Machine learning model for building seismic peak roof drift ratio assessment

Federico Mori, Daniele Spina, Flavio Bocchi, Amerigo Mendicelli, Giuseppe Naso & Massimiliano Moscatelli

Published online: 28 Feb 2023.

Submit your article to this journal ⬚

Article views: 664

View related articles ⬚

View Crossmark data ⬚

Machine learning model for building seismic peak roof drift ratio assessment

Federico Mori[a], Daniele Spina[b], Flavio Bocchi[b], Amerigo Mendicelli[a], Giuseppe Naso[b] and Massimiliano Moscatelli[a]

[a]Italian National Research Council (CNR) Institute of Environmental Geology and Geoengineering (IGAG), Consiglio Nazionale delle Ricerche, Monterotondo, Italy; [b]Civil Protection Department (DPC), Dipartimento della Protezione Civile, Rome, Italy

ABSTRACT

The peak roof drift ratio is one of the most important engineering parameters to describe the expected seismic damage in a building. A predictive model of the drift ratio was developed using a machine learning approach (Gaussian process regression model) on a dataset of approximately 11,800 records from 34 monitored buildings in Japan. Four predictors for ground motion and three predictors for building vulnerability are used in the machine-learning modelling. The residual analysis shows a reduction of 50% compared to the state of the art. The Gaussian process regression model is applied in a second analysis on an original dataset of approximately 4,500 records for 127 monitored buildings in Italy. A satisfactory comparison emerges by comparing the drift ratio prediction map with the observed damage pattern produced by satellite imagery for a test site in central Italy after the 2009 earthquake. The drift ratio map plays an important role in the simulation of an earthquake scenario at regional scale, which is needed by Civil Protection for emergency planning and management activities.

Introduction

Estimating the expected structural damage to buildings and infrastructure in the event of an earthquake is important for emergency planning. The peak roof Drift ratio (Dr), defined as the maximum relative displacement between the top and bottom of a building during an earthquake, normalized to the height of the building (Figure 1a, inset), is one of the most important engineering parameters for relating the seismic response to the damage level of a structure (Astorga et al. 2020). Dr prediction is therefore very important in earthquake emergency preparedness: a Dr map is a seismic scenario that indicates the areas where the greatest damage to buildings is expected, with consequent physical/economic losses and possibly obstruction of roads

CONTACT Federico Mori ✉ federico.mori@cnr.it, federico.mori@igag.cnr.it

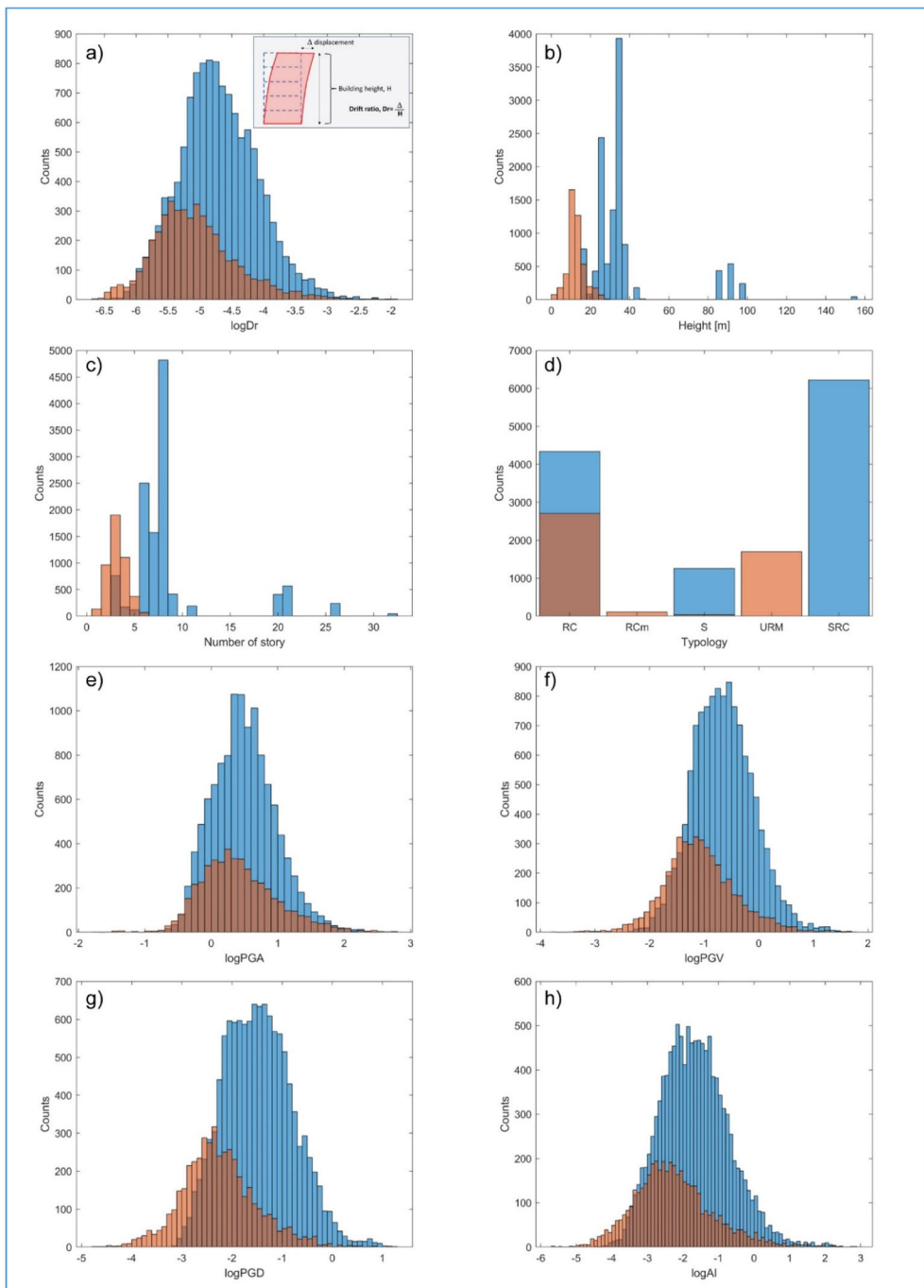**Figure 1.** Distribution of the response parameter Dr and the seven predictors in JapanDB (blue colour) and ItalyDB (brown colour): (a) logDr, (b) building height, (c) number of stories, (d) building typology, (e) logPGA, (f) logPGV, (g) logPGD, (h) logAI. It is important to note that some distribution classes of vulnerability predictors have little or no data. The significance of Dr is displayed in the inset of (a).

in case of collapse. However, at the planning stage, it is necessary to assess all possible levels of damage and not only the severe ones. In order to develop a model capable of predicting all levels of damage, it is necessary to analyse data containing different values of Dr. A machine learning approach was adopted using data from the Japanese literature and the original Italian data. This allowed an excellent prediction of Dr, as demonstrated for a reference seismic event in central Italy, through comparison with satellite data observations. The main use for the community of this predictive model is both the assessment of earthquake impacts on the whole built-up area and the stochastic estimation of losses for a portfolio of spatially distributed objects, which is fundamental to ensure the functioning of the earthquake emergency management system (Mori et al. 2020).

