Data-Driven Earthquake Multi-Impact Modelling: A Comparison of Approaches

Hamish Patten[1,2,*], Max Anderson Loake[1], and David Steinsaltz[1]

[1]University of Oxford, Department of Statistics, Oxford, OX1 3LB, United Kingdom
[2]International Federation of Red Cross and Red Crescent Societies (IFRC), Geneva, CH-1211, Switzerland
[*]hamish.patten@ifrc.org

ABSTRACT

Natural and people-provoked hazards such as earthquakes have the potential to cause widespread devastation, loss of life, and significant economic damage. Within the first few days of the event, it is vital for disaster management organisations to utilise effective methods to prepare and allocate resources to mitigate the hazard-impact. Resource allocation within the time period immediately post-disaster (hours and days) relies heavily on secondary data: population demography, physical infrastructure, hazard intensity and vulnerability datasets, as well as historical hazard data and remote-sensing based products[1]. Primary data consists of field-based survey data such as the number of displaced persons or damaged buildings in a given area[2]. When conducting a post-hazard needs assessment, decision makers will combine different secondary datasets and often, when possible, will also compare to historical hazards that occurred nearby the affected area[3]. The affected sites are generally identified by overlaying hazard intensity data over population count, Areas Of Interest (AOI - schools, hospitals, etc) and vulnerability data, where available[4]. This is therefore a simple, unvalidated model: if two cities, city A of population 100,000 and city B of population 1,000,000, are exposed to severe hazard intensities of 9 and 6 on the Modified Mercalli Intensity (MMI) scale, respectively, then the way that the two datasets are overlaid will drastically change the impact-risk. Continuing the example of city A and B, if an impact-risk map is produced by multiplying the population by the hazard intensity, then the relative risk of city B compared to A would be 6.7. However, by taking the exponential of the hazard intensity (MMI is on a log-scale), then the relative risk of city B to A would then be 0.5. If vulnerability data is also mixed into the model, then this simple, unvalidated model loses credibility even more. Furthermore, when comparing to historical events in the same country or area, it is difficult to know whether the current event can reasonably be compared to any of the others. In addition to this, what estimate should be given if there is no available historical hazard data, such as when the last hazard occurrence was before the era of digitalisation began? The aim of Disaster Risk Modelling (DRM) is to parameterise, test and validate models such as those described above. By integrating secondary data and historical events, models are developed with the aim to accurately predict the potential impact of hazards. DRM could serve as an invaluable tool to estimate hazard impact in the short-term, allowing disaster management organisations to allocate aid more effectively, and make informed decisions to minimise long-term effects.

In this article, we focus on earthquake impact modelling using data-driven models. There are two main types of hazard impact data used in the modelling: aggregated impact data and geo-located remote-sensing derived building damage assessment data. For the aggregated impact data, these estimate the total number of each impact type over a spatial region such as a province or city. There are four different impacts: mortality, population displacement, building damage and building destruction. Note that population displacement is the maximum population displaced throughout the event. For the satellite-image derived building damage assessment data, images before and after the hazard occurrence are compared to estimate whether a building is damaged or not. This research aims to explore, using a range of different Machine Learning and advanced statistical models, the best-performing models in terms of predicting a) the total mortality, human displacement, building damage and building destruction over a given region, and b) whether an exposed building would be classed as damaged or unaffected. With the recent advances in Machine Learning (ML) models, built for performance when a large amount of data is available, this research aims to address whether ML models perform better than simple regression-based models. For the aggregated impact data models, the impact estimate data come from 161 earthquakes since 2010 (inclusive). This includes 61 countries from Africa (9), Asia (107), Europe (20), North America (23), South America (15) and Oceania (6). These impact estimates were taken from a range of different sources, such as the Global Internal Displacement Database (GIDD)[5] produced by the Internal Displacement Monitoring Center (IDMC), the EMergency events DATabase (EM-DAT)[6] produced by the Center for Research on the Epidemiology of Disasters (CRED) based at the University of Louvain, and Desinventar[7] which is produced by United Nations Disaster Risk Reduction (UNDRR). For the satellite-image derived building damage assessment data, 228,765 buildings were classified for damage, covering a total of 13 separate earthquake events. This data came from either the Copernicus

