



Data-Driven Earthquake Multi-impact Modeling: A Comparison of Models

Hamish Patten^{1,2} · Max Anderson Loake² · David Steinsaltz²

Accepted: 29 May 2024 / Published online: 14 June 2024
© The Author(s) 2024

Abstract

In this study, a broad range of supervised machine learning and parametric statistical, geospatial, and non-geospatial models were applied to model both aggregated observed impact estimate data and satellite image-derived geolocated building damage data for earthquakes, via regression- and classification-based models, respectively. For the aggregated observational data, models were ranked via predictive performance of mortality, population displacement, building damage, and building destruction for 375 observations across 161 earthquakes in 61 countries. For the satellite image-derived data, models were ranked via classification performance (damaged/unaffected) of 369,813 geolocated buildings for 26 earthquakes in 15 countries. Grouped k-fold, 3-repeat cross validation was used to ensure out-of-sample predictive performance. Feature importance of several variables used as proxies for vulnerability to disasters indicates covariate utility. The 2023 Türkiye–Syria earthquake event was used to explore model limitations for extreme events. However, applying the AdaBoost model on the 27,032 held-out buildings of the 2023 Türkiye–Syria earthquake event, predictions had an AUC of 0.93. Therefore, without any geospatial, building-specific, or direct satellite image information, this model accurately classified building damage, with significantly improved performance over satellite image trained models found in the literature.

Keywords Disaster risk modeling · Earthquake impact models · Machine learning · Disaster statistics · Satellite image-derived building damage

1 Introduction

Earthquakes are one of the most devastating hazards that affect people and infrastructure around the world. According to events recorded in the EM-DAT hazard-impact database (Guha-Sapir 2023), earthquakes killed 725,678 people during the years 2000–2022, which is almost four times larger than any other natural hazard recorded in EM-DAT. In certain areas of the world, increasing urbanization and population growth is occurring in earthquake-prone regions, which is expected to increase the risk of human and economic losses from earthquakes in countries with poor earthquake protection-related building code regulation (He et al. 2021).

The significant loss of life, population displacement, and damage to buildings and other structures caused by earthquakes underscores the need for effective strategies to mitigate their impact. This includes the need to develop impact prediction models and to understand direct and indirect indicators of vulnerability (Albulescu 2023).

Many organizations have developed global earthquake impact models, each with varying complexity. One example is the Prompt Assessment of Global Earthquakes for Response (PAGER) (Earle et al. 2009), developed by the United States Geological Survey (USGS). The impact variables chosen as outcome variables are mortality and economic cost, by applying linear log-normal regression-based models to each impact type, tested against several loss functions. One of the limitations of the PAGER fatalities model is that only one mathematical model formulation is used (see Eqs. 1 and 2 in Jaiswal and Wald 2010). The models are also not spatially explicit, as they aggregate over the geospatial dimensions, reducing the problem to a zero-(spatial) dimensional model. Furthermore, no earthquake vulnerability variables are directly integrated as covariates

✉ Hamish Patten
hamish.patten@ifrc.org

¹ Information Management, International Federation of Red Cross and Red Crescent Societies (IFRC), 1211 Geneva, Switzerland

² Department of Statistics, University of Oxford, Oxford OX1 3LB, UK

in the model, but country-wise vulnerability values are used to cluster certain countries together. OpenQuake (Silva et al. 2014), RiskScape (Paulik et al. 2022), and HAZUS (Rozelle 2018), products developed by the Global Earthquake Model (GEM), National Institute of Water and Atmospheric Research (NIWAR) Ltd., and Geological and Nuclear Sciences (GNS) Ltd., respectively, all apply engineering-based approaches, based on historical engineering experiments on the average earthquake tolerance per building type beyond which damage occurs. Although these avoid some pitfalls of purely data-driven approaches, these methods rely entirely on the accuracy of the building type and building type tolerance data over the given spatial region, which can be erroneous, outdated or not integrate the experimental uncertainty (Ehrlich and Zeug 2008; Shultz 2017; Silva et al. 2019). Furthermore, the building type data generally have low spatial resolution (mostly administrative level one or zero (Yepes-Estrada et al. 2016)) and are often biased toward certain regions of the world (Ehrlich and Zeug 2008). Recent research has also applied deep-learning models to predict building damage by comparing satellite image data before and after floods, earthquakes, and conflict, using purely data-driven deep learning models (Zhang et al. 2022; Xia et al. 2022). This requires satellite image data, where the models in these studies (Zhang et al. 2022; Xia et al. 2022) were trained on commercial satellite data. This approach limits the outcome predictor to building damage classification, and not mortality or displacement. The predictions are made at the level of individual buildings, thus requiring significant computing power to scale up an estimate to a large spatial region. Finally, recent work on the Geospatial Data Integration Framework (G-DIF) (Loos et al. 2020) integrates different primary and secondary impact data types into a single model via the use of geospatial Gaussian process regression (GPR), also referred to as kriging (Hengl et al. 2007). The G-DIF model was trained on only four different events, thus the accuracy in different countries or earthquake events is uncertain. There is also recent research published on jointly estimating secondary and cascading hazards triggered by earthquakes and the associated building damage (Xu et al. 2022; Li et al. 2023), by the use of satellite imagery data. Due to the high spatial resolution and satellite image requirement, these approaches also require access to satellite imagery data and large computing resources to scale up an estimate to a large spatial region. As previously described, earthquake impact models have been developed to predict mortality, economic cost, and building damage using historical earthquakes. However, what is not currently present in the literature is research that compares different models, covariates (such as proxies for vulnerability), and different impact types such as human displacement.

This article presents a variety of different, globally representative, fully-supervised earthquake impact models and

model formulations, applied to four impact types: mortality, population displacement, building damage, and building destruction, including data-driven vulnerability proxies. A comparison was made of a wide range of models of varying complexity: from basic linear statistical models to some of the more advanced machine learning (ML) approaches. The performance of non-geospatial versus geospatial models was also explored. After an evaluation of model predictive performance, the top-performing models were then retrained on all data other than the Türkiye–Syria 2023 earthquakes. This earthquake was chosen to provide an estimate of model performance on outlier earthquake events of extreme severity. The models presented in this article are developed to be entirely open source, negligible cost, computationally efficient (predictions delivered within seconds), and validated in a statistically rigorous manner. Ultimately, our goal was to provide insights that can inform the development of more accurate and robust models for earthquake impact prediction, which can support effective disaster risk reduction and management strategies.

2 Data and Methods

In this section, the data integrated into this research as well as the methods applied to the data are presented. For the data, a description is provided on the observational impact estimate data and the model covariates, including methods for data reproducibility. A brief description is also provided of the different models explored in this research.

2.1 Historical Earthquake Impact and Disaster Risk Related Data

The data integrated into this analysis can be split into two components: observed impact estimates and the background covariates (including spatial information).

2.1.1 Impact Estimate Data

There are two types of spatial impact data: aggregated polygon and point data. Aggregated polygon data are the most commonly known type of hazard impact observation, whereby an estimate is made for the total impact counts over the entirety of a given spatial region (such as a country or city). The point data consist of single points in space whereby the object in question is categorised by hazard-induced impact (or damage). Please note that we refer to the impact observations as estimates, which is separate from model predictions in that they are primary or secondary data and not a model prediction. The spatial polygon data presented here include 375 observations for 162 earthquakes that occurred since 2010 (inclusive) in 61 countries

of Africa (8), Asia (104), Europe (17), North America (20), South America (13), and Oceania (6). The impact types are mortality, displacement, building damage, and building destruction. Please note that all but the first have subjective definitions, differing substantially even between world leading hazard-impact monitoring databases such as EM-DAT (Guha-Sapir 2023) and DesInventar (UNDRR 2023).