Data recorded on real structures describe the response of buildings to earthquakes more adequately than laboratory tests or numerical modelling. Therefore, it is important to build a prediction model based on a comprehensive database of earthquake records on real structures. Based on this premise, Astorga et al. (2020) compiled the NDE1.0 database containing 11,810 Japanese and 572 Californian building records with information on vulnerability, shaking and structural response. The parameters identified as representative of building vulnerability were overall height, number of stories and structural type. The parameters representative of ground motion or intensity measurements were: Peak Ground Acceleration (PGA), Velocity (PGV) and Displacement (PGD), Arias Intensity (AI), spectral values for pre-seismic (1) and co-seismic (2) fundamental frequencies (Sa1_2, Sv1_2, Sd1_2), mean spectral values between pre-seismic and co-seismic fundamental frequencies (AvgSa, AvgSv, AvgSd), Cumulative Absolute Velocity (CAV) and Destructive Potential (DP). The drift ratio (Dr) was taken as the structural response parameter.

Ghimire et al. (2021), starting from the original NDE1.0 database with the addition of the Romanian dataset, proposed an empirical Dr prediction equation that depends on the building class and considers one intensity measurement at a time. Iaccarino et al. (2021), using a machine learning approach, tested the performance of different Dr prediction models. Based on a dataset of almost 6,000 waveforms from sensors inside buildings registered in Japan and California, they used three different Earthquake Early Warning P-wave parameters, calculated considering three-time window lengths, for a total of nine predictors.

Our analysis was conducted first on the literature NDE1.0 Japanese dataset and then on the original Italian dataset published and described in this paper for the first time. The publication of this Italian dataset is very important because it allows a detailed analysis of the Italian building heritage that is obviously different from that of Japan or California, think for example of historic buildings and buildings of cultural significance. The data from California were excluded from the analysis because they were much less numerous than those from Japan (34 buildings, 11,810 records) and Italy (127 buildings, 4,500 records) and their performance with a machine learning approach was not robust. For the Japanese database, Ghimire et al. (2021) reported that the ground motion parameters that minimize the standard deviations of Dr's residuals are in order: Sv1, Sv2, PGD, PGV, IA and PGA. Unfortunately, since Sv1 and Sv2 cannot be calculated for the Italian database for now, only seven

parameters of the Japanese database were used for analysis and comparison: PGD, PGV, PGA, IA, overall height, number of floors and type of structure. We called this modified NDE1.0 database 'JapanDB'.

The Italian database collects data from the Seismic Observatory of Structures (OSS) of the Italian Civil Protection Department (Dolce et al. 2017). The OSS dataset consists of 1,389 seismic events recorded from 127 buildings, of which 81 are in Reinforced Concrete (RC), 38 are in UnReinforced Masonry (URM), 1 is in Steel (S) and 7 are mixed RC/URM (RCm). This database has been named 'ItalyDB'.

A preliminary analysis was conducted on JapanDB to determine which machine learning model performs best and whether this model improves the performance reported in the literature. The best model was then applied to ItalyDB by verifying its performance. The results were validated by comparison with the damage map observed at a test site affected by the 2009 earthquake in Central Italy.

In the preliminary analysis, several machine learning methods were tested. The results were compared by means of the Root Mean Square Error (RMSE), R-squared and residual analysis against each predictor. The best results were obtained using Gaussian Process Regression (GPR) with an exponential kernel model. The GPR model is a supervised probabilistic machine learning framework that has been widely used for regression and classification tasks. It is a non-parametric Bayesian approach to regression. A GPR model can make predictions by incorporating prior knowledge (kernels) and provide measures of uncertainty on predictions. The suggested reference for a complete description of the GPR method is Rasmussen and Williams (2006). In addition to providing the best results, the GPR model has the advantage of providing standard deviation and confidence intervals.

Because of these very qualities, it has already been used for regional seismic risk and consequence assessment (Park and Jung 2021; Sheibani and Ou 2021; Ahmad et al. 2022; Bodenmann et al. 2022; Ghimire et al. 2022).

To understand the order of importance of the predictors, the Ensemble bagged tree model[7] was used, which also gave good results.

Figure 1 shows the distributions of the response parameter Dr to be predicted and of the seven predictors used. In blue are the distributions of the JapanDB, and in brown are those of the ItalyDB. It can be clearly seen that the JapanDB refers to buildings of 6-9 stories on average with Steel Reinforced Concrete (SRC) and RC types, while the ItalyDB refers to URM (the most vulnerable) and RC buildings of 2-4 stories. It is also visible how the Japanese shaking predictors are on average higher than the Italian ones, reflecting the different seismological contexts of the two areas. The distribution of the Dr parameter in the JapanDB reflects this difference.

The objective of the present work is to develop a machine learning model for estimating the mean value and associated uncertainty of the drift ratio parameter that can be used in the near future for the assessment of expected losses at the regional scale after the occurrence of an earthquake or in prevention.

There are two key innovations:

- a unique dataset consisting of 127 buildings monitored in Italy over 20 years that have experienced earthquakes of all types (from the smallest to the largest and at

distances of all types) which can implement the NDE1.0 dataset found in the literature;

- the application of machine learning techniques that can provide the prediction of structural response at the regional scale by simultaneously integrating factors controlling the vulnerability of structures and factors controlling ground surface shaking.

Methods

The methods and tools for the reproducibility of the results are described in the following. The Regression Learner App (https://it.mathworks.com/help/stats/regression-learner-app.html) was used as a tool with the following workflow: import of the dataset and choice of the validation method, training and testing of models, comparison of model performance, and export of the model for the prediction on new data.

Opening a new session, the dataset is imported from the workspace, and the response variable (logDr) and the seven predictors (Height, NoOfStory, Typology, logPGA, logPGV, logPGD, logAI) are selected. The validation scheme is a Cross-Validation with 5 folds. The percentage for the test is 20%.

From the ML model selection menu, we opt for the exponential model from Gaussian process regression (GPR) models. A detailed description of GPR method terminology, development, and application is outside the scope of this article.

Briefly, a GP is a nonparametric method given a stochastic process considering $f(x)(x\ \varepsilon\ R^d)$, where $f(x_1), f(x_2), \ldots, f(x_n)$ is a multivariate Gaussian random variable for all combinations of input variables $(x_1, x_2, \ldots, x_n)$. Thus, the GPR model can be defined by introducing a mean function of the form $\mu(z) = E(f(x))$ and a covariance function of the form $k(x, x') = cov(f(x)), f(x_0))$. Consider our case in which the inputs x are a vector of the seven predictors, whereas y is the response variable logDr. Hence, the response variable can be modelled as

$$y(x) = h(x)^d\ \beta + f(x) +\ \epsilon \qquad (1)$$

where $h(x)$ is a vector of (deterministic) basis functions, β is a vector of basis function coefficients, $f(x)$ is a GP with zero mean and covariance function $k(x, x')$, and ε is Gaussian noise. The first term of equation 1 denotes the mean behaviour of the GP model. The GP term builds a nonlinear relationship between input and the response variable as well as elated uncertainties in the data.

