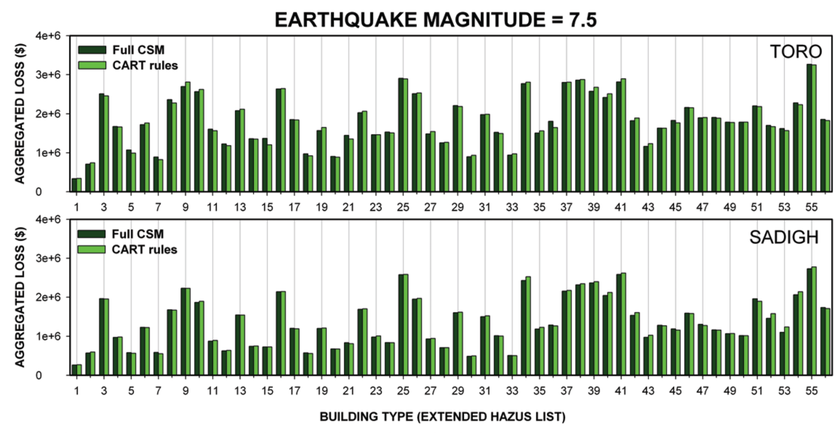


Figure 3: Aggregated building losses for each building type for each method (decision tree rules and full CSM) for a magnitude 7.5 earthquake, and modelled using the Toro and Sadigh ground motion models.

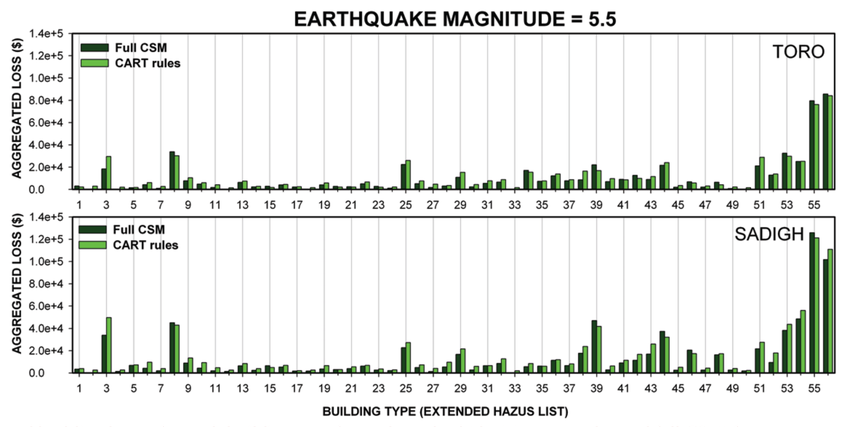


Figure 4: Aggregated building losses for each building type for each method (decision tree rules and full CSM) for a magnitude 5.5 earthquake, and modelled using the Toro and Sadigh ground motion models.

- Comparison of total absolute losses predicted using decision trees and the CSM illustrates that agreement between the two approaches improves as earthquake magnitude increases (Figure 5).
- When applied to ground motions calculated using the Atkinson and Boore ground motion model, the CART rules perform poorly for most earthquake magnitudes (Figure 5). Thus it appears CART rules are specific to the variables and their range of values used to create a synthetic loss dataset (the Atkinson and Boore model was not used in the rule generation).

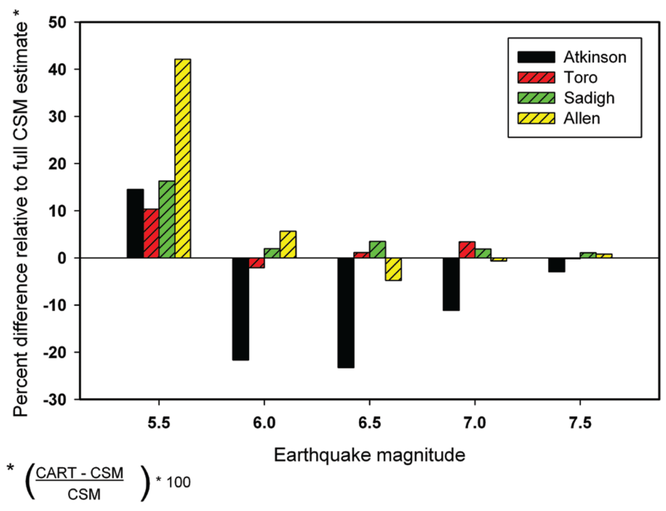


Figure 5: Percentage difference in total aggregated losses between the CART and full-CSM approaches, for different earthquake magnitudes and ground motion models.

- To produce more 'generally applicable' decision-tree rules (likely to be associated with a decrease in predictive accuracy), a larger synthetic test dataset should be used in the CART analysis. Conversely, creating a unique rule set for each ground motion model should be associated with increased predictive accuracy.

Conclusions

- CART is a powerful and useful software tool for revealing complex relationships between variables in multi-variate datasets.
- The best splitter variables (i.e. those that are most closely related to structural damage) determined by CART were periods of ground motion in the earthquake response spectra (eg: response spectral acceleration at T = 0.5 seconds), rather than variables such as Joyner-Boore distance, site-class (soil type), or earthquake magnitude.
- The decision-tree approach is able to produce loss estimates within 6% of the loss estimate produced using the full CSM in EQRM. The rules perform best (i.e. are able to predict loss) when applied to ground motions that are calculated using the ground motion models originally used to develop the rules.
- The decision-tree approach for generating loss estimates is computationally more efficient than the full CSM, and by accepting the trade-off of a small decrease in the accuracy of loss estimates (for earthquake magnitudes greater than 6), it allows modelling of significantly more earthquakes, thus should produce more rigorous estimates of earthquake risk and event impact.

References

[1] Robinson, D., Fulford, G., and Dhu, T., (2005) EQRM: Geoscience Australia's Earthquake Risk Model. Technical Manual Version 3.0. Geoscience Australia Record 2005/01.
[2] http://sourceforge.net/projects/eqrm
[3] Patchett, A.M., Robinson, D., Dhu, T. and Sanabria, A., (2005) Investigating Earthquake Risk Models and Uncertainty in Probabilistic Seismic Risk Analyses. Geoscience Australia, Record 2005/02.
[4] Freeman, S.A., (2004) Review of the development of the capacity spectrum method. ISET Journal of Earthquake Technology, Paper No. 438, Vol. 41, No. 1, pp1-13.
[5] Kircher, C.A., Nassar, A.A., Kustu, O. and Holmes, W.T., (1997) Development of Building Damage Functions for Earthquake Loss Estimation. Earthquake Spectra, Vol 13(4), pp663-682
[6] Toro, G.R., Abrahamson, N.A., and Schneider, J.F., (1997) Model of strong ground motions from earthquakes in Central and Eastern North America: Best estimates and certainties. Seismological Research Letters, Vol 68(1), 41-57.
[7] Sadigh, K., Chang, J.A., Egan, J.A., Makdisi, F. and Youngs, R.R., (1997) Attenuation relationships for shallow crustal earthquakes based on California strong motion data. Seismological Research Letters, Vol.68(1), pp180-189.
[8] Atkinson, G.M. and Boore, D.M., (1997) Some comparisons between recent ground- motion relations. Seismological Research Letters, Vol. 66(1), pp24-40.

For further information contact: Bridget Ayling
Ph: 02 6249 9415 Email: bridget.ayling@ga.gov.au
www.ga.gov.au