Earthquake Damage Prediction of Buildings in Nepal using Machine Learning tools.

Subash Ghimire*, Philippe Gueguen**,** Danijel Schorlemmer

Abstract

Decision-makers and stakeholders need rapid assessment of the potential damage following earthquake events to develop and execute disaster risk reduction strategies and to systematically react in post-disaster situations. Classical risk assessment methods are resource- and time-consuming. In this study, the Mw 7.8 Gorkha, 2015 Nepal Earthquake crowd-sourced building damage data is used to explore the efficiency of various machine-learning techniques in rapid earthquake-induced building damage assessment. The Random Forest Regressor showed the best performance among several machine learning methods considered in this study. For rapid seismic damage assessment in Nepal, for a given earthquake scenario, the building features data collected from the existing built-up environment can be used as an input to this model and the output will help decision-makers to take an appropriate decision.

Key words: *Machine learning, seismic risk assessment, building damage portfolios, earthquake damage*

1. Introduction

It is crucial for decision-makers and stakeholders to have rapid assessments of potential damage due to earthquake events (Bommer and Crowley, 2006). For successful emergency response planning before and after an earthquake, the spatial distribution of damage over the built environment is required (Ranf et al. 2007; Earle et al., 2009). Various classical methods exist for estimating earthquake-induced building damage based on ground shaking. These methods require a lot of information on building portfolios and earthquake ground motion. Generally, the collection of such information and damage assessment is costly and time-consuming. For the last decade, the progress in artificial intelligence (AI) tools and their application in various domains has increased. Yet, there is only a very limited number of applications of AI for rapid seismic risk assessment. Riedel et al. (2016, 2018) showed the ability of the Support Vector Machine for seismic vulnerability assessment at urban or regional scales. Mangalathu et al. (2019) showed an application of the machine learning technique in rapid seismic risk assessment using an earthquake damage data portfolio of the 2014 South Napa earthquake. They concluded that the use of the rapidly growing machine-learning technique in the field of rapid seismic risk assessment provides a reliable estimate of the earthquake-induced potential building damage.

**Subash Ghimire, PhD scholar, University of Grenoble Alpes, France, subash.ghimire@univ-grenoble-alpes.fr*
Philippe Gueguen, Institut des Sciences de la Terre, philippe.gueguen@univ-grenoble-alpes.fr
DanijelSchorlemmer, GFZ German Research Centre for Geosciences, Potsdam, Germany, ds@gfz-potsdam.de

Moreover, building damage portfolios of earthquake events are starting to become openly accessible. For example, Nepal suffered on 25 April 2015 from a devastating 7.6 magnitude earthquake in 2015 (named hereafter as 2015 Nepal earthquake) with the epicentre about 80km northwest of Kathmandu and a rupture length of 120 km towards east. Thousands of buildings were damaged, resulting in around 9000 casualties. The Nepal government carried out a massive household survey to map the damage in the eleven mostly affected districts. The household data survey is published as open access through the official website of the National Planning commission of Nepal (http://eq2015.npc.gov.np/). This article presents the results on the performance of various machine-learning models in rapid damage earthquake assessment using the Nepal earthquake damage portfolio.

2. Description of the Damage Database

The 2015 Nepal earthquake building-damage database consists of 762,106 building datasets collected in eleven districts of Nepal (Fig. 1). Information about each building features: number of stories, age of the building, height, plinth area, construction material, ground slope condition, building position with respect to another building, and roof type. In addition to the database, in each district, the ground motion is computed using the ShakeMap tool from the United States Geological Survey USGS. Macroseismic intensity is considered as an input ground motion in this study. Fig. 2 and Fig. 3 show the distribution of intensity of the 2015 Nepal earthquake and different building features and damage grades in the dataset, respectively. The severity of damage is grouped into five grades observed by visual inspection. The detailed description of these five grades is available on the same website (http://eq2015.npc.gov.np/docs/#/faqs/faqs).