Training data comprise input-output pairs such as $\{(x_i,\ y_i);\ i = 1,\ 2, \ldots, N\}$. Supposing $y_i'\ s$ are output of the considered model (i.e. $y_i = y(x_i)$), and $X = (x_1, x_2, \ldots, x_p)$ are inputs for which the predictions are calculated. Accordingly, $Y = (y_1, y_2, \ldots, y_N)$ and $Y = [y(x_1), y(x_2), \ldots, y(x_p)]$ are both Gaussian. The conditional distribution of Y' based on Y may be defined as

$$p(Y|Y') = N\left(\mu_{Y'} + \sum_{Y'Y}^{-1} \sum_Y^{-1} (Y - \mu_Y), \sum_{Y'} + \sum_{Y'Y} \sum_Y^{-1} \sum_{YY'}\right) \qquad (2)$$

where μ_Y and Σ_Y indicate the mean and covariance of Y, respectively, and Σ_{YY} is the cross-covariance of Y and Y'.

The means and covariances can be calculated by plugging in the model as

$$\mu_{Y'|Y} = h(z')^T \ \beta + k(X, X') \ (k(X, X) + \sigma_\epsilon^2 \ )^{-1} \ (Y - h(X)^T \beta \quad (3)$$

$$\sum\nolimits_{Y'|Y} = k(X', X') - k(X', X) \ (k(X, X) + \ \sigma_\epsilon^2 \ I)^{-1} \ k(X, X') \quad (4)$$

where h (X') and k (X', X) are simplified symbolizations for the vector and matrix with the components h (x_i) and k (x_i, x_j), respectively. Using the conditional mean function (equation 3), a prediction value for Y' can be computed along with a confidence approximation given by the conditional covariance (equation 4).

Choosing the functional form of the covariance (kernel) function (k (x, x')) is mostly based on assumptions about the main function to be modelled. In this study, the most widely used covariance function, namely, the 'exponential function', was used. This kernel function can be written as

$$k(x, x') = \ \sigma^2 \ e^{-(x-x')}/_\lambda \quad (5)$$

where σ^2 is the variance, and λ is the length scale for each input (hyperparameters).

To apply Gaussian processes in regression fitting, the hyperparameters of the selected covariance function must be optimized with respect to the experimental data. In this way, the MATLAB's *fitrgp* function estimates hyperparameters of θ (β, σ_ε^2, σ^2, λs) by minimizing the negative log-likelihood

$$\mathcal{L} \ (\Theta) = \ - \log \left(p(y|X, \Theta) = 0.5 \ y^T \ \sum\nolimits_\Theta^{-1} y + 0.5 \log \det \sum\nolimits_\theta + \frac{n}{2} \log 2\pi \quad (6)$$

where $\Sigma_\theta = k$ $(X, X; \theta) + \sigma_\varepsilon^2 I$. The default values of the parameters of the *fitrgp* command are chosen for optimizing the hyperparameter process in equation 6.

Results

The results of two analyses are presented.

In the first analysis, the machine learning approach was applied to the JapanDB and aims to:

- understand which machine learning model is best suited to the problem;
- assess whether the machine learning models produce a real improvement over the state of the art (Ghimire et al. 2021, Iaccarino et al. 2021) in terms of the standard deviation of residuals and R-squared.

In the second analysis, the best machine learning model identified for the JapanDB was applied to the ItalyDB to assess the following:

- whether the performance of the models is comparable to that of the first analysis (applying state-of-the-art methods and new method);

- the global ranking of the predictors;
- the residuals with respect to each predictor;
- the predictors of three representative buildings that suffered damage through Shapley's analysis.

Finally, a Dr prediction map is calculated for the 2009 L'Aquila (Italy) earthquake using the developed GPR model. This map was compared with a map obtained from an independent remote sensing method.

First analysis, applied to JapanDB

The JapanDB was analysed with 3 predictors of vulnerability (overall height, number of stories, and structural typology) and 4 predictors of ground motion intensity (PGD, PGV, PGA, IA).

The results were obtained by operating with 80% training and 20% testing; even when dividing the dataset by 60%/40%, the results are very similar. The models that performed best are the exponential GPR and the ensemble bagged tree as shown in Table 1). The GPR model shows a reduction in RMSE compared to previous studies. In fact, RMSE = 0.23 for the test leads to a reduction of approximately 70% compared to the results (0.76) shown by Ghimire et al. (2021) and approximately 50% compared (0.45) to Iaccarino et al. (2021). The R-squared values obtained with the GPR model (0.86) also improve on the performance reported in the literature (0.47 in Iaccarino et al. 2021). It is important to note that for values of Dr > 0.001 (log −3.0), the model predicts robustly.

Referring to the best prediction model (i.e. GPR with an exponential kernel), Figure 2a and 2b shows the performance of training and test datasets respectively referring to the response parameter Dr.

The results of all models are shown in Table 1.

Second analysis, applied to ItalyDB

As previously done for the JapanDB, the ItalyDB was analysed by operating with 80% training and 20% testing and using the same seven predictors. The two best models

Table 1. Results in terms of RMSE and Rsquared for the ML models used for both datasets in the 80% training-20% testing configuration.