The aggregated spatial polygon data analyzed in this study are a collection of observed impact estimates from a variety of different sources. An effort was made to ensure that the source databases came from legitimate and credible entities that were mostly from international nongovernmental organizations (INGOs), national government disaster management agencies (NDMAs), or international organizations (IOs), although for many events we also relied on other sources. Often, these aggregated observed impact estimates are referred to as “ground-truth data,” a term that we propose should be avoided. Some observed impact estimates that were aggregated over the same spatial location, by different organizations, were found to be inconsistent. We may quantify this inconsistency by the coefficient of variation (CV), the ratio of standard deviation to mean. The CV estimates of the impact estimate data employed in this study are 0.15, 0.48, 0.46, and 0.59 for mortality, displacement, building damage, and building destruction, respectively, illustrating that the relative disagreement between observed impact estimates is generally lower when estimating mortality, and estimated to be the largest for building destruction. When selecting between conflicting observations to avoid producing bias in the results, curated observed impact estimates produced by organizations such as the Centre for Research on the Epidemiology of Disasters (CRED) and the Internal Displacement Monitoring Centre (IDMC) were prioritized, and, when not curated by these organizations, all observed impact estimates that had a very low variance-to-mean ratio were discarded. In total, we have 171, 105, 285, and 231 data points for mortality, displacement, building damage, and building destruction. When combining multiple outcome variables, the sample size reduces significantly, such that earthquake events with recorded mortality and displacement estimates, building damage and destruction, and all four impact types is 94, 224, and 59, respectively.

The spatial point data integrated into this work are satellite image-based building damage assessment provided by both the Copernicus Emergency Management Service (Svatonova 2015) and the United Nations Organisation for SATellite center (UNOSAT) (Miura et al. 2016). After an event, such as an earthquake, occurs, these two organizations access satellite images before and after the event and compare the images to look for infrastructural damage, predominantly in buildings. The entire area exposed to the hazard is not generally covered in the analysis, but a spatial region is chosen based on where the highest infrastructural impact

is expected. Damage is classified into several categories, mostly from the EMS-98 classification standard (Copernicus Emergency Management Service n.d.) but also including several defined in the UNOSAT event data. The classifications are “unaffected,” “possible damage,” “damaged,” “negligible to slight damage,” “moderate damage,” “substantial to heavy damage,” “destroyed,” “very heavy damage,” and “destruction.” This research integrated 369,813 buildings (as spatial points) assessed for damage from a total of 26 separate earthquakes that took place in 15 different countries since 2010 (including 2010), taken from both the UNOSAT and the Copernicus databases. The events involved countries from Asia (8), Europe (3), North America (2), South America (1), and Oceania (1). We discarded all entries labelled as “possible damage” from the analysis, and grouped all data points not classified as “unaffected” as “damaged.” There were 15,361 buildings classified as damaged, and 354,452 as unaffected.

Note that the aforementioned earthquakes were also purposefully chosen to not have additionally had tsunamis, to avoid convolving the different hazards and the types of damage that tend to result from them. We also included all earthquake pre- and after-shocks that occurred within a 5-day and a 3-week period before and after the principal earthquake, respectively. For example, for the recent 2023 Türkiye–Syria earthquakes, 16 aftershocks were included.

2.1.2 Earthquake Hazard and Exposure Data

To enhance the predictive power of the models developed in this study, many different covariates were used. These may be split broadly into three groups: hazard, exposure, and vulnerability data. The hazard data applied to this work come from the United States Geological Survey (USGS) database on historical earthquakes and their ShakeMaps (Wald 2005). A ShakeMap is a measure of the intensity of the shaking and ground motion due to an earthquake, measured on the MMI scale. Across all earthquakes (with $MMI > 5$) that occurred during the event, the variable *hazMax* corresponds to the maximum earthquake intensity experienced over the entire event, aggregating spatially. However, this must not be confused with the variable *max_MMI*, used in the classification models of the geolocated building damage data, which refers to the maximum earthquake intensity that the individual building was exposed to. For exposure data, this study integrated population count data from the Socio-economic Data and Applications Center (SEDAC) (CIESIN 2018), adjusted to correspond to the population on the day of the event via interpolation from the World Bank national population count (World Bank 2022). More recent sources could have been used, but due to our events going back to 2010 we preferred to use data that provide larger temporal coverage in a multi-country form.

2.1.3 Inferring Disaster Vulnerability

The choice of variables is driven mainly by data availability and extent of geographic and temporal coverage. Many of the socioeconomic vulnerability data included in this research come from the Global Data Lab (Smits and Permanyer 2019), which provides subnationally aggregated data from census and survey. There is a very strong coverage of different countries around the world, and the data often have coverage up to administrative level 2, making this dataset a strong candidate for global hazard impact modeling. *GNIc* is the Gross National Income (GNI) per capita. *ExpSchYrs* is the number of school years a child is expected to have by the time they finish school. Although not a direct vulnerability indicator for earthquake impact, this variable is expected to reflect current improvements in public infrastructure. An estimate of the average lifespan of the population is provided by the variable *LifeExp*, reflecting population health and medical infrastructure/quality. Population density (from SEDAC-WB) of the given area is included as a proxy for urban-rural differences, named *Population*. We also included several environmental-based vulnerability variables that are provided by other organizations, with data available in a high spatial resolution gridded format. *EQFreq* is the expected Peak Ground Acceleration (PGA) value that has a 10% probability of occurring within a period of 50 years, which corresponds to a 475-year return period, produced by SEDAC (CHRR and CIESIN 2005). Note that this dataset was produced using earthquake data from before 2002, and therefore is completely independent of the earthquake events analyzed in this study. The time-average shear-wave velocity in the upper 30 m of ground level, produced by USGS, provides an indicator of soil stiffness (Heath et al. 2020). The ShakeMap intensity uncertainty, *hazSD*, is the maximum ShakeMap uncertainty term across all pre-, principal, and after-shocks, per pixel, and was used as a proxy for the presence/absence of seismic centers and thus may correlate with historical occurrence of severe earthquakes, as seismic centers are strategically located in areas prone to earthquake impacts (Wald 2005). This only represents one element of the uncertainty (Wald 2008) and thus this variable is not expected to have a high feature importance relative to other covariates. Time in days since the earliest event in the data is also included as a covariate (21st February 2011—Christchurch earthquake, NZL). Finally, a function of the World Inequality Database (WID) first to ninth deciles were included as covariates, as a proxy for societal inequality and disparity.

2.2 Methodology

In this study, we combined the aforementioned disaster risk covariates in various model formulations to predict the

different earthquake impacts (for example, mortality). However, we have a variety of data sources that are of raster, point, and polygon data types. This section describes how we combined the different types of data to allow for impact prediction and to infer the important covariates that can be used to model earthquake impacts. A description is then given of the different ML and traditional statistical models. Once the models had been developed using grouped (by earthquake event) k-fold, 3 repeat cross validation, we then retrained the top-performing regression and classification models on all earthquake events other than the 2023 Türkiye–Syria earthquakes, to measure predictive performance on this severe outlier earthquake event.