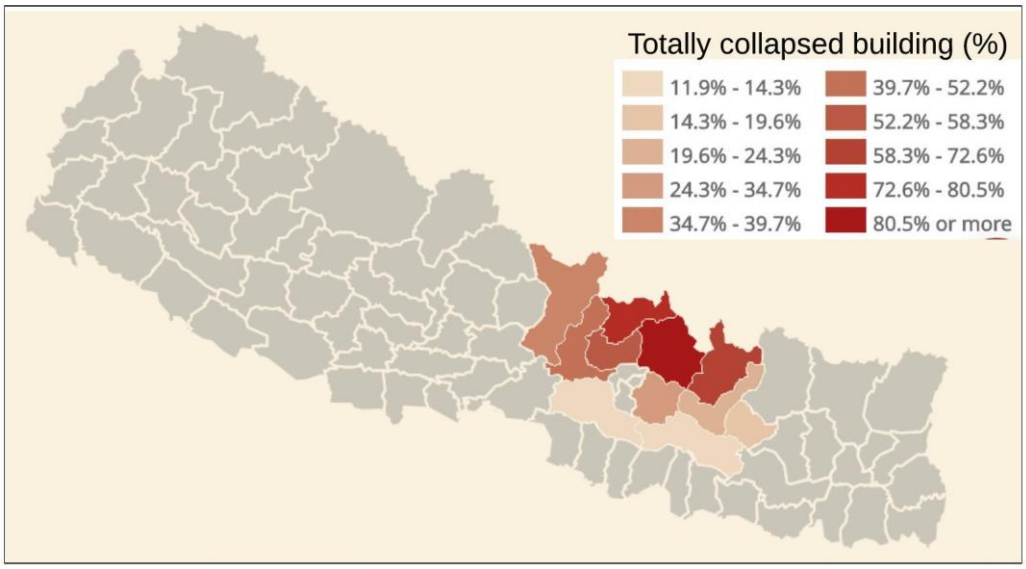

Figure 1. Location of 11 districts where the 2015 Nepal earthquake building damage data are available. It also illustrates the severity of the earthquake effect in each district in terms of the collapsed buildings. (Source: http://eq2015.npc.gov.np/#/compare).

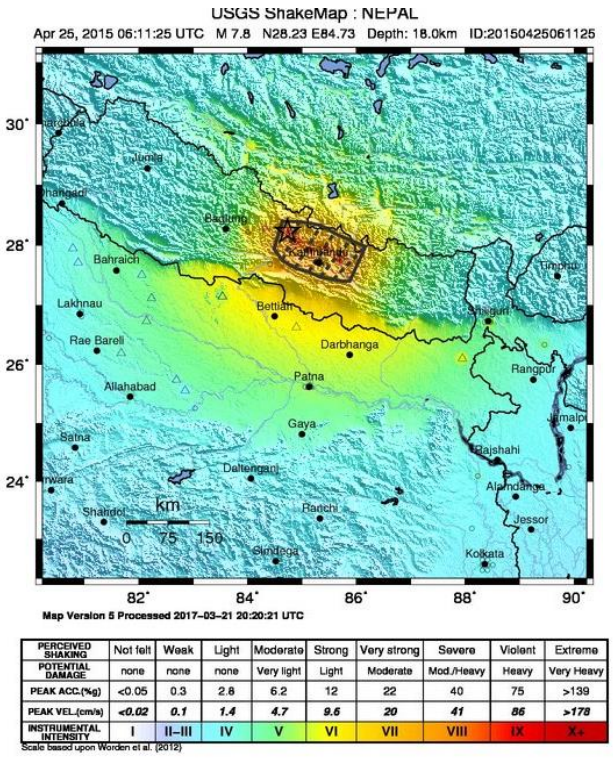

Figure 2. Spatial distribution of 2015 Nepal earthquake ground motion intensity. (Source: https://earthquake.usgs.gov/earthquakes/eventpage/us20002926/shakemap/intensity)