| DATASET	Model	RMSE training	RMSE test	R-squared training	R-squared test
JapanDB	Exponential GPR	0.23	0.23	0.86	0.86
	Ensemble Bagged	0.23	0.23	0.86	0.86
	Fine Gaussian SVM	0.24	0.24	0.84	0.85
	Coarse Tree	0.25	0.25	0.83	0.83
	Ensemble Boosted	0.32	0.32	0.73	0.72
ItalyDB	Exponential GPR	0.22	0.22	0.88	0.88
	Ensemble Bagged	0.26	0.26	0.84	0.84
	Coarse Tree	0.3	0.3	0.8	0.8
	Fine Gaussian SVM	0.33	0.33	0.75	0.75
	Ensemble Boosted	0.34	0.34	0.73	0.73
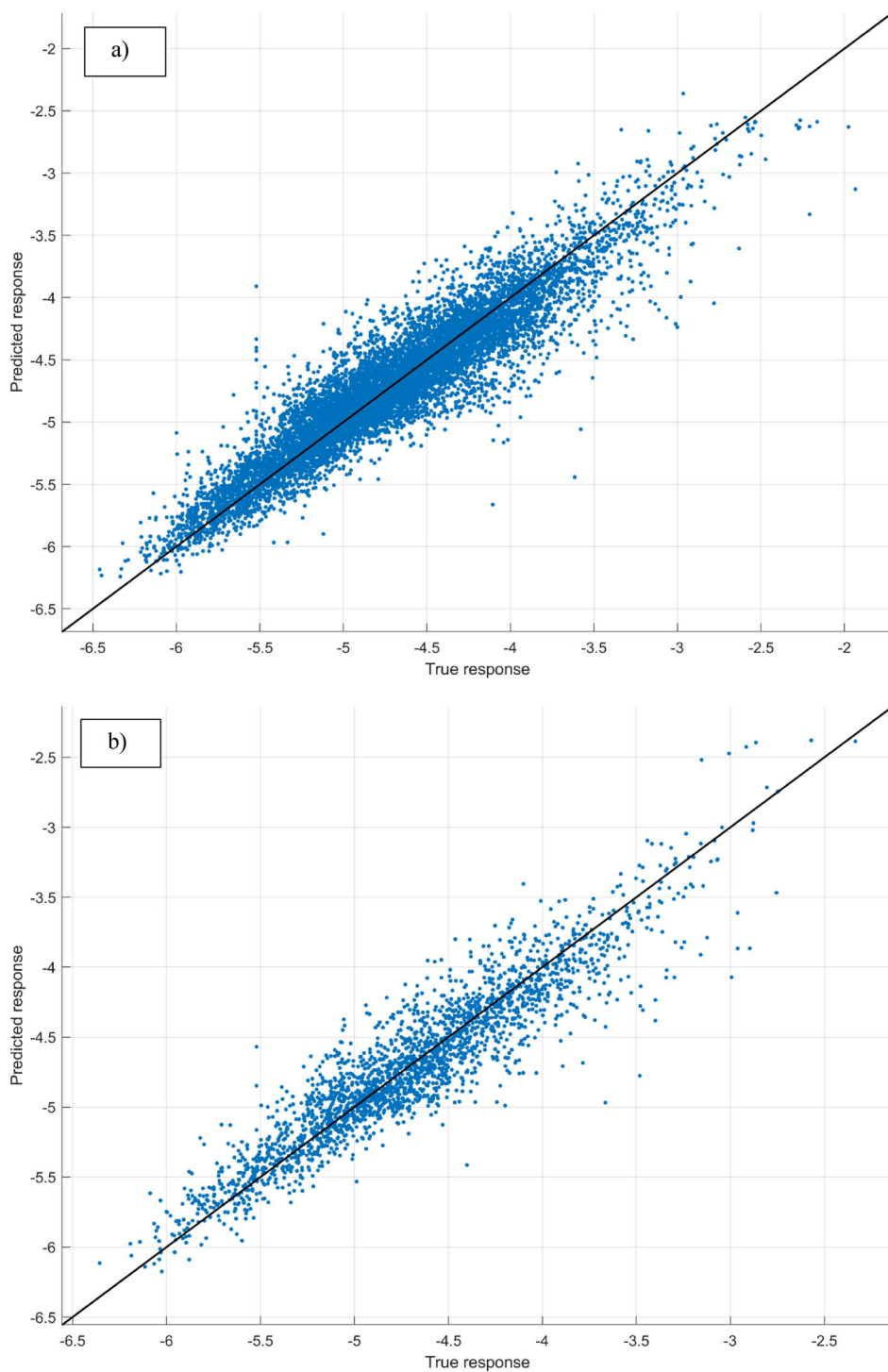
Figure 2. Performance of GPR model on the: (a) training dataset, (b) test dataset for JapanDB in terms of predicted Dr vs true Dr.
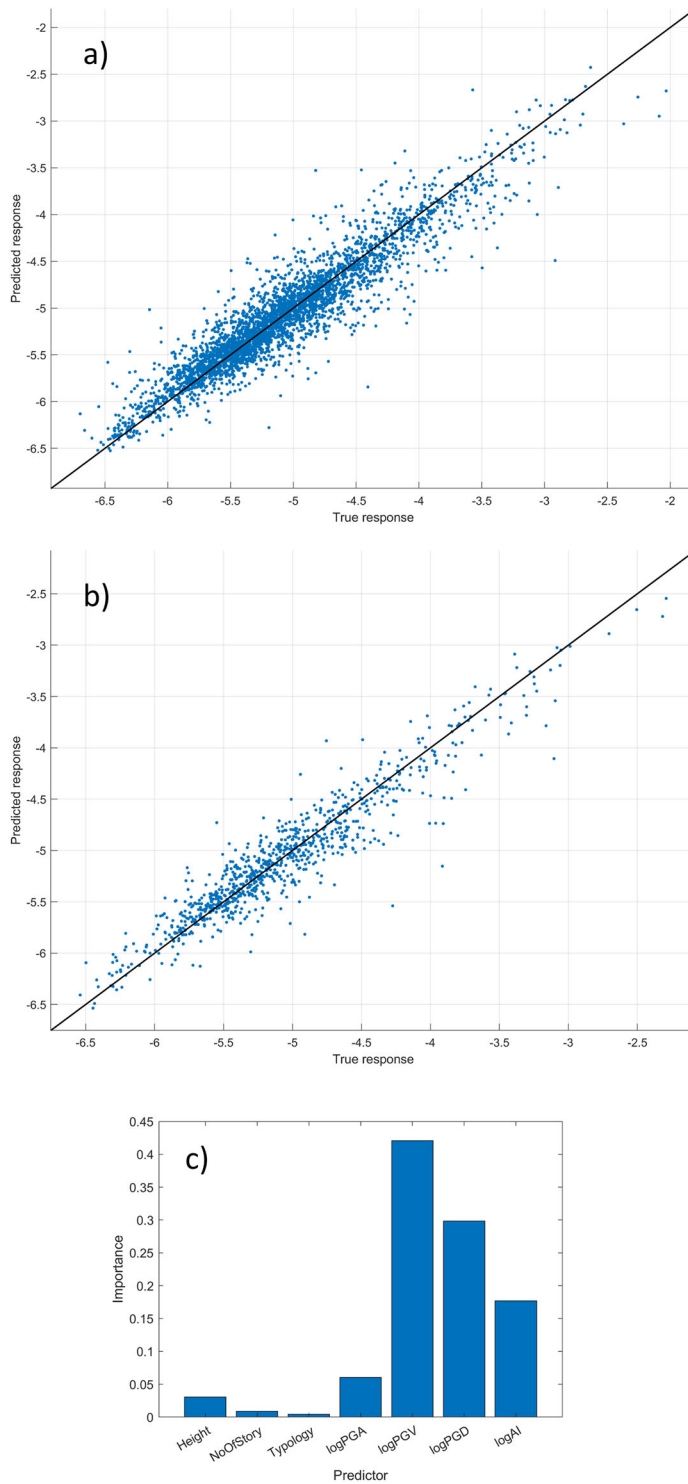
Figure 3. Performance of GPR model on the: (a) training dataset, (b) test dataset for ItalyDB in terms of predicted Dr vs true Dr, (c) ranking of predictors for ItalyDB.

Table 2. Main characteristics of the three analysed buildings that suffered damage (ID building code in the first column).