2.2.1 Data Manipulation

The methods used to combine the observed impact estimates with the hazard, exposure, and vulnerability data are as follows: the variables *hazMax*, *max_MMI*, *hazSD*, *EQFreq*, and *Vs30* were interpolated using cubic splines onto either the spatial points directly or the population count data grid-point centroid in the case for the aggregated spatial polygon impact data. These gridded data are of 30 arc-second spatial resolution. For the polygon data-based covariates *GNIc*, *ExpSchYrs*, and *LifeExp*, the nearest-neighbor value was calculated at either each spatial point or population count data grid-point, again, depending on the impact spatial data type. For the satellite imagery-based building damage assessment data, we also interpolated the population density data onto each building location as an additional covariate. For the spatial point data, models were then directly applied to the data. However, for the aggregated spatial polygon-based observed impact estimate data, we had to aggregate the gridded data to the same polygons as the observed impact estimate data. When the observed impact estimate was made over a region that was well defined as a specific administrative level, we used the (Database of Global Administrative Areas (GADM) (University of California, Berkley 2022) for missing data, and then extracted any further missing data from the OpenStreetMaps (OpenStreetMap contributors 2023). For the non-geospatial regression models, the gridded background data were then aggregated into each of the polygons that contained observed impact estimates. The hazard-related variables: *hazSD*, *EQFreq*, and *Vs30* were averaged over each polygon. For the socioeconomic variables: *GNIc*, *WID*, *ExpSchYrs*, and *LifeExp*, a weighted mean was calculated over each polygon, weighted by the population count at each of the grid points. The hazard exposure was then calculated for the aggregated impact polygon. For each observed impact estimate polygon, the number of people exposed to at least {5, 5.5, 6, ..., 9} MMI was calculated by summing the population count data per hazard-exposed boundary. Calculating the hazard exposure produced several additional

covariates, that we named *Exp5*, *Exp5.5*, ..., *Exp9*. As the purpose of this research was not only to improve predictive performance but also conduct inference on important covariates for disaster risk, the covariate data had to be modified to reduce the influence of variable collinearity. The covariates that strongly correlated with one another were the WID deciles (*p0p100*, ..., *p90p100*), as well as the hazard exposure variables *Exp5*, ..., *Exp9*. In order to resolve this issue, we applied Principal Components Analysis (PCA) separately to the WID deciles and the hazard exposure. The first two dimensions were retained (such that at least 80% of the variance was explained by the two variables), and are referred to as *WIDDim1*, *WIDDim2* and *ExpDim1*, *ExpDim2*. After adjusting these variables, the maximum Variance Inflation Factor (VIF) across all covariates was less than 5, implying low risk of collinearity. The covariates were then normalized, for both the spatial point and polygon data.

2.2.2 Predictive Models

The two impact estimate data forms of spatial point and polygon data also have different data formats: the point data are binary categorical variables (unaffected and damaged), whereas the polygon data are integers. For the point data, classification models must be used, and for the polygon data regression models. We applied 9 and 17 different model frameworks to the point and polygon data, respectively, including geospatial models for each. For the point data, thus the classification models, we applied support vector machines (SVM) with linear, radial, and polynomial basis functions, referred to in figures as *svmLinear*, *svmRadial*, and *svmPoly*, respectively. We also applied naive Bayes, AdaBoost, random forest, and a generalized linear model (GLM) with elastic net regularization (GLM-net), referred to in figures as *naive_bayes*, *ada*, *rf*, and *glmnet*. For the polygon data, we applied GLMs using linear, poisson, log-normal, hurdle-poisson, hurdle-negative binomial, zero-inflated poisson, and the zero-inflated negative binomial regression, referred to as *LM*, *pois*, *lognorm*, *HurdlePois*, *HurdleNegBin*, *ZIpois*, and *ZInegbin*, respectively. Other than the linear model (*LM*), the other models are specifically built to perform well for count response data. The last five models are robust against heteroscedasticity, and the last four models are built to accommodate zero-inflation in the models. For the polygon data, we also applied a standard random forest; linear-, radial-, and polynomial-basis SVM; log-normal regression; feed-forward neural network with Bayesian regularization; elastic net regularization (GLM-net), and a linear model with stepwise AIC-based regularization. These are referred to in the figures as *rf*, *svmLinear*, *svmRadial*, *svmPoly*, *lognorm*, *nnet*, *brnn*, *glmnet*, and *lmStepAIC*, respectively. All regression and classification models are fit with grouped (by earthquake event) 10-fold

and 8-fold cross validation, respectively, repeated at least 3 times. 8-fold was chosen for the classification models to save on computation and to reduce overfitting due to the small (26) total number of earthquake events.

In addition to the models mentioned above, we included two more model classes in our study: convolutional neural networks (CNNs) (LeCun et al. 2015) and Gaussian process regression (GPR) (Hengl et al. 2007). The GPR tunes a spatial covariance matrix with two or three additional parameters on top of the covariate parameters. Exponential, spherical, and Gaussian covariance models were applied, with the best performance from applying the Gaussian covariance. The maximum distance to be considered by the variogram was fixed at 100 km, beyond which it is assumed that the spatial covariance between points does not vary. The variogram range value was estimated by the model as 56 km. The CNN regression model was applied to the aggregated spatial polygon impact data. All covariate values (population, hazard intensity, and other covariates) in grid points that lie outside of the impact spatial polygon region are set to zero. Standard CNN frameworks accept gridded data with a given, pre-specified size. The size is dictated by the largest range of longitude and latitude values across all events, with the remaining events filled up with extra columns and rows of zeros. A padding layer was also added that is equal to the maximum CNN-filter array size, minus one. Additionally, different resizing on the filled, padded, final array was done both to reduce computational costs and to infer the influence of the minimum scale-length on the predictive performance of the model. Due to the low sample size of the number of events included in this work, the number of trainable parameters in the CNN have to be kept as low as possible. The general structure of the CNN is as follows: first a data augmentation layer was applied with a random horizontal flip and random rotation by an amount in the range $\in [-0.4\pi, 0.4\pi]$, then a 2D convolutional layer with N_c layers, a 2D max pooling layer with S_p as the symmetric pooling size, and we flattened the layers, then a dropout layer with p_D as the dropout rate, and we then applied a densely connected layer with N_d as the number of densely connected layers. The covariates input into the model were only the hazard intensity and population. These two variables were selected as the minimum required to predict any of the impacts. A large number of models were trained and tested consistently with the previous models applied in this study, using grouped 10-fold cross validation with 3 repeats. The aforementioned CNN hyperparameters were simultaneously varied over $N_c \in \{1, \dots, 5\}$, $S_p \in \{1, \dots, 5\}$, $p_D \in \{0.1, 0.2, \dots, 0.8\}$, and $N_d \in \{1, \dots, 10\}$, resulting in a total of 2,250 models evaluated for each of the four impacts.