3. Method

This study assessed the efficiency of Linear Regression (LR), Support Vector Regressor (SVR), Gradient Boosting Regression (GBR), Random Forest Regression (RFR), Gradient Boosting Classification (GBC) and Random Forest Classification (GBC) in damage prediction. A brief description of these methods is provided in the annex. Interested readers are suggested to refer to Friedman et al. (2001) and scikit-learn machine learning in Python (https://scikitlearn.org/stable/) for detailed information on these machine-learning methods. For each machine-learning model, the features of buildings (number of stories, height, age, plinth area, ground slope condition, position, roof material, construction material), as well as the intensity of ground motion, are defined as input features and damage grades are used as response variables. About 0.48% of the dataset was observed with missing/outlier values. The missing data points associated with damage grades were removed and the outliers associated with the number of stories, age, the height of buildings were replaced by their respective mean value. The entire dataset is randomly divided into training and testing subsets. Following the recommendation of Friedman et al. (2001), 70% of the data are used as a training set and 30% are used as testing set. The training set is used to train the machine learning model and the testing set is used to observe the predictive performance of the machine learning model. The machine learning model is trained on the previously defined features. The performance of each machine learning model is evaluated through the coefficient of determination (R^2 scores) and Root Mean Square Error (RMSE) scores for

regression and accuracy scores for classification problems. Higher the value of R^2 and lower the RMSE value, better is the performance of the model.