Building ID	Event Date	Mw	Epicentral distance [km]	Height [m]	# story	Typology	Log PGA	Log PGV	Log PGD	Log AI	Dr
15SNO	2016-10-30	6.5	4.39	9	3	'RC'	2.75	1.62	0.77	2.73	0.0031
BC037	2016-08-24	6	28.58	14	4	'URM'	2.50	1.41	0.35	1.89	0.0021
BC039	2016-10-30	6.5	46.59	10.7	2	'URM'	2.04	1.26	0.51	1.48	0.0020

The second, third and fourth columns refer to the parameters of the seismic event; the other columns refer to the predictors and the response parameter (Dr) of the buildings.

were used, the exponential GPR (Rasmussen ad Williams 2006) and the ensemble bagged tree (Breiman 1996). The performance is confirmed to be very good (RMSE = 0.22 and R-square = 0.88 for the test), and the model is also confirmed to be robust for values greater than −3.0 on a logarithmic scale. RMSE for the 5 folds of the cross validation are stable as they show variability from a low of 0.209 to a high of 0.228. Figure 3a and 3b show the prediction quality of the GPR model compared to the real data for both the training and testing datasets respectively for ItalyDB. The ranking of the predictors is shown in Figure 3c.

The most important predictors of Dr are logPGV, logPGD and logAI, while the predictor of ground motion with the poorest performance is logPGA. It is emphasized that all ground motion parameters are more important predictors than vulnerability parameters.

The family of ensemble boosted trees was also tested for this dataset, but it gave worse results (RMSE = 0.34, Rsquared = 0.73).The results of all models are shown in Table 1.

Residuals analysis for ItalyDB

The error residuals with respect to the individual predictors are illustrated inFigure 4 (a to g). The GPR model has residuals scattered symmetrically around 0, and there are no clear patterns. However, with regard to the 'typology' predictor, it can be seen that the highest residuals are in RC buildings.

Analysis of damaged individual buildings for ItalyDB