The distance metric, or cost function, is what the algorithm optimizes over in order to train the model. For the point data, we measured the performance of predicting the

binary damage-unaaffected classification using the distance metric of the area under the receiver operating characteristic curve (AUC). We chose this metric as it is not affected by the fact that our point dataset is imbalanced with respect to the number of unaffected and damaged buildings. For the polygon data, we used a bespoke distance metric that we refer to as the mean absolute deviation of logs (MADL), calculated using Eq. 1. Note that the addition of 10 to the outcome variables is to accommodate impacts that are equal to zero, with 10 chosen to shift the asymptote of the log function sufficiently to prevent over-penalizing differences between small impacts, such as between $y = 0$ and $\hat{y} = 1$. Note that some models do yield negative predictions of \hat{y} , and we truncated these to 0. Additionally, for the log-normal GLM, we first modified the response data $y^* = \log(y+10)$ and then used the standard mean absolute deviation (MAD) metric, so that the model predictive error can be compared.

$$MADL = \sum_{i=1}^N \left| \log(y_i + 10) - \log(\max(\{\hat{y}_i, 0\}) + 10) \right| \quad (1)$$

When applying the GLM regression models to the polygon impact data, we took a brute-force approach to ensuring that the covariate inclusion ensures maximum predictive performance. With 12 possible covariates there are $2^{12} - 1 = 4,095$ distinct model covariate formulas, excluding the null model. Multiplied by 7 different GLM models and 4 impact types, and adding in the multivariate models and the impact-as-covariate models (which add either 1 or 2 extra covariates), there were more than 1.3 million GLMs trained in this work. Each was evaluated by the performance of the output model distance under grouped 10-fold cross-validation. Once all the models were trained and cross-validated, the predictive performance (as measured by either the AUC or the MADL) was then compared between the best

performing GLM model formulations, per model type (for example, SVM or linear regression). In order to estimate the feature importance across the different models, we applied a model-agnostic feature importance method. For this, we used the feature importance ranking measure (FIRM), which is a variance-based method. The FIRM was combined with individual conditional expectation (ICE) curves, and then scaled, to provide the final feature importance values (Greenwell et al. 2020).

3 Results

The results section is split into three parts: aggregated impact data model results, building damage point data model results, and a section dedicated to applying these models to held-out extreme events via application to the 2023 Türkiye–Syria earthquake event.

3.1 Aggregated Impact Data

The MADL error value, as calculated using Eq. 1, for the best performing covariate formulation of each model is shown in Fig. 1, for each impact type. Note that the y-axis is on a log-scale. The y-axis error bar is calculated from the standard deviation across the grouped k-fold cross validation error values. Via use of the Welch's t-test for unequal variances, we compared every model only to the overall top model for each impact type. This ad hoc approach focuses entirely on estimating the out-of-sample prediction error of the expected prediction values, and thus does not include the model uncertainty. Note that this is a non-standard method of model comparison, but likelihood ratios or alternative methods are not possible due to the mixture of machine learning and likelihood-based models. For mortality, the

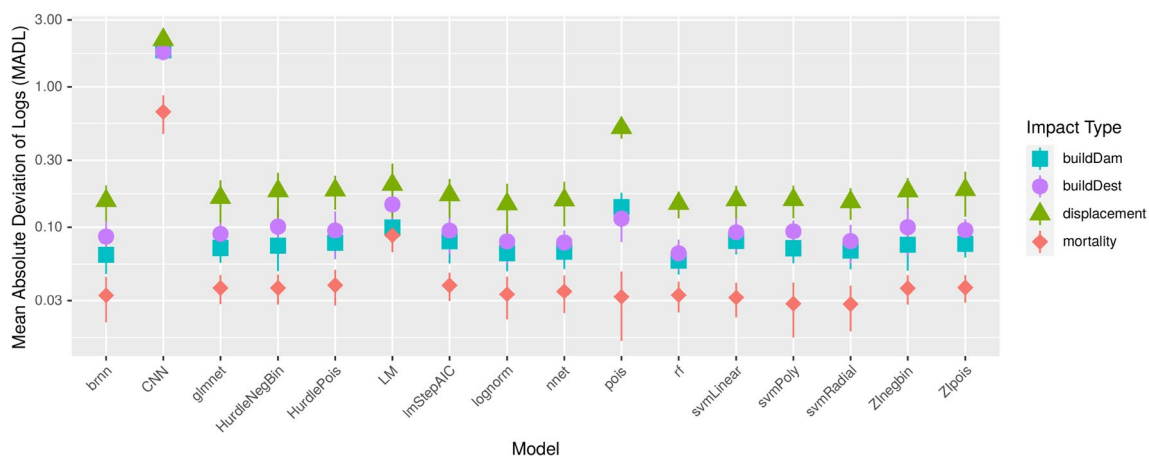


Fig. 1 Performance of each of the models on the spatial polygon impact data, on a log-scale. The color and symbols of the points separate the different impact types. The error bars represent the variance of the MADL distances when applying the grouped 10-fold cross validation

polynomial basis and radial basis function based SVM models were statistically significantly different ($p = 1.4 \times 10^{-5}$) from all other models. For displacement, the random forest, radial basis-based SVM, and log-normal models were found to be statistically significantly different ($p = 0.038$) from all other models. For both the building damage and destruction models, the random forest model was found to be significantly better ($p = 1.4 \times 10^{-7}$ and $p = 1.5 \times 10^{-22}$, respectively) than all other models. Multiplying the MADL error value for each impact together provides an estimate for the overall performance. Ranked by this overall error value, normalized to the minimum error, the top five models were the random forest (1.00), radial-basis SVM (1.27), log-normal (1.38), feed-forward neural network with Bayesian regularization (1.51), and the simple feed-forward neural network (1.56). For the geospatial model using convolutional neural networks (CNN), the results (Fig. 1) indicate poor performance compared to other models. This is most likely due to the low ratio of the sample size to number of model parameters that is required to model each impact. The most successful CNN structures applied one convolutional filter and one densely connected layer (N_c and N_d , respectively), a 2D symmetric max pooling layer of size three-by-three (S_p), and a dropout layer of approximately 0.5 (p_D). This reflects the influence of the low sample size on model overfitting, which is the reason why model performance was generally low as compared to the other models applied to the aggregated spatial polygon impact data.

Figure 1 illustrates model performance in predicting each impact type. Unfortunately, due to the presence of zero-inflation in the mortality data (25% of the data—42 data points in total—had a mortality estimated to be 0, where there were less than 5% for all other impact types), the MADL value for mortality cannot easily be compared with the other impact types. The average MADL values indicate that human displacement is the most difficult to predict, then building destruction and then building damage. This is unsurprising, given the inconsistent definitions of human displacement, as well as the variability of displacement over time. Figure 2 shows the model performance, not as a function of the MADL cost but of a direct comparison between the observed and predicted impact values, using the random forest model. For this specific model, it can be seen that it over-estimates impacts for small observed impacts (low relative to each specific impact type), and under-estimates impacts for larger observed impacts. Therefore, the authors propose that, to maximize performance, a simple quadratic regression could be added to minimize the residual error when using the random forest model.

For each impact type, we also calculated the feature importance using the FIRM-ICE method, shown as a percentage in Fig. 3. The largest feature importance for all impact types other than population displacement was the

first dimension of the exposed-population PCA variable, for all impacts. An encouraging result is the correlation between the feature importance of building damage-based and destruction-based impact models. The expected number of years of schooling is estimated to be the third-most important variable for both building damage and destruction impact types. This variable reflects investment in improving schools and increasing national schooling capacity, and thus may also generally reflect investment in national social and physical infrastructure, such as disaster preparedness or welfare development. Please note that the average feature importance shown in Fig. 3 was calculated as the weighted mean, using the CV estimates mentioned in the Data and Methods section.