Emergency Management Service[8], or the United Nations SATellite center (UNOSAT)[9]. To help build robust models, we include a range of background data. For each earthquake event, the United States Geological Survey (USGS)[10] gridded earthquake shakemaps of foreshocks, the principal shock and the aftershocks are used to infer hazard intensity. The Socioeconomic Data and Applications Center (SEDAC) population count data[11] is adjusted by the World Bank national population estimate[12] on the day of the event to infer the hazard-exposed population on the day of the earthquake. Using datasets from the Global Data Lab[13], proxies for socio-economic vulnerability are from the Gross National Income (GNI), average lifespan from birth, and the expected number of schooling years of the current youth generation. One additional proxy for vulnerability was from the World Inequality Database (WID)[14], where all income deciles were extracted for each country at the year of the event occurrence. We also include three environment-based vulnerabilities. The first is earthquake frequency distribution based on historical earthquake data, estimated by the expected Peak Ground Acceleration (PGA) that has a 10% probability of occurring within a period of 50 years, corresponding to a 475-year earthquake return period, produced by SEDAC[15]. We also include the time-average shear-wave velocity from the ground level to minus thirty-metres[16], a gridded dataset produced by USGS. Finally, the standard deviation from the earthquake shakemap is also used as a proxy for earthquake-impact vulnerability.

An initial focus of the research was to understand the relationship between the different aggregated impact variables. For example, a common assumption is that the number of displaced persons post-disaster can be approximated by the total number of buildings damaged and destroyed, multiplied by the average household size. Forming part of the Exploratory Data Analysis (EDA), we found strong correlation ($R_{adj} = 0.98$) between the population displacement and household size multiplied by total damaged buildings for 45 earthquake events, when the logarithm was taken of both sides of the equation. Log-population displacement is estimated to equal 0.93 times the logarithm of the total building damage and destruction multiplied by household size. However, looking at the residuals of the model, there is generally a significant difference (generally orders of magnitude) between the two variables. Additionally, for population displacement estimated to be approximately 10,000 people or below, building damage is observed to always be larger than what would be expected based on displacement. This might reflect the difference in humanitarian aid (thus the displaced population surveys or Post Disaster Needs Assessment - PDNA - quality). Furthermore, we provide evidence against the use of the term ‘ground-truth data’, when referring to the impact estimate data. By gathering hundreds of different estimates, including for the same earthquake events, of the same impact types, we can measure the variance between estimates. This provides an estimate of the estimation error of the different impacts, as measured by different organisations. What is found is that, as compared to the mean value over all estimates for the same event and same impact type, the standard deviation of the error is smallest for the mortality data as compared to population displacement, building damage and building destruction. This is expected to be mostly due to government regulation of death certificates in many of the countries included in the data.

For the aggregated impact data, we apply a variety of different statistical regression and Machine Learning models. For the statistical regression, we apply linear (normal), log-normal, Poisson, negative binomial, hurdle-Poisson, hurdle-negative binomial, zero-inflated Poisson, zero-inflated negative binomial, elastic-net regularisation log-normal and stepwise-AIC-regularised log-normal models. For the Machine Learning, we apply AdaBoost; feed-forward neural network; bayesian regulated feed-forward neural network; linear, radial and polynomial basis function-based Support Vector Machines (SVM) and Random Forest (RF) models, additionally with the use of Convolutional Neural Networks (CNN) for the geospatial models. All of the above models are applied as univariate response multiple regression models. In addition to these, multivariate response multiple regression models are applied using the log-normal regression models, as well as applying univariate response multiple regression models with other impact types included as covariates. For the latter, an example is to have mortality as the response variable, but with displacement included as a covariate in the model. For the geolocated satellite-derived building damage data, we apply elastic-net regularisation log-normal; stepwise-AIC-regularised log-normal; AdaBoost; feed-forward neural network; bayesian regulated feed-forward neural network; linear, radial and polynomial basis function-based Support Vector Machines (SVM) and Random Forest (RF) models, additionally with the use of Gaussian Process Regression (GPR, otherwise known as kriging for 2D geospatial applications) for the geospatial models. To infer the feature importance of each of the hazard, exposure and vulnerability variables included in our models, we require a model-agnostic method. We choose to use the Feature Importance Ranking Measure (FIRM), which is calculated based on the variance between the response and independent variables. Furthermore, we choose to apply the FIRM method which is estimated via the Individual Conditional Expectation (ICE) curves[17].

The predictive performance of the models trained on the aggregated impact response data reflect strong performance on the mortality and building damage and building destruction data, with the population displacement-based models performing poorly in comparison. Furthermore, models trained on building damage performed better than those trained on building destruction. The top three models, in terms of the 10-fold, 5-repeat cross validated residual Mean Absolute Deviation of