An additional analysis was performed to further investigate the importance of the predictors when considering high values of Dr, for which structural damage is expected. For this purpose, the right tail of the Dr distribution in Figure 1a (logDr > −3.0), which represents damaged buildings, was considered. Three buildings from ItalyDB were chosen to test this part of the distribution. The main characteristics of the buildings and the seismic events that affected them are shown in Table 2; a full description is also given in Cattari and Magenes (2022).

For the three buildings that suffered damage due to seismic events, Shapley (Lundberg and Lee 2017) analysis was used. The Shapley method works for both classification (when dealing with probabilities) and regression. The Shapley value of a feature for a query point explains the deviation of the prediction for the query point from the average prediction due to the feature. As shown in Figure 5, the analysis
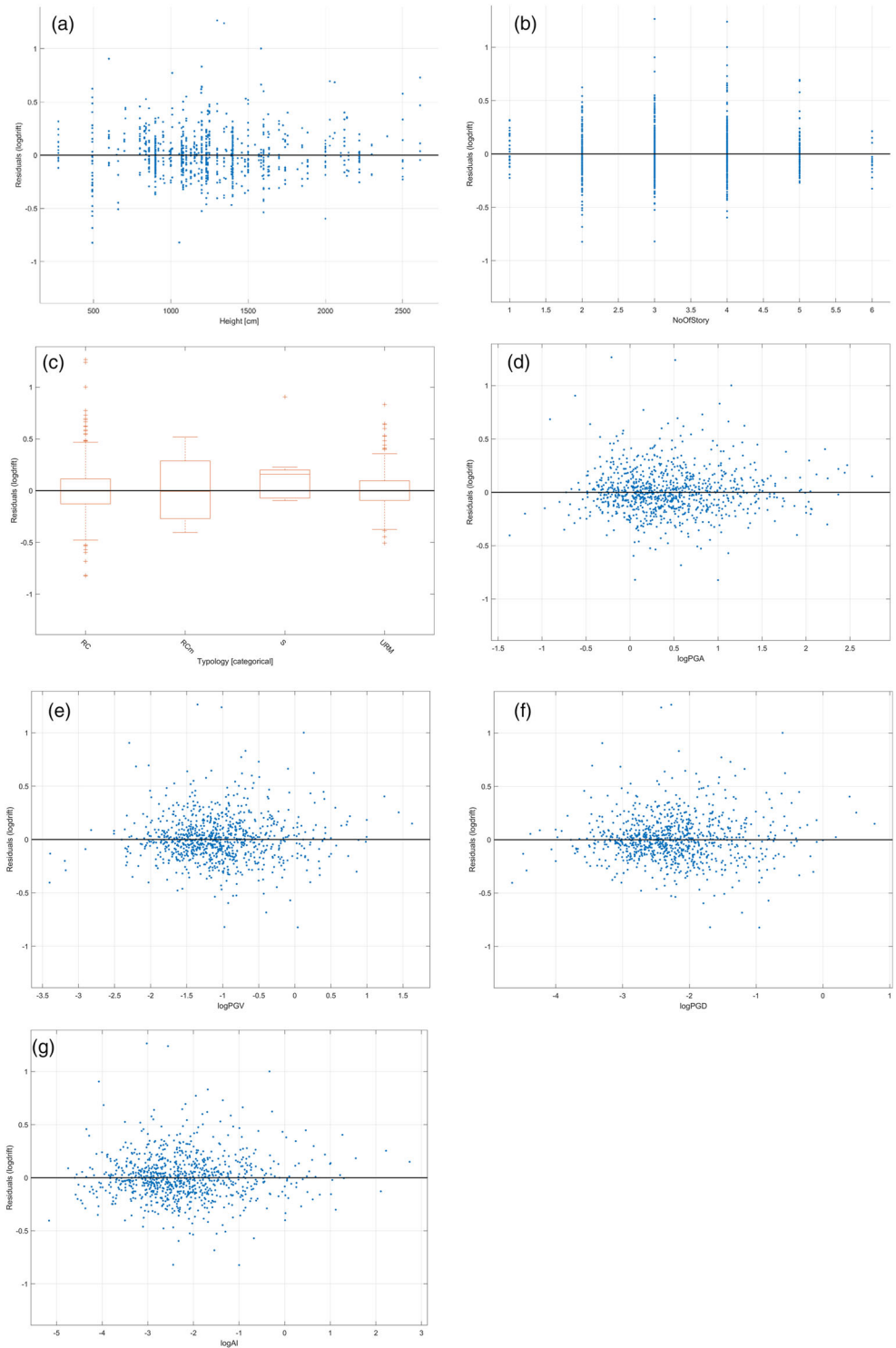
Figure 4. Residuals vs. predictors for the ItalyDB test: (a) height; (b) number of stories; (c) typology; (d) logPGA; (e) logPGV; (f) logPGD; (g) logAI. The analysis generally shows an absence of trend, and the residuals are symmetrically distributed around 0.
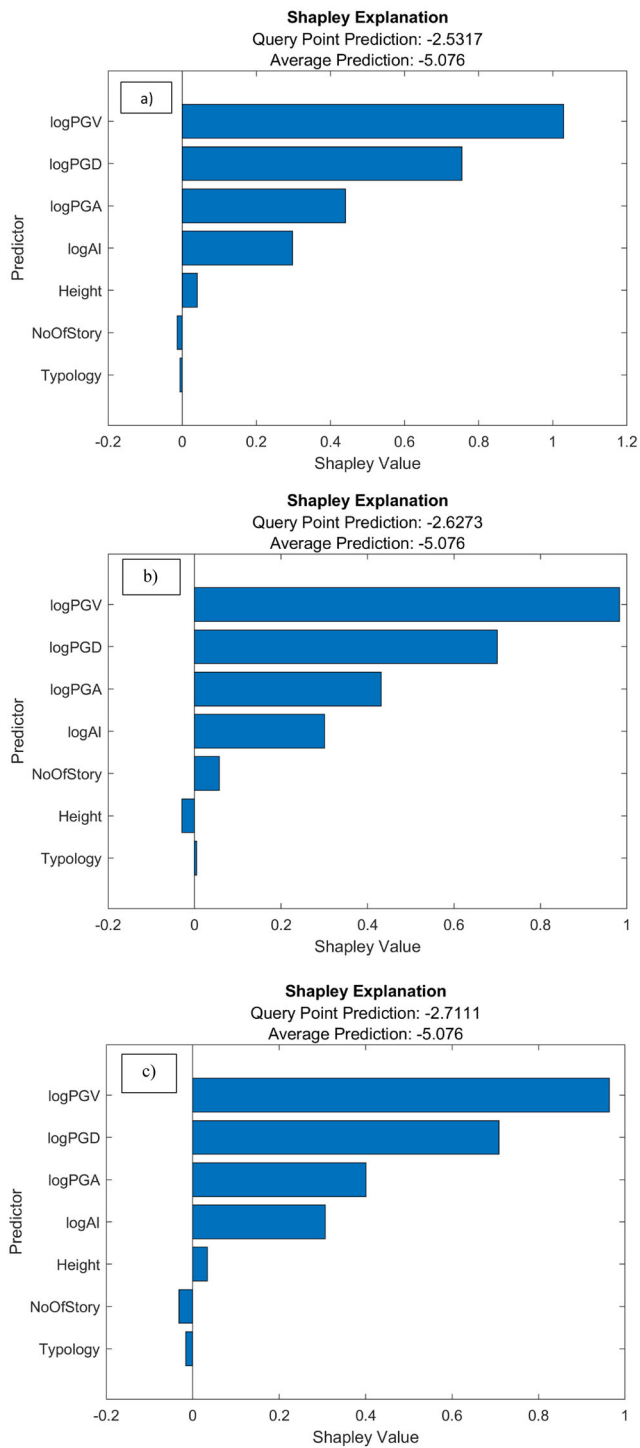
Figure 5. Shapley values of the predictors for the three buildings of Table 2: a) 15SNO, b) BC037, c) BC039.
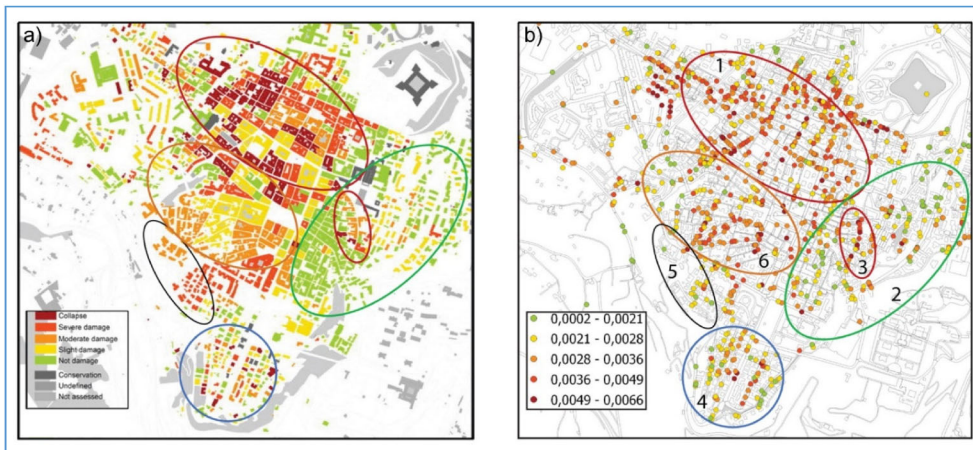
Figure 6. (a) Remote sensing damage map (slightly modified and reprinted from Int. Journ. Disaster Risk Reduct., 125-142 (2014), Contreras D., Blaschke T., Kienberger S., & Zeil P., Myths and realities about the recovery of L'Aquila after the earthquake, with permission from Elsevier) following the 6th April 2009 earthquake for the historic centre of L'Aquila (Italy) *vs* b) prediction Dr map with GPR exponential model; base map from http://opendata.regione.abruzzo.it/. In b) the colored dot represents the presence of the damage information in the post-earthquake survey. Not all buildings have been studied post-earthquake.