For each impact model, we can also include, as a model covariate, any of the other impact types. This reflects the correlation between the different impact types as compared to the other hazard, hazard exposure, and vulnerability covariates. Table 1 reflects the improvements in the model performance when including the impact covariate in the model formulation. Due to the variance in the MADL values per cross validation fold (10 in total), only model prediction improvements that are statistically significantly different from the no-impact-covariate model will be designated as “improved,” as shown in the column “ p value (95%).” Note that there are differing sample sizes for each of the simulations, which means that the MADL values should not be compared between different models in the table (unless they have the same two impacts). Population displacement, building damage, and building destruction are shown to significantly improve the model predictions as a covariate, but not mortality. However, an exception to this is the displacement response-based model with building damage as a covariate, whereby the MADL value is shown to be improved over not including the building damage as a covariate, but is not statistically significant.

3.2 Building Damage Point Data

Using the AUC value as a metric for model performance, Fig. 4 shows that the top model for classifying building damage from point data was AdaBoost, followed by GPR and then random forest. The total number of different events and the total number of countries included in the building damage point data led to several of the vulnerability covariates having less than 30 unique values. Therefore, to avoid overfitting, we removed the following vulnerability covariates from the model formulation: WIDDim1, WIDDim2, and time. Using the FIRM-ICE method, the feature importance is shown in Fig. 5. The feature importance seems to correspond well with the average of the aggregated impact data from Fig. 3: earthquake frequency is the most important feature, but *hazSD* and *Vs30* are the least important. The variables

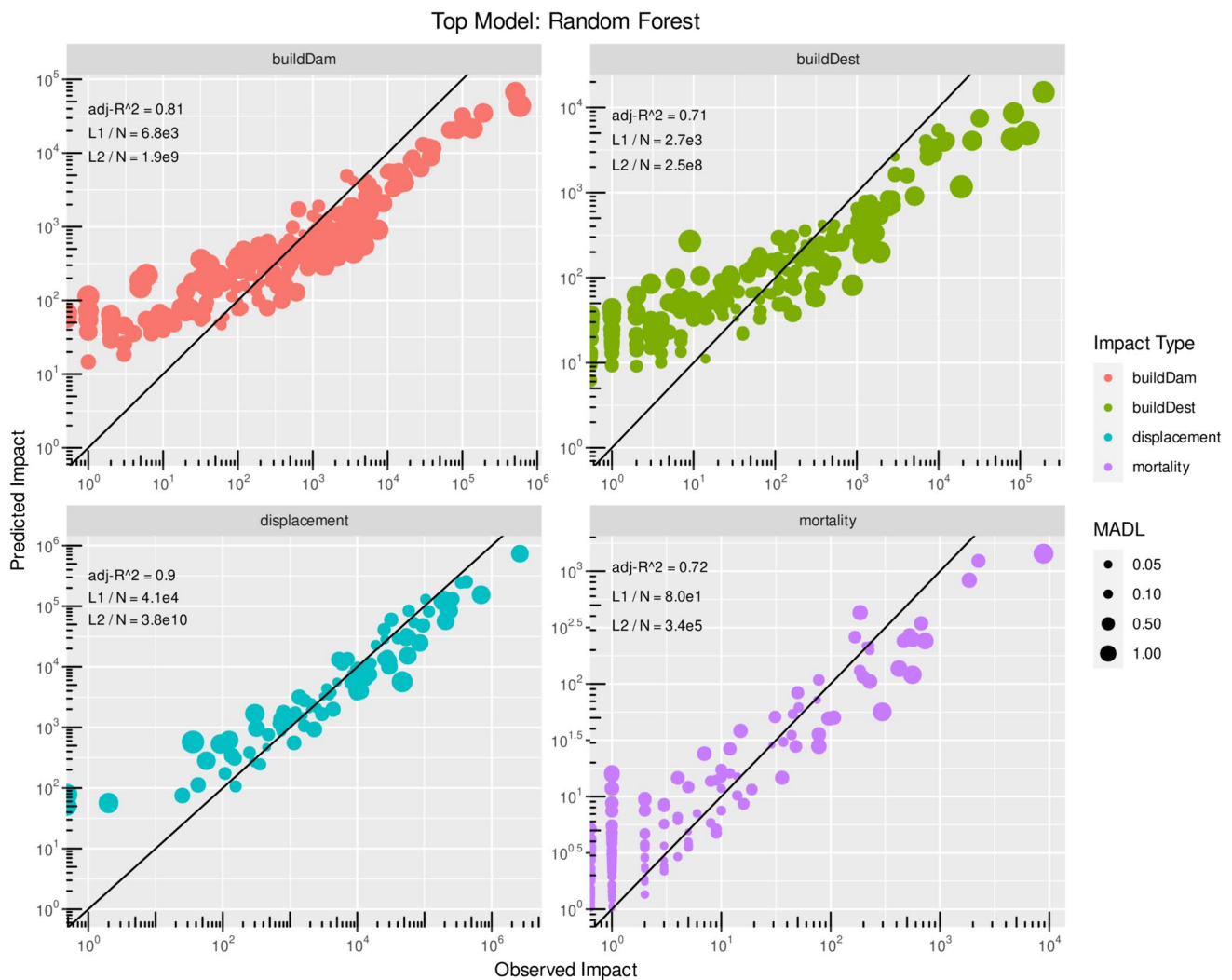


Fig. 2 A direct comparison between the predicted and observed impact values, for each impact type, using the random forest model, which was one of the top models for all impact types. The size of the

points corresponds to the MADL cost. The text in each plot shows the adjusted R-squared metric and the L1 and L2 norms divided by the sample size for each impact type

GNic, *LifeExp*, and *ExpSchYrs* all seem to lie within the middle range of feature importance, indicating potential correlation but potentially not statistically significant.

We present here an additional model applied to the building damage point data, that uses the spatial information: the longitude and latitude position of each building evaluated. This model applies Gaussian process regression (GPR), which relates each building to others via a spatial covariance matrix: the closer one building is to another, the more likely it will be to have the same classification. The mean value is dictated by the covariates included in the model, with the additional correction from the spatial covariance. The benefits of GPR over the Ada model is that, like GLMs, the model parameterization is intuitive and interpretable. However, the benefit of using the spatial covariance is lost when applying to a new location that is further than the 100

km maximum distance limit. What this implies is that this model is not expected to predict well on events that occur in locations far from the locations included in this analysis (which is only 26 events in total). On the contrary, if an initial analysis has already taken place of some of the building damage of the new event using satellite imagery data, then this GPR model could be used to predict in areas with few impact estimates.

