COMPENSATING THE LACK OF BIG DATA IN CONSTRUCTION INDUSTRY WITH EXPERT KNOWLEDGE: A CASE STUDY

Zoran Stojadinović

Faculty of Civil Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia
e-mail: joka@grf.bg.ac.rs

Abstract:

Due to various reasons, there is a lack of big data in the construction industry, one of the main obstacles to a broader implementation of AI. Another obstacle is adhering to analytical methods in fields more suitable for AI solutions. If appropriately used, multidisciplinary expert knowledge can compensate for these problems and enhance the application of AI techniques in construction. The case study refers to rapid earthquake loss assessment. The problem with traditional systems is their low accuracy, making them unreliable and unusable in the recovery process, which is the purpose of loss assessment systems. Low accuracy is caused by too much uncertainty in analytical and insufficient data sets to create vulnerability curves in empirical methods. The contribution of this research is designing a new kind of rapid earthquake loss assessment system using multidisciplinary expert knowledge and AI methods. The problem of small data sets was solved using the procedure of representative sampling, which makes a small sample informative and sufficient to use. The low accuracy of analytical methods is caused by assuming *theoretical* vulnerability relations *before* an earthquake. The new approach uses trained assessors to perform on-the-ground observation of *actual* damage on the representative sample *after* an earthquake. AI methods are then used to predict damage to the remaining building portfolio, which is more accurate and still rapid enough. Another contribution is using a building representation without earthquake data which eliminates the need for analytical methods, shake maps and robust ground motion sensor networks, making the proposed framework unique and applicable in any region.

Keywords: big data, expert knowledge, machine learning, representative sampling

1. Introduction

The construction industry lags in implementing AI in scientific research. Some specific problems lead to that. Due to various reasons, there is a lack of big data (big enough data sets to implement AI techniques) in the construction industry, one of the main obstacles to a broader implementation of AI. Typical realistically obtained data sets are: 30 highway sections contracts, 1797 earthquake damaged buildings, or 38 residential buildings with actual financial data. The most common reasons for the lack of data are data ownership, the sensitive nature of financial information, and incomplete data sets. Such small or medium-sized data sets are insufficient for AI-based scientific research and limit the application of different AI techniques in the construction industry. It is not uncommon that researchers sometimes give up on ideas when faced with small data sets. In addition, researchers sometimes adhere to traditional engineering solutions in fields where AI could outperform analytical methods.

When it comes to implementing AI in construction, there is another phenomenon worth mentioning. There are sources of larger data sets of which researchers are still not aware. For instance: real-time spatial labor data from construction sites, minutes of meetings (text mining), or digital twin-enabled evaluations. The scientific community should find ways to

use these IT-enabled big data sources. This intriguing topic is outside the scope of this research.

This paper presents a case study that shows a way forward for overcoming two obstacles - the lack of big data (data sets undersized for machine learning techniques) and adhering to underperforming analytical methods (which can be outperformed by AI solutions). The case study refers to rapid earthquake loss assessment. The purpose of any loss assessment system is to be accurate enough to be used for budgeting the recovery process. Due to establishing *theoretical* vulnerability relations *before* an earthquake, the accuracy of traditional systems is low. The research goal is to overcome the accuracy **problem** by using a completely different approach – establishing vulnerability relations *after* an earthquake by observing *actual* damage. If the new system is accurate and rapid enough, it would be a significant scientific and practical **contribution**.

2. The traditional approach to rapid earthquake loss assessment and what's wrong with it?

When an earthquake occurs, buildings are damaged, and society strives to recover as soon as possible. The first step in a recovery process is determining the damage to each building. On-the-ground surveys have to be conducted to record and verify damage officially, but it takes a long time to visit every building – usually a couple of months. So, a rapid assessment (near real-time prediction of damage and loss) is needed to start planning recovery. The primary goals of a rapid system are to assess safety and occupancy issues (detect higher levels of damage) and to monetize loss at the community level (so the local authorities can plan the budget and other resources). The accuracy and speed of prediction are the main quality features of rapid loss assessment systems.

Rapid earthquake loss assessment consists of two phases: preparing the system (pre-earthquake phase) and activating the system when an earthquake happens to predict loss (co-earthquake phase).

2.1 Pre-earthquake phase – preparing the system before an earthquake

The preparation of traditional loss assessment systems comprises five steps:

1. Creating the building portfolio - a database of buildings in a city or region. The database is populated with characteristics of the building (e.g. building type, age, and geometry features) and its location (e.g. geo coordinates and soil type). Classifying the portfolio of buildings into building types (BT) depends on the materials, construction methods, structural elements, and other factors influencing their seismic behavior.

The problem with this step is that many buildings don't fit the exact building type due to design modifications, improvisations, additions, or omissions. It is not always easy to determine the actual BT, so mislabeling can occur.

2. Defining damage states (DS) for buildings. Various classifications exist globally to describe the severity of the damage, ranging from slight damage to collapse.

The problem with this step is that actual damage states are not clear cut and often coexist in a building. Also, DS does not contain information about the quantity of damage, which is critical for budgeting the recovery process.