confirms that the two most important predictors are logPGV and logPGD, in agreement with the ranking in Figure 3c for the ItalyDB dataset.

Dr prediction map for the l'Aquila test site (Italy)

After observing the robust performance of the GPR exponential model, a real application was carried out to test the method in a scenario mode:

1. select magnitude and hypocentral coordinates of the 2009 L'Aquila earthquake (Central Italy) from Italian accelerometric database (Russo et al. 2022);
2. prediction of the surface ground motion parameters (PGA, PGV, PGD, AI) according to Mori et al. (2022);
3. select the building locations and vulnerability parameters (total height, number of floors, building type) from the Da.D.O. database (Dolce et al. 2019) only for the buildings with post-earthquake damage surveys;
4. prediction of Dr-related scenario with GPR model from ItalyDB.

A first qualitative application was carried out by comparing the Dr prediction map produced by GPR model with a spatial damage model elaborated with remotely sensed data of Figure 6 (Contreras et al. 2014). As already written, the comparison is only qualitative, but the authors decided to show it to clarify what could be a possible application of the work. In this key one should read the figure. A direct quantitative comparison between the two maps is not possible because they refer to different entities, i.e. damage and drift ratio. However, the good agreement between the areas with
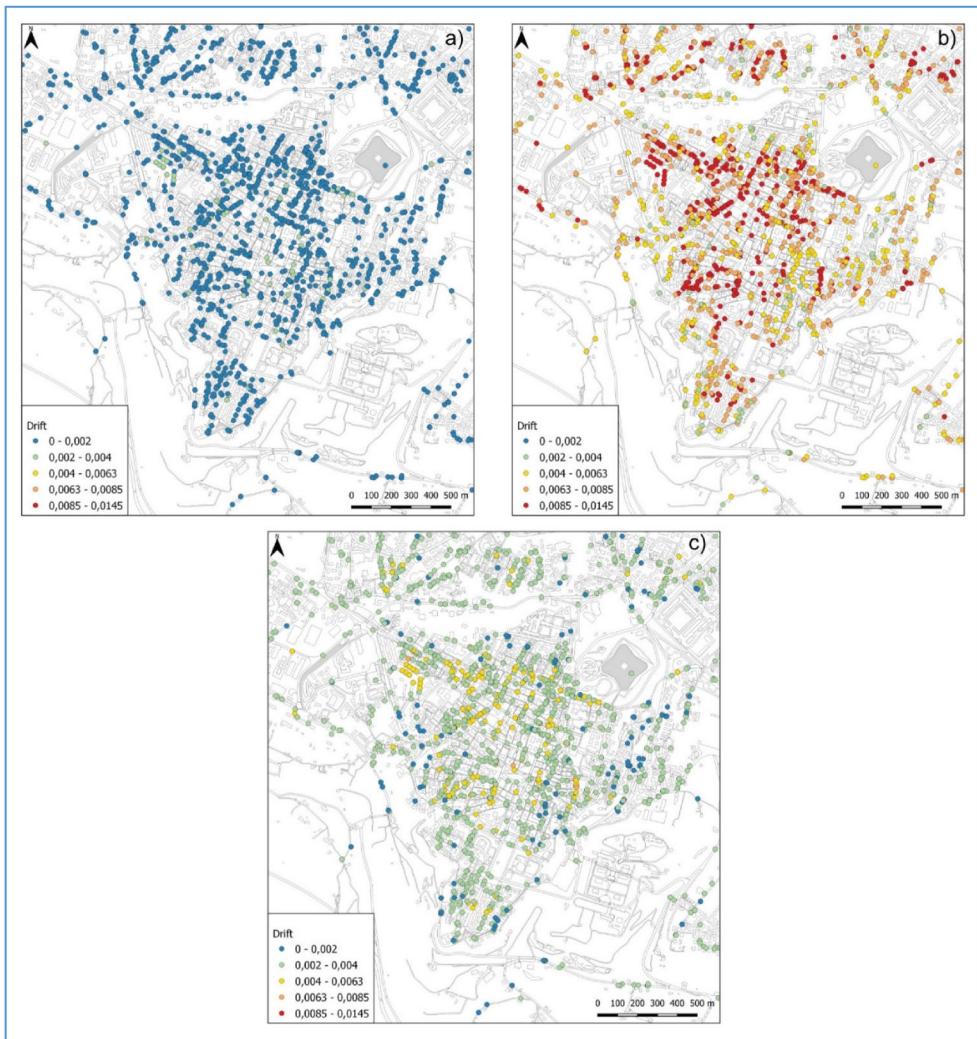
Figure 7. a) lower bound of the 68% confidence interval of Dr prediction map relative to Figure 6b, (b) upper bound of the 68% confidence interval of Dr prediction map relative to Figure 6b, (c) Dr prediction map obtained with GPR exponential model and shown in Figure 6b.

the highest observed damage and those with the highest Dr confirms that the proposed estimation model is useful for prevention purposes.

In particular, the GPR model map shows a short-range variability in terms of Dr patterns that reproduces well the variability of the damage pattern reconstructed from satellite images. In Figure 6, macro-areas with homogeneous damage levels are highlighted and numbered. Area 1 is characterized by collapses (high damage) and high Dr values. Area 2 has relatively less damage and Dr, except for a long stretch highlighted in area 3. In area 4, there is great variability in damage and Dr. The GPR model fails to predict in area 5, probably because the observed damage is due to strong topographic amplification phenomena that the ground motion model (Mori et al. 2022) used to derive shaking predictors was apparently unable to grasp. In area 6, there is medium to severe damage, as confirmed by map Dr. The confidence

interval for the Dr map is given in the Figure 7 (model uncertainties, on the other hand, are in the analysis of the residuals, RMSE). The transition from expected value of drift ratio to damage is not the subject of this work but of future developments that will lead to mappings with probabilistic distributions of degrees of damage.

Discussion

A machine learning GPR model with an exponential kernel was trained and tested on records from a Japanese building database for the prediction of Dr. The proposed model performed very well, and the standard deviation of the residuals improved compared to the literature. The same model was applied to the Italian database of buildings, again obtaining a robust prediction of the Dr response parameter.

In this work for the first time the Italian OSS database (20 years of records on 127 buildings) has been used for these purposes and the first time a machine learning approach has been used to predict Dr values.

The proposed GPR model is able to predict Dr from 4 predictors related to ground motion intensity (PGV, PGD, PGA, AI) recorded at the base of the building and thus including site effects and 3 predictors related to vulnerability (total height, number of floors, building type). The standard deviation of the residuals results in a 50% improvement over the state of the art. Furthermore, for high values of Dr leading to the occurrence of damage, the model is still robust.

The results show that predictors describing ground motion are always more important than those relating to vulnerability. This is not unexpected because the latter are categorical descriptors (attributes) of the building and do not describe the physical-mechanical and geometric characteristics with which it was built (Luo and Paal 2019). According to the authors, this is the biggest problem for this type of analysis, and unfortunately, the solution of replacing the vulnerability predictors with others that describe the strength of the building materials is not adoptable today due to the enormous lack of data.

The degree of importance of the predictors was studied both with the entire Italian database (weights calculated using the ensemble bagged tree method) and with three buildings in the database that suffered damage during seismic events using the Shapley method.

In the first case, the ranking of predictors (of ground motion) is PGV, PGD, AI, PGA, while in the second case, it is PGV, PGD, PGA, AI. These results show that PGA is the ground motion parameter least correlated with Dr (and therefore with damage), as discussed extensively in many other works (Vacca et al. 2022), despite being the most widely used parameter for damage estimation in earthquake engineering (i.e. the parameter most used to construct fragility curves for risk analysis). It is interesting to note that many damage indices proposed in the literature are linked to the number of plastic cycles and thus to the earthquake's energy content; among others, Manfredi and Cosenza (2000) proposed damage indices that are functions of IA, PGV and PGA.

Shapley's analysis conducted on three damaged buildings shows that the joint use of PGV and PGD parameters is essential to estimate Dr. Conversely, the analysis shows that vulnerability parameters have less weight in predicting Dr.

The method was also tested by comparing it with the satellite damage maps available for L'Aquila (Central Italy) after the 2009 earthquake. The predicted Dr map adequately reproduces the damage levels and patterns observed in the town.

In terms of application, the Dr map contributes significantly to the assessment of an earthquake scenario because it is a good proxy for damage. Communities living in earthquake-prone areas would undoubtedly benefit from it. As a tool for emergency planning, Dr maps could also be very useful to improve the estimation of the efficiency of the emergency management system (Mori et al. 2020). The emergency management system is a complex network consisting of structural components (such as buildings, infrastructure and emergency areas) and non-structural elements (such as procedures, emergency workers, and information management). The basis of the design of this system is the calculation of an appropriate damage scenario to calibrate human and instrumental resources for a prompt and effective civil protection activity. Furthermore, starting from near real-time seismological information to calculate shaking parameters (Mori et al. 2022), Dr maps could be useful to estimate the impact on structures in the first minutes after an earthquake. Dr maps could be used in conjunction with the Copernicus Emergency Management Service's EMS maps (Cotrufo et al. 2018) and NASA ARIA Team JPL's Damage Proxy Maps (Yun et al. 2015), which are produced several hours to several days after the event, depending on the availability of satellite imagery.