3.3 Application to the 2023 Türkiye–Syria Earthquake Event

In this section we apply the top regression and classification models to the 2023 Türkiye–Syria earthquakes, to evaluate model performance on severe, outlier earthquake events. Compared to all other 161 earthquakes, this specific event

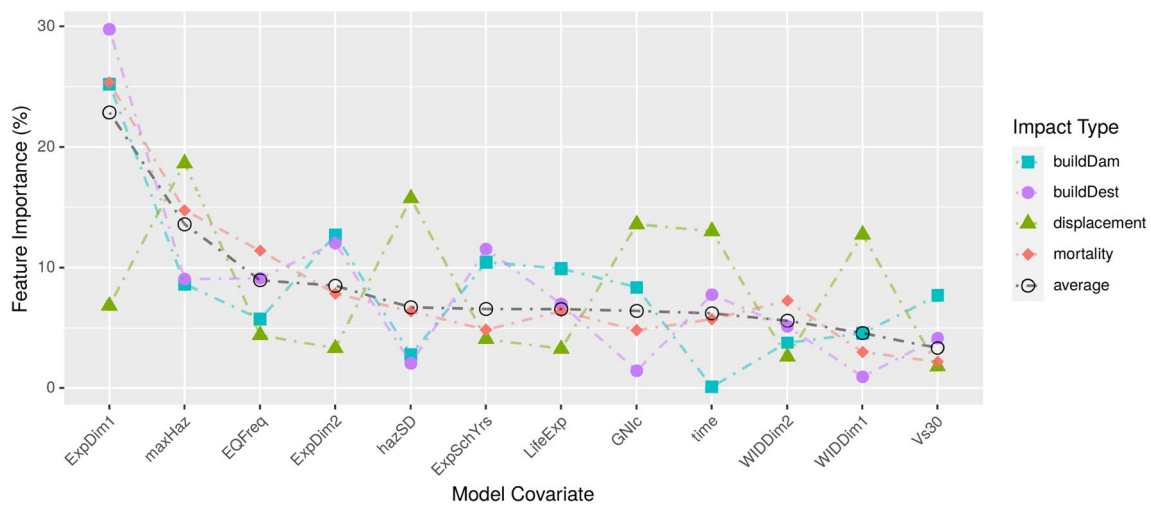


Fig. 3 Feature importance for each impact type for the aggregated spatial polygon data, as a weighted combination of the best performing models and the models that were not statistically significantly different (Welch’s t-test with $p > 0.05$) from the best performing model per impact

Table 1 Increase in model performance by including other impact types as covariates

Impact outcome variable	Impact covariate	MADL with covariate	MADL without covariate	Improved? (Stat. Sign.)	p value (95%)
buildDam	buildDest	0.042	0.076	TRUE	7.70×10^{-35}
buildDam	Displacement	0.173	0.233	TRUE	1.87×10^{-5}
buildDam	mortality	0.190	0.183	FALSE	0.335
buildDest	buildDam	0.045	0.083	TRUE	4.54×10^{-67}
buildDest	Displacement	0.171	0.261	TRUE	7.30×10^{-7}
buildDest	mortality	0.174	0.185	FALSE	0.323
Displacement	buildDam	0.170	0.178	FALSE	0.355
Displacement	buildDest	0.175	0.243	TRUE	0.016
Displacement	mortality	0.148	0.160	FALSE	0.101
Mortality	buildDam	0.064	0.066	FALSE	0.386
Mortality	buildDest	0.058	0.069	FALSE	0.057
Mortality	Displacement	0.066	0.070	FALSE	0.252

The MADL value for the model including the impact covariate must be more than that of the best model without an impact covariate, as well as the p value to be statistically significant (95%) in order for the “Improved?” column to appear as “TRUE.”

was above the 95th percentile for all population exposure levels between 5 and 9 MMI. Without taking into account the 2023 Türkiye–Syria earthquake event, the 95th percentile of the MADL error using the out-of-training predictions with the random forest regression model over all of the 161 individual earthquake events for mortality, displacement, building damage, and building destruction was 0.8, 1.8, 2.2, and 1.9, respectively. The MADL error values for the same random forest regression model when applied to the 2023 Türkiye–Syria earthquakes event was 5.1, 5.2, 5.0, and 4.0 for mortality, displacement, building damage, and building destruction

respectively. Therefore, even for the top performing random forest model, the large MADL error compared to all of the other earthquake event out-of-training prediction values reflects the limitations of extrapolating the aggregated impact data models to severe, outlier earthquakes. For the classification models, the predictive performance, in AUC values, of the AdaBoost and GPR models on the held-out Türkiye–Syria earthquakes event was 0.93 and 0.65, respectively. This held-out validation data included 27,032 buildings with 1,840 classified as damaged, which is a representative sample size to validate the model against.

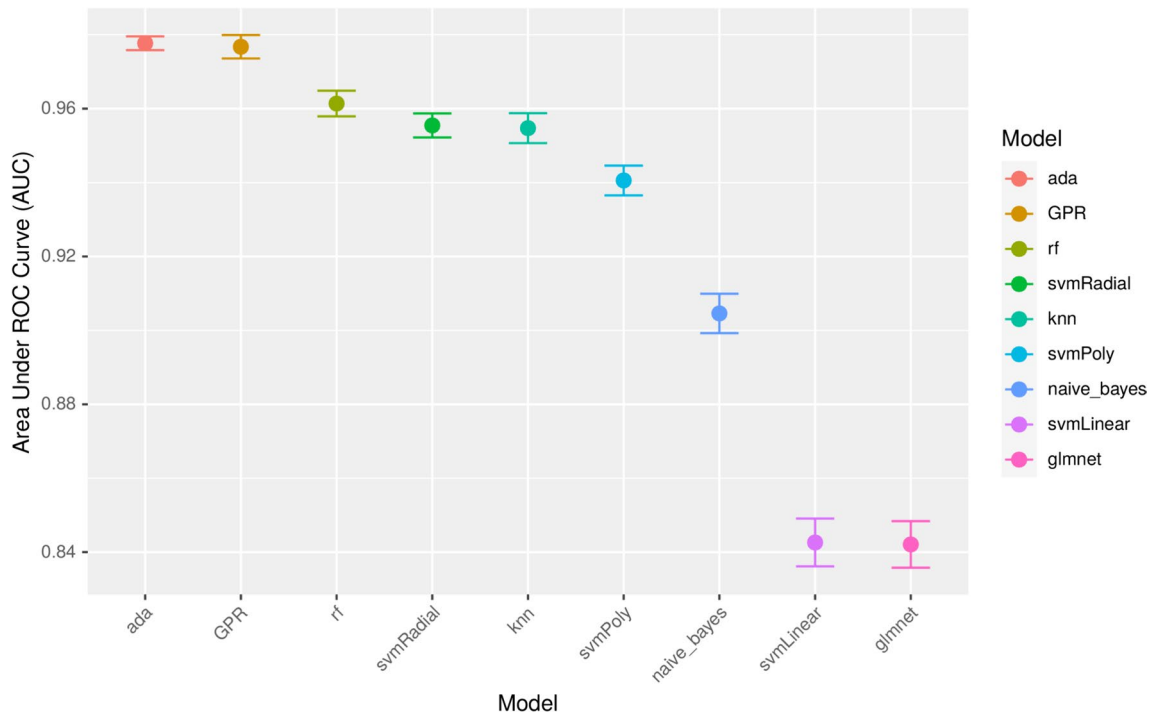


Fig. 4 Performance of the different machine learning models applied to the satellite image-based building assessment spatial point data, with respect to the average area under the receiver operating characteristic curve (AUC) values from the cross validation