Log-values (MADL) error multiplied across all impact types, were the Random Forest, radial basis-based SVM and then the log-normal regression models, respectively. This interesting result illustrates that statistical regression models can have similar performance to black-box Machine-Learning models, due to the low sample size of 161 earthquakes. With respect to the geospatial CNN model, the performance was significantly lower than almost all of the non-geospatial models, across all impact types, which further reflects the low sample size. This did not change when varying the neural network hyperparameters. By including other impact types as covariates in the models, we find that the only impact type which does not have a statistically significant improvement in predictions when either a) other impact types are included as covariates or b) when included as a covariate in the models based on other impact data types, is mortality. Therefore, by having an estimate of the number of damaged or destroyed buildings, or of the population displacement, this significantly improves the performance of models trained on these three impact types. The feature importance estimates clearly illustrate that the exposed population (first dimension of the PCA) was the most important covariate to model mortality, building damage and destruction. However, for population displacement, this was estimated to be time since the first event in the dataset. The importance of each covariate was found to be similar between the building damage and destruction data. The population displacement feature importance was found to be significantly different from the other three impact types. This inconsistency in the model covariates may be explained by several factors: the quality of the primary data of the population displacement estimates may be lower than that of the other impact types, the temporal effect of population displacement may not always be sufficiently considered when the curative organisations (such as IDMC) evaluate their estimate, or that the covariates used in this research may not be as appropriate for displacement predictions as for the other impacts. For the geolocated building damage data, the performance of the models varied significantly. As measured by the Area Under the Receiver Operating Characteristic (AUROC) value, which is ideal for unbalanced classification data, the top three performing models, with respect to the 8-fold, 5-repeat cross-validated AUC values, were the AdaBoost, GPR and Random Forest models, with 0.97, 0.96 and 0.95, respectively, which are very high. Note that the GPR is the only model that included geospatial variation, which was shown to significantly improve on all models except AdaBoost. This is an important discovery, as AdaBoost models are black-box, whereas the user can easily and intuitively interpret the regression coefficients from GPR models.

The authors recommend using a bagged-model approach if any of the models developed in this research would be applied in a post-disaster scenario: taking the top-3 or 5 models to provide estimate uncertainty. Additionally, unless further research goes into significantly increasing the sample size across a broader range of countries with diverse vulnerabilities, the authors would recommend using statistical regression-based methods to ensure that the parameterisations are plausible and intuitive. For example, that increasing hazard intensity always results in an increase in impact risk. For future work, the authors would recommend parameterising the top-performing models using a Bayesian framework to ensure that estimates of model uncertainty are also provided. Finally, future work should also study the possibility of data-driven but physics-constrained gridded models, whereby model parameterisations can be constrained to physically-plausible and intuitive values.

References

1. Assessment Capacities Project (ACAPS). Disaster summary sheet, earthquakes. (2012).
2. Thurstans, S. *et al.* 2011 edition of the sphere handbook humanitarian charter and minimum standards in humanitarian response. *Field Exch. 41* 36 (2011).
3. Assessment Capacities Project (ACAPS) and Emergency Capacity Building Project (ECB). Humanitarian needs assessment: The good enough guide. DOI: 10.3362/9781780448626 (2014).
4. Koedam, A. Rapid estimation of affected population figures, don't make acronym. **73** (2012).
5. Internal Displacement Monitoring Center (IDMC), Global Internal Displacement Database (GIDD). *website* https://www.internal-displacement.org/ (2023).
6. D., G.-S. *et al.* Emergency events database (em-dat) - cred/ofda international disaster database. *website* https://public.emdat.be/ (2023).
7. United Nations Disaster Risk Reduction (UNDRR). Desinventar - disaster information management system. *website* https://www.desinventar.net/DesInventar/download.jsp (2023).
8. Svatonova, H. Aerial and satellite images in crisis management: Use and visual interpretation. In *International Conference on Military Technologies (ICMT) 2015*, 1–5 (IEEE, 2015).
9. Miura, H. *et al.* Building damage assessment using high-resolution satellite sar images of the 2010 haiti earthquake. *Earthq. Spectra* **32**, 591–610 (2016).
10. Wald, D. *et al.* Shakemap manual: technical manual, user's guide, and software guide. Tech. Rep. (2005).

11. for International Earth Science Information Network CIESIN, C. Gridded population of the world, version 4.11 (gpwv4): Population count. DOI: https://doi.org/10.7927/H4JW8BX5 (2018).

12. Population estimates and projections. *website* https://datacatalog.worldbank.org/search/dataset/0037655/Population-Estimates-and-Projections.

13. Smits, J. & Permanyer, I. The subnational human development database. *Sci. data* **6**, 1–15 (2019).

14. Solt, F. The standardized world income inequality database. *Soc. science quarterly* **97**, 1267–1281 (2016).

15. Yetman, G. Natural disaster hotspots: a global risk analysis. *Int. Bank for Reconstr.* (2005).

16. Allen, T. & Wald, D. On the use of high-resolution topographic data as a proxy for seismic site conditions (vs 30). *Bull. Seismol. Soc. Am.* **99**, 935–943 (2009).

17. Greenwell, B. *et al.* A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755* (2018).