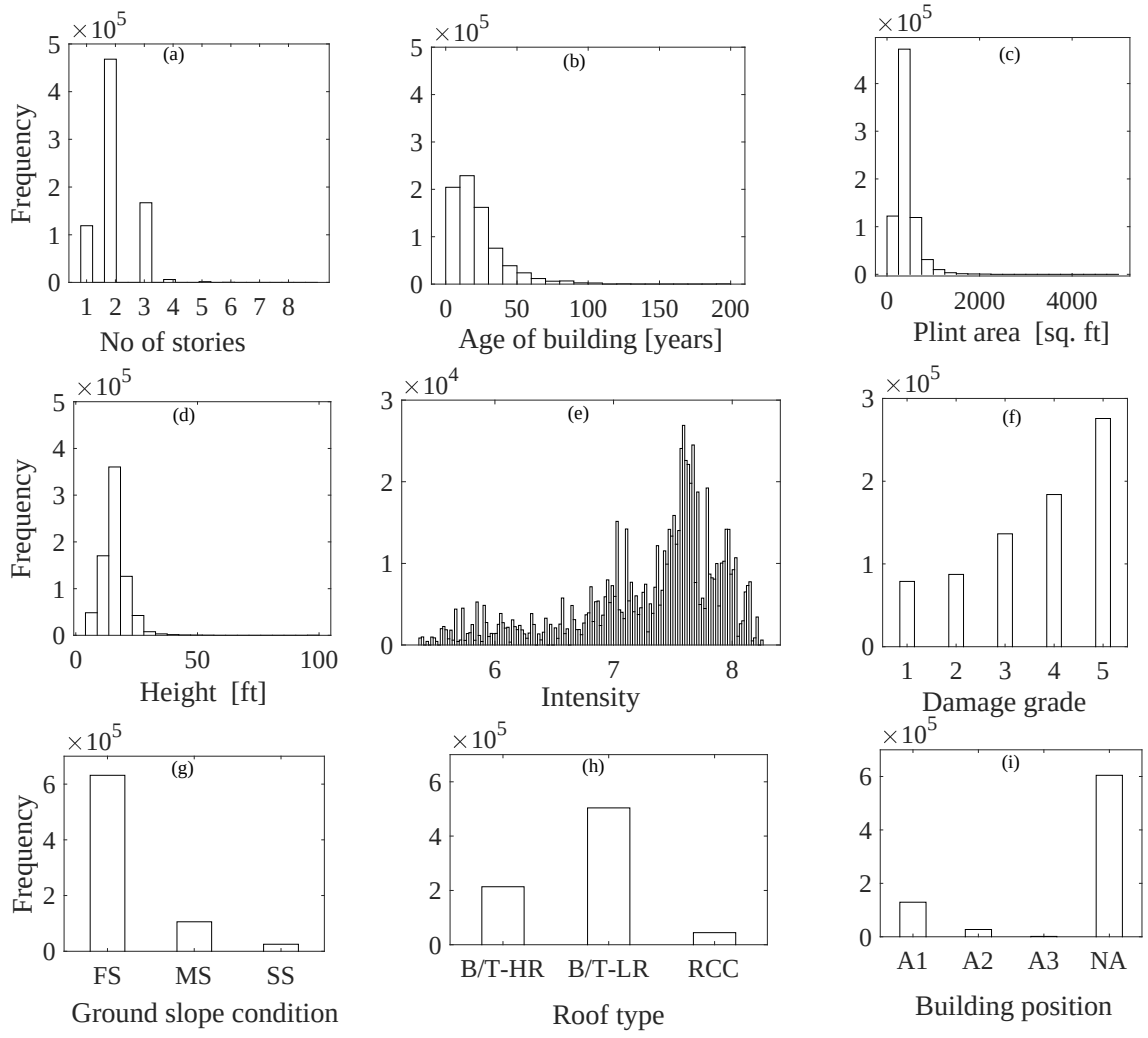

Figure 3. Distribution of different features in the dataset. The y-axis is the frequency and the x-axis in frame is (a) number of storey, (b) age of the building, (c) plinth area of building, (d) height of the building (e) Macroseismcic intensity, (f) damage grade, (g) ground slope condition at building location (h) type of construction material used in roof, and (i) position of building with respect to another building. In frame (g) FS/MS/SS represents flat/mild/steep slope, respectively. In frame (h) B/T-HR, B/T-LR, represents bamboo/timber-heavy-roof, bamboo/timber- light-roof and RCC represents reinforced cement concrete. In frame (i) A1/A2/A3 and NA represents attached with one/two/three sides and not attached, respectively.

4. Results and Discussion

The LR and SVR are observed to have the values of R^2 score equal to 0.41 and 0.38 and RMSE score equal to 1.06 and 1.08, respectively. The lower R^2 value and the higher RMSE value for LR and SVR methods prove less suitable for this dataset. They oversimplified the

complex non-linear interaction present in the dataset. Similarly, the GBC and RFC methods are observed to have an accuracy score of 0.33 and 0.55, respectively. GBC and RFC are also unable to classify the true damage grade with high accuracy. The higher values of R^2 score are 0.58 and 0.56, and the lower RMSE values are 0.88 and 0.87 are observed for GBR and RFR, respectively. These methods give higher efficiency in the damage prediction. GBR and RFR can reproduce the stronger non-linear interaction that exists among different features present in the dataset.

The performance, effectiveness, and computational time of these methods are very sensitive to the value of model parameters (hyperparameters). The GBR method requires careful tuning of a greater number of hyperparameters as compared to RFR. Fig. 4 shows the results of the RFR method in the test dataset. The high score associated along with the diagonal element of the confusion matrix and the location of the median value of the predicted damage grade (Fig. 4) proves that the RFR method is the most efficient method applied to our dataset. This study shows the applicability of the RFR method with the 2015 Nepal earthquake building-damage portfolio in rapid seismic risk assessment in Nepal i.e. using the RFR model trained on the 2015 Nepal earthquake building-damage dataset, we can predict potential damage for a given earthquake scenario by considering the same input features data collected from the existing built-up environment.