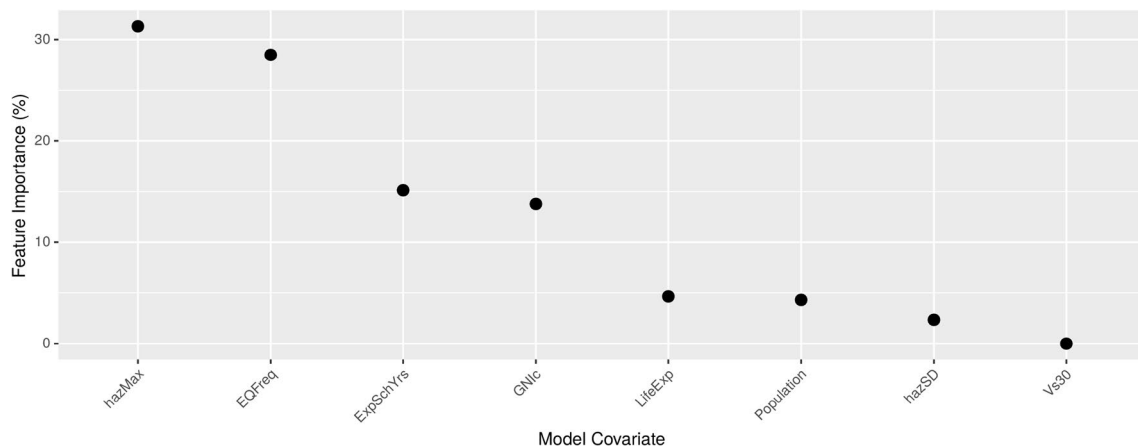


Fig. 5 Feature importance for the satellite image-based building assessment spatial point data

4 Discussion and Conclusion

The research presented in this article compared the predictive power of various data-driven models and their appropriateness for earthquake impact modeling, and tried to uncover characteristic information about hazard impact data and covariates. In the context of impact data with a relatively low number of events, which is generally typical of historical disaster data for which hazard and impact

data have been paired and curated, a large variety of different models have been presented. Mortality models were found to have smaller (MADL) error, followed by building damage, then building destruction, then population displacement. Evidence suggests that overall improved model performance for mortality was not entirely linked to the presence of zero- or low-counts in the model (thus resulting in a smaller margin-for-error). This evidence was provided by showing that observed mortality estimates also have the lowest coefficient of variation between same event

and same spatial extent observed impact estimates made by different organizations. Population displacement model predictions were shown to have the worst performance. This may result from errors not so much at the model level but at the level of the displacement data. Such an error could come from temporal effects not being properly taken into account when estimating the maximum population displacement caused by an event, or that estimates of displaced populations is a process itself that results in large errors due to the inaccuracy of the measurement process. However, it was found that population displacement, building damage, and building destruction can be used to significantly improve on model predictions when they are included as model covariates for one another. This may reflect that often the population displacement is estimated by curative data sources based on the building damage and destruction data rather than from primary data collection. A possible explanation for why conditional mortality-based models (as both response or covariates) do not improve predictive performance is the presence of zero-inflation.

With respect to the different supervised machine learning and parametric statistical models applied to the impact data, the predictive performance of the machine learning models is generally superior to that of parametric statistical models. However, an intuitive understanding of the relationship between the covariates on the impact magnitude then becomes convoluted, and is not easily interpretable in a physically intuitive manner. We found that the top three models for predicting aggregated earthquake impact were random forest, the radial basis-based support vector machine, and a log-normal-based linear regression model. For mortality data, however, which often have zero or low counts, the performance gap between the count-based parametric statistical and machine learning models was much smaller. With respect to geospatial models, there was insufficient data in the aggregated impact data to train a convolutional neural network (CNN) geospatial model to have comparable performance with the alternative models. However, the use of Gaussian process regression (GPR) for the spatial point data was shown to perform nearly as well as the best non-geospatial model (AdaBoost). Adding the influence of geospatial variance in the model therefore improves performance in predicting damaged and unaffected buildings.

For the covariate importance of the vulnerability variables, reasonable agreement was found between the (weighted) average of the aggregated impact data and the geolocated building data. The results illustrate that earthquake frequency is the most important feature, and that soil stiffness and ShakeMap uncertainty were the least important. In between these two extremes were the importance of the GNI, life expectancy, and expected number of schooling years covariates. With respect to the feature

importance on the different impact types in the aggregated data, the earthquake-exposed population was the covariate most strongly correlated with mortality, building damage, and building destruction. This is an intuitive result for these three impacts, as without the stable presence of a population no population or building impact is possible, but the fact that population displacement has the earthquake intensity (ShakeMap) average uncertainty and maximum ShakeMap intensity value as the two (approximately) joint most important covariates is likely to imply error in the observation/measurement process.

By exploring the severe outlier event of the 2023 Türkiye–Syria earthquakes, this study also explored model limitations for extreme events. For the regression models trained on the aggregated impact data, performance was poor. This highlights the need to check whether a comparison can be made between any given event with the historical earthquakes in the data. However, for the classification models, the remarkable performance (AUC = 0.93) of the AdaBoost model provides an unexpected result: with no building-specific information (for example, building height or ground floor surface area), no geospatial covariance information, and no satellite image data, we can train a highly accurate model to classify building damage. This outperforms the models that were trained directly on the satellite images from Xu et al. (2019) and Lee et al. (2020), which had AUC values between 0.62 and 0.73 for models tested on untrained events. In combination with open building datasets such as OpenStreetMaps (OpenStreetMap contributors 2023), this model could significantly improve earthquake-related building damage classification in the immediate phase post-disaster, by identifying potential damage hotspots to aid workers and government agencies significantly faster and without incurring significant costs, as opposed to models built on the satellite imagery directly.

For future work, we suggest devoting more effort toward developing hybrid physics-based machine learning models whereby monotonic relationships between variables is inherently built-in, and to dedicate more effort into calculating model uncertainty, rather than just predictive performance, through the use of Bayesian hierarchical model frameworks.

Acknowledgments The authors would like to acknowledge the attentive curation that goes into databases such as the Global Internal Displacement Database, Internal Displacement Monitoring Centre (GIDD—IDMC); Desinventar, United Nations Office for Disaster Risk Reduction (UNDRR); and the Emergency Events Database (EM-DAT), Centre for Research on the Epidemiology of Disasters. Furthermore, this work was accelerated significantly by the availability of a range of R-packages related to parametric statistical regression, machine learning, and deep learning. This work relied significantly on the caret, tensorflow, keras, gstat, and fastknn packages. In particular, the authors would like to thank Justin Ginnetti, Sylvain Ponserra, Maria-Teresa

Miranda Espinosa, and Bina Desai for the fruitful discussions with respect to population displacement. This research was funded by the Engineering & Physical Sciences Research Council (EPSRC) Impact Acceleration Account Award EP/R511742/1.