Comforted by the robustness of the dataset and results, in the future, we plan to:

- standardize the Italian database as the NDE1.0 database, implementing intensity measurements related to building pre and co-seismic frequencies;
- establish a semi-automatic procedure for the elaboration of Dr maps for both stochastic probabilistic scenario analyses and near-real-time forecasts for civil protection purposes;
- moving from the prediction of expected drift ratio to the probabilistic distribution of degrees of damage also working with the prediction of the Maximum Interstory Drift Ratio;
- integrate age of construction and vulnerability characteristics more representative of the material and geometric properties of buildings; in Italy, this is very difficult to achieve because this type of data is dispersed in numerous databases.

These integrations and refinements of the methodology are of fundamental importance for improving safety policies and therefore potentially for all citizens in the event of seismic events. This method can also integrate the large amount of data available from dynamic soil-structure monitoring in order to build a robust and comprehensive data-driven analysis programme.

Finally, the predictive model proposed here is an example of how artificial intelligence can increase the ability to predict the impacts in urban areas.

Author contributions

F.M. developed the proposed method. D.S. supervised the method and the results. F.B. prepared the dataset, A.M. prepared the figures, G.N. supervised the method and the results, M.M. supervised the research. All wrote.

Competing interests

The authors declare no competing interests.

Disclosure statement

No potential conflict of interest was reported by the authors.

Data availability

The dataset is provided in the file 'datasets_rev.xls' with the following specifications: JapanDB for NDE1.0 Japanese dataset, ItalyDB for the original dataset. JapanDB is modified from NDE1.0 flatfile downloaded from https://www.isterre.fr/annuaire/pages-web-du-personnel/philippe-gueguen/new-earthquake-data-recorded-in-buildings-nde1-0/

References

Ahmad M, Amjad M, Al-Mansob RA, Kamiński P, Olczak P, Khan BJ, Alguno AC. 2022. Prediction of liquefaction-induced lateral displacements using gaussian process regression. Appl Sci. 12(4):1977.

Astorga A, Guéguen P, Ghimire S, Kashima T. 2020. NDE1.0: a new database of earthquake data recordings from buildings for engineering applications. Bull Earthquake Eng. 18(4): 1321–1344

Bodenmann L, Reuland Y, Stojadinovic B. 2022. Using gaussian process models for dynamic post-earthquake impact estimation with regional risk predictors. [Preprint].

Breiman L. 1996. Bagging predictors. Mach Learn. 26:123–140. doi:10.1007/BF00058655.

Cattari S, Magenes G. 2022. Benchmarking the software packages to model and assess the seismic response of unreinforced masonry existing buildings through nonlinear static analyses. Bull Earthquake Eng. 20:1901–1936. doi:10.1007/s10518-021-01078-0.

Contreras D, Blaschke T, Kienberger S, Zeil P. 2014. Myths and realities about the recovery of L'Aquila after the earthquake. Int J Disaster Risk Reduct. 8:125–142. doi:10.1016/j.ijdrr.2014.02.001.

Cotrufo S, Sandu C, Tonolo FG, Boccardo P. 2018. Building damage assessment scale tailored to remote sensing vertical imagery. Eur J Remote Sensing. 51(1):991–1005. doi:10.1080/22797254.2018.1527662.

Dolce M, Nicoletti M, De Sortis A, Marchesini S, Spina D, Talanas F. 2017. Osservatorio sismico delle strutture: the Italian structural seismic monitoring network. Bull Earthquake Eng. 15(2):621–641 doi:10.1007/s10518-015-9738-x.

Dolce M, Speranza E, Giordano F, Borzi B, Bocchi F, Conte C, Di Meo A, Faravelli M, Pascale V. 2019. Observed damage database of past Italian earthquakes: the Da. DO WebGIS. Boll Geofis Teorica Appl. 60(2):141–164. doi:10.4430/bgta0254.

Ghimire S, Guéguen P, Astorga A. 2021. Analysis of the efficiency of intensity measures from real earthquake data recorded in buildings. Soil Dyn Earthquake Eng. 147:106751. doi:10.1016/j.soildyn.2021.106751.

Ghimire S, Guéguen P, Giffard-Roisin S, Schorlemmer D. 2022. Testing machine learning models for seismic damage prediction at a regional scale using building-damage dataset compiled after the 2015 Gorkha Nepal earthquake. Earthquake Spectra. 38(4):2970–2993.

Iaccarino A, Gueguen P, Picozzi M, Ghimire S. 2021. Earthquake early warning system for structural drift prediction using machine learning and linear regressors. Front Earth Sci. 9: 666444. doi:10.3389/feart.2021.666444.

Lundberg SM, Lee S. 2017. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. Vol. 30; p. 4765–4774. Curran Associates. doi:10.48550/arXiv.1705.07874.

Luo H, Paal SG. 2019. A locally weighted machine learning model for generalized prediction of drift capacity in seismic vulnerability assessments. Comput-Aided Civ Infrastruct Eng. 34(11):935–950. doi:10.1111/mice.12456.

Manfredi G, Cosenza E. 2000. Damage indices and damage measures. Prog Struct Mater Eng. 2(1):50–59. doi:10.1002/(SICI)1528-2716(200001/03)2:1<50::AID-PSE7>3.0.CO;2-S.

Mori F, Gena A, Mendicelli A, Naso G, Spina D. 2020. Seismic emergency system evaluation: the role of seismic hazard and local effects. Eng Geol. 270:105587. doi:10.1016/j.enggeo.2020.105587.

Mori F, Mendicelli A, Falcone G, Acunzo G, Spacagna RL, Naso G, Moscatelli M. 2022. Ground motion prediction maps using seismic-microzonation and machine learning. Nat Hazards Earth Syst Sci. 22(3):947–966. doi:10.1016/j.enggeo.2020.105587.

Park S, Jung K. 2021. Gaussian process regression-based structural response model and its application to regional damage assessment. ISPRS Int J Geo-Inf. 10(9):574. doi:10.3390/ijgi10090574.

Rasmussen CE, Williams CKI. 2006. Gaussian processes for machine learning. Int J Neural Syst. 14(2).

Russo E, Felicetta C, D, Amico M, Sgobba S, Lanzano G, Mascandola C, Pacor F, Luzi L. 2022. Italian accelerometric archive v3.2. Istituto Nazionale di Geofisica e Vulcanologia (INGV). doi:10.13127/itaca.3.2.

Sheibani M, Ou G. 2021. The development of Gaussian process regression for effective regional post-earthquake building damage inference. Comput Aided Civ Inf. 36:264– 288. doi:10.1111/mice.12630.

Vacca V, Occhipinti G, Mori F, Spina D. 2022. The use of SMAV model for computing fragility curves. Buildings. 12. doi:10.3390/buildings12081213.

Yun SH, Hudnut K, Owen S, Webb F, Simons M, Sacco P, Gurrola E, Manipon G, Liang C, Fielding E, et al. 2015. Rapid damage mapping for the 2015 Mw 7.8 Gorkha Earthquake Using synthetic aperture radar data from COSMO-SkyMed and ALOS-2 satellites. Seismol Res Lett. 86(6):1549–1556. doi:10.1785/0220150152.