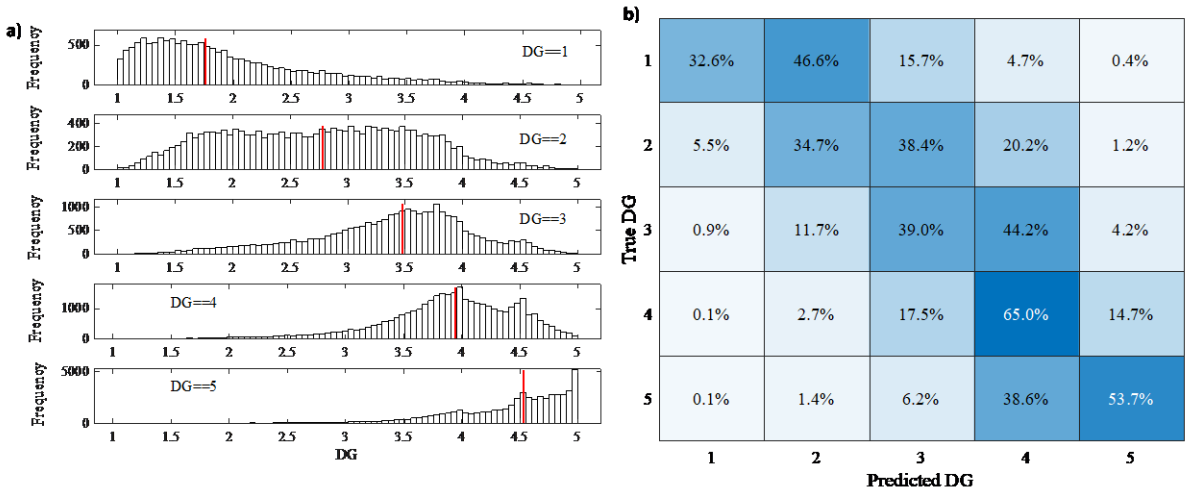

Figure 4. Graphical representation of the predictive performance of the RFR model on the test dataset. In frame (a) the x-axis is the predicated damage grade (DG) and the y-axis is the frequency. The red vertical line represents the median value. The true damage grade is noted in the same subplot. In frame (b) the x-axis is the predicted DG and the y-axis is the true DG.

5. Conclusion

The suitability of machine learning techniques in rapid seismic risk assessment is studied using the 2015 earthquake building damage data from Nepal. Performance of Linear Regression, Support Vector Regression, Gradient Boosting Regression, Random Forest Regression, Gradient Boosting Classification, and Random Forest Classification was tested. The Random Forest Regression is observed to be the most efficient in damage prediction.

This model may assist stakeholders and decision-makers in rapid seismic risk assessment in order to formulate and implement new plans and policies in earthquake disaster risk reduction. Further investigation should be carried out for a better understanding of the applicability of the machine learning model in earthquake-induced rapid building damage prediction based on the need and interests of the decision-makers and stakeholders. As a future perspective, further investigation in rapid seismic risk assessment should be carried out by considering the key building features that are easily accessible and could be used as a good proxy to predict building damage using the most suitable machine learning technique. Investigation of the applicability of the machine learning model with other open-data platforms like OpenStreetMap (OSM) should be investigated for rapid seismic risk assessment.

6. Acknowledgement

This work is part of the URBASIS program led by P.G at ISTerre/Université de Grenoble Alpes. We thank LabEx OSUG@2020 (Investissements d'avenir-ANR10LABX56 and the ITN-MSCA URBASIS project, a project funded by the EU Horizon 2020 program under Grant Agreement Number 813137. Part of this work was supported by the Real-time earthquake rIsk reduction for a reSilient Europe (RISE) project, funded by the EU Horizon 2020 program under Grant Agreement Number 821115. S.G. would like to thank Kathmandu Living Labs for their most valuable assistance.

7. References

Bommer, J. J., & Crowley, H. (2006). The influence of ground-motion variability in earthquake loss modelling. *Bulletin of Earthquake Engineering*, *4*(3), 231-248.

Earle, P. S., Wald, D. J., Jaiswal, K. S., Allen, T. I., Hearne, M. G., Marano, K. D., ... & Fee, J. M. (2009). Prompt Assessment of Global Earthquakes for Response (PAGER): A system for rapidly determining the impact of earthquakes worldwide. *US Geological Survey Open-File Report*, *1131*, 15.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Guettiche, A., Guéguen, P., & Mimoune, M. (2017). Seismic vulnerability assessment using association rule learning: application to the city of Constantine, Algeria. *Natural hazards*, *86*(3), 1223-1245.

Mangalathu, S., Sun, H., Nweke, C. C., Yi, Z., & Burton, H. V. (2020). Classifying earthquake damage to buildings using machine learning. *Earthquake Spectra*, *36*(1), 183-208.

OECD (2018), Financial Management of Earthquake Risk, *www.oecd.org/finance/Financial-Management-of-Earthquake-Risk.htm*.

Preciado, A., Ramírez-Gaytán, A., Salido-Ruiz, R., Caro-Becerra, J. L., & Lujan-Godinez, R. (2015). Earthquake risk assessment methods of unreinforced masonry structures: Hazard and vulnerability.

Riedel, I., Gueguen, P., Dunand, F., & Cottaz, S. (2014). Macroscale vulnerability assessment of cities using association rule learning. *Seismological Research Letters*, *85*(2), 295-305.

Riedel, I., Guéguen, P., Dalla Mura, M., Pathier, E., Leduc, T., & Chanussot, J. (2015). Seismic vulnerability assessment of urban environments in moderate-to-low seismic hazard regions using association rule learning and support vector machine methods. *Natural hazards*, *76*(2), 1111-1141.