Code Availability The GitHub repository of the code is openly available at <https://github.com/hamishwp/ODDRIN>, in the files ML_ODD-data.R and ML_BD-data.R. The custom class definitions are in ODDobj.R and BDobj.R.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Albulescu, A.C. 2023. Open source data-based solutions for identifying patterns of urban earthquake systemic vulnerability in high-seismicity areas. *Remote Sensing* 15(5). <https://doi.org/10.3390/rs15051453>.
- CHRR (Center for Hazards and Risk Research, Columbia University) and CIESIN (Center for International Earth Science Information Network, Columbia University). 2005. Global earthquake hazard frequency and distribution. Palisades, New York: NASA Socio-economic Data and Applications Center (SEDAC). <https://doi.org/10.7927/H4765C7S>.
- CIESIN (Center for International Earth Science Information Network). 2018. *Gridded population of the world, version 4.11 (GPWv4): Population count*. <https://doi.org/10.7927/H4JW8BX5>.
- Copernicus Emergency Management Service. n.d. *Damage assessment*. <https://emergency.copernicus.eu/mapping/book/export/html/138313>. Accessed 29 May 2024.
- Earle, P.S., D. Wald, K. Jaiswal, T. Allen, M. Hearne, K. Marano, A.J. Hotovec, and J. Fee. 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(2009): Article 15.
- Ehrlich, D., and G. Zeug. 2008. Assessing disaster risk of building stock: Methodology based on earth observation and geographical information systems. JRC Scientific and Technical Reports. Luxembourg: European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizen.
- Greenwell, B.M., B.C. Boehmke, and B. Gray. 2020. Variable importance plots—An introduction to the vip Package. *The R Journal* 12(1): 343–366.
- Guha-Sapir, D. 2023. *Emergency Events Database (EM-DAT) – CRED/OFDA International Disaster Database*. <https://public.emdat.be/>. Accessed 12 May 2024.
- He, C., Q. Huang, X. Bai, D.T. Robinson, P. Shi, Y. Dou, B. Zhao, and J. Yan et al. 2021. A global analysis of the relationship between urbanization and fatalities in earthquake-prone areas. *International Journal of Disaster Risk Science* 12(6): 805–820.
- Heath, D.C., D.J. Wald, C.B. Worden, E.M. Thompson, and G.M. Smoczyk. 2020. A global hybrid VS 30 map with a topographic slope-based default and regional map insets. *Earthquake Spectra* 36(3): 1570–1584.
- Hengl, T., G.B.M. Heuvelink, and D.G. Rossiter. 2007. About regression-kriging: From equations to case studies. *Computers & Geosciences* 33(10): 1301–1315.
- Jaiswal, K., and D. Wald. 2010. An empirical model for global earthquake fatality estimation. *Earthquake Spectra* 26(4): 1017–1037.
- LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521(7553): 436–444.
- Lee, J., J.Z. Xu, K. Sohn, W. Lu, D. Berthelot, I. Gur, P. Khaitan, K.-W. Huang, et al. 2020. *Assessing post-disaster damage from satellite imagery using semi-supervised learning techniques*. <https://doi.org/10.48550/arXiv.2011.14004>. eprint: 2011.14004.
- Li, X., P.M. Bürgi, W. Ma, H.Y. Noh, D.J. Wald, and S. Xu. 2023. DisasterNet: Causal Bayesian networks with normalizing flows for cascading hazards estimation from satellite imagery. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6–10 August 2023, Long Beach, CA, USA, 4391–4403.
- Loos, S., D. Lallemand, J. Baker, J. McCaughey, S.-H. Yun, N. Budhathoki, F. Khan, and R. Singh. 2020. G-DIF: A geospatial data integration framework to rapidly estimate post-earthquake damage. *Earthquake Spectra* 36(4): 1695–1718.
- Miura, H., S. Midorikawa, and M. Matsuoka. 2016. Building damage assessment using high-resolution satellite SAR images of the 2010 Haiti earthquake. *Earthquake Spectra* 32(1): 591–610.
- OpenStreetMap contributors. 2023. *Planet dump retrieved from https://planet.osm.org*. <https://www.openstreetmap.org>. Accessed 21 Apr 2024.
- Paulik, R., N. Horspool, R. Woods, N. Griffiths, T. Beale, C. Magill, A. Wild, and B. Popovich et al. 2022. RiskScope: A flexible multi-hazard risk modelling engine. *Natural Hazards* 119: 1573–1840.
- Rozelle, J.R. 2018. International adaptation of the HAZUS earthquake model using global exposure datasets. Ph.D. dissertation. University of Colorado Denver, Denver, CO, USA.
- Shultz, S. 2017. Accuracy of HAZUS general building stock data. *Natural Hazards Review*. [https://doi.org/10.1061/\(ASCE\)NH.1527-6996.0000258](https://doi.org/10.1061/(ASCE)NH.1527-6996.0000258).
- Silva, V., H. Crowley, M. Pagani, D. Monelli, and R. Pinho. 2014. Development of the OpenQuake engine, the Global Earthquake Model's open-source software for seismic risk assessment. *Natural Hazards* 72: 1409–1427.
- Silva, V., S. Akkar, J. Baker, P. Bazzurro, J.M. Castro, H. Crowley, M. Dolsek, and C. Galasso. 2019. Current challenges and future trends in analytical fragility and vulnerability modeling. *Earthquake Spectra* 35(4): 1927–1952.
- Smits, J., and I. Permanyer. 2019. The subnational human development index database. *Scientific Data* 6(1): 1–15.
- Svatonova, H. 2015. Aerial and satellite images in crisis management: Use and visual interpretation. In *Proceedings of the International Conference on Military Technologies (ICMT) 2015*, 19–21 May 2015, Brno, Czech Republic. <https://doi.org/10.1109/MILTECHS.2015.7153705>.
- UNDRR (United Nations Office for Disaster Risk Reduction). 2023. *DesInventar – Disaster Information Management System*. <https://www.desinventar.net/DesInventar/download.jsp>. Accessed 29 May 2024.
- University of California, Berkley. 2022. Global administrative areas version 4.1. <https://www.gadm.org/>. Accessed 13 Jan 2023.
- Wald, D.J. 2005. ShakeMap manual: Technical manual, user's guide, and software guide. *U.S. Geological Survey Techniques and Methods*. Reston, VA: U.S. Geological Survey.
- Wald, D.J. 2008. Quantifying and qualifying USGS ShakeMap uncertainty. Reston, VA: US Geological Survey.

- World Bank. 2022. *Population estimates and projections*. <https://datacatalog.worldbank.org/search/dataset/0037655/> Population-Estimates-and-Projections. Accessed 14 Feb 2023.
- Xia, Z., Z. Li, Y. Bai, J. Yu, and B. Adriano. 2022. Self-supervised learning for building damage assessment from large-scale xBD satellite imagery benchmark datasets. In *Database and expert systems applications*, ed. C. Strauss, A. Cuzzocrea, G. Kotsis, A.M. Tjoa, and I. Khalil, 373–386. Cham: Springer.
- Xu, S., J. Dimasaka, D.J. Wald, and H.Y. Noh. 2022. Seismic multi-hazard and impact estimation via causal inference from satellite imagery. *Nature Communications* 13(1): Article 7793.
- Xu, J.Z., W. Lu, Z. Li, P. Khaitan, and V. Zaytseva. 2019. *Building damage detection in satellite imagery using convolutional neural networks*. <https://doi.org/10.48550/arXiv.1910.06444>. eprint: 1910.06444.
- Yepes-Estrada, C., V. Silva, T. Rossetto, D. D'Ayala, I. Ioannou, A. Meslem, and H. Crowley. 2016. The global earthquake model physical vulnerability database. *Earthquake Spectra* 32(4): 2567–2585.
- Zhang, H., M. Wang, Y. Zhang, and G. Ma. 2022. TDA-Net: A novel transfer deep attention network for rapid response to building damage discovery. *Remote Sensing* 14(15): 3687.