Riedel, I., & Guéguen, P. (2018). Modeling of damage-related earthquake losses in a moderate seismic-prone country and cost–benefit evaluation of retrofit investments: application to France. *Natural hazards*, *90*(2), 639-662.

Ranf, R. T., Eberhard, M. O., & Malone, S. (2007). Post-earthquake prioritization of bridge inspections. *Earthquake Spectra*, *23*(1), 131-146.

Šipoš, T. K., & Hadzima-Nyarko, M. (2017). Rapid seismic risk assessment. *International journal of disaster risk reduction*, *24*, 348-360.

Wieland, M., Pittore, M., Parolai, S., Zschau, J., Moldobekov, B., & Begaliev, U. (2012). Estimating building inventory for rapid seismic vulnerability assessment: Towards an integrated approach based on multi-source imaging. *Soil Dynamics and Earthquake Engineering*, 36, 70-83

Annex

Linear Regressor

Linear Regression (LR) explains the relationship between target variables through a linear combination of input (predictors) variables. The functional form of the LR is given below as:

$$Y = \sum_{i=0}^{n} w_i x_i = w^{\mathrm{T}} x$$

Here, the weight w_0 represents the y-axis intercept and w_i is the weight coefficient of the input variable, and Y is the target variable. The LR fits a linear model with coefficients $w = (w_1, \ldots, w_p)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. The LR has simple analytical and computational properties. They provide an adequate interpretable description of how the input affects the output. This method is computationally efficient. The weight associated with each input variable helps in features importance identification. The LR is oversimplified (unable to capture the complexity of the problem), and is very sensitive to outliers. The LR assume that data are linearly separable, special attention should be paid with multicollinearity issues, not very efficient to nonlinear data (https://scikit-learn.org/stable/modules/linear_model.html).

Support Vector Regressor

Support vector machines (SVM) is a set of supervised learning methods used for classification, regression, and outlier detection. In SVM, the input features are transformed into a higher-dimensional space where two classes can be linearly separated by a high dimensional space called a hyperplane. The SVM was originally used for classification problems and then extend to regression problems called Support Vector Regression (SVR). SVR maintains all features of SVM. The model produced by SVR depends only on the subsets of the training dataset because the cost function ignores samples whose prediction is close to their target. Three types of implementation are possible for SVR: SVR, Nu-SVR, and Linear SVR. SVM is effective in high dimensional spaces, memory efficient, versatility in kernel functions. This method is more suitable when the number of features in more than the number of data. SVM is less suitable when the number of data points is so large, they do not provide direct probability estimate, overfitting could be an issue when the number of features is larger than the of data points (https://scikit-learn.org/stable/modules/svm.html).

Gradient Boosting

Gradient Boosting (GB) is a generalization of boosting to the arbitrary differentiable loss function. The GB is based on an ensemble of several decision trees. A decision tree represents a set of conditions or restrictions that are hierarchically organized and successively applied from a root to a lead of the tree. The GB is an accurate and effective procedure that can be used for both regression and classification. It is shown that both the approximation accuracy and execution speed of the GB can be substantially improved by incorporating randomization into the procedure. Specifically, at each iteration, a subsample of the training data is drawn at random (without replacement) from the full training data set. This randomly selected subsample is then used in place of the full sample to the base learner and compute the model update for the current iteration. This randomized approach also increases robustness against the overcapacity of the base learner. The GB has lots of flexibility in terms of the loss function. They can easily handle missing data, often works great with categorical and

numerical data. This is sometimes computationally expensive, requires careful tuning of hyperparameters (model input parameters). (https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting).

Random Forest

Random Forest (RF) ensemble the performance of several decision trees to classify or predict the value of variables, which is based on bagging and boosting. Bagging and boosting are two most known and used methods for the classification of the trees. Decision trees are trained by using a random subset of the original features. The RF can model complex relationships in the data and account for non-linear relationships between predictor and response variables by the adaptive nature of the decision rules. The RF has better generalization performance, less sensitive to outliers, does not require tuning of many hyperparameters. It works with continuous and also categorical predictors and also can handle missing data (https://scikit-learn.org/stable/modules/ensemble.html).