3. Establishing a Ground Motion Model - to estimate ground motion intensities in the affected region and map the effects of an earthquake on buildings locations. The estimated distribution of the ground motion intensity field, updated with the observations recorded during the event, is called a shake-map. It shows intensity measures such as Peak Ground Acceleration, Spectral Acceleration, Peak Ground Velocity, or macroseismic intensity. Hence, generating shake maps is a procedure that maps the intensity measures of an earthquake to specific building locations. The most common algorithms used to generate shake-maps are the USGS ShakeMap® algorithms [1] and the Bayesian inference method [2]. In the pre-earthquake phase, a region has to build a network of ground motion sensors and establish an organizational unit to be prepared to use the system.

The problem with a ground motion model is that it introduces significant uncertainty and requires building a network of ground motion sensors and establishing an organizational unit to be prepared to use the system.

4. Damage prediction - determining the relationship between ground motion and damage states for each building type. Damage prediction involves methods for creating seismic vulnerability models (vulnerability curves) for different building types independently of the earthquake under assessment. Prediction methods can be analytical, empirical, and hybrid [3]. For a particular earthquake and observed ground motion, the system predicts the probability of a building type being in a certain damage state.

4.1 In analytical methods, the assessment of expected damage states is based on dynamic modelling and analysis [4]. Analytical methods are capacity spectrum based, collapse mechanism-based, or fully displacement-based. The strength of analytical methods is the provable and quantifiable accuracy of seismic damage prediction for a given, well-defined building type. However, the actual building stock is quite diverse and numerous, such that it cannot be easily classified and separately investigated. As-built buildings often differ from theoretical building types. In addition, their seismic performance capacity diminishes due to ageing or poor maintenance. Thus, there is significant uncertainty in applying the analytical approach to a diverse building portfolio, making for less precise earthquake loss assessment systems.

Machine learning (ML) methods are also used in research to create vulnerability relations but with no significant scientific or practical added value other than to verify already explored methods. Using AI this way does not solve implementation issues on a diverse portfolio.

4.2 In empirical approaches, the evaluation of damage state probabilities for each building type is based on the observed damage from previous earthquakes. The outcomes are presented in the form of damage probability matrices or continuous vulnerability curves fitted to the data [3]. The weak side of this approach, when implemented for a particular locality, is the lack of a sufficiently large set of reliable empirical data due to the limited number of damaging earthquakes affecting the locality in the recent past [5]. Data from a single earthquake has a limited range of intensities, thus producing just a part of the vulnerability curve. As in analytical methods, combining data from different affected areas to develop vulnerability relations introduces significant uncertainty concerning the equivalence of the related building types – the ones used to create vulnerability models and the ones to be assessed for damage after a future earthquake. Typical vulnerability curves derived from empirical data are shown in Fig.1.

5. Loss quantification is essential for the successful management of the recovery process because it provides the local authorities with a first estimate of the needed budget. Loss quantification usually implies using matrices with repair or replacement costs per unit area of a building, covering all possible combinations of building type and building damage state. Experts make these cost estimates considering the current construction technologies and costs. Typically, such cost estimates are normalized by the replacement value of the building [6], [7]. If not updated regularly, unit repair cost matrices may introduce significant uncertainties regarding both current repair costs and building replacement values. Furthermore, market conditions in different regions of the world may lead to very different ratios of the market value of an existing building, which may be in a poor state of function and maintenance, and the market cost to build an equivalent new building in its place. There is a need to include construction management experts to upgrade the current practice of loss quantification.

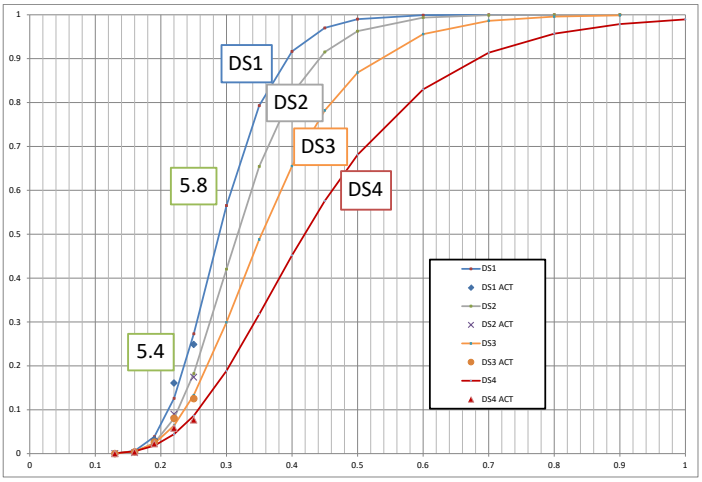

Fig.1. Vulnerability curves derived from empirical data (Kraljevo 2010 earthquake)

2.2 Co-earthquake phase - activating the system immediately after an earthquake

Since all components are set in advance, the co-earthquake phase is simple to perform. When an earthquake occurs, sensors detect ground motions, shake maps are generated and, using vulnerability relations, probabilities for damage states are estimated for each building in near real-time. Loss is computed according to pre-determined relations and replacement values. Based on different combinations of shake-map algorithms, damage prediction software, and methods for loss assessment, various local or global operational rapid earthquake loss assessment systems exist worldwide [4], [8].

Therefore, the existing loss assessment systems introduce uncertainties in all three stages: in generating shake maps (both in algorithms and in measurements of ground motion), in fragility/vulnerability relations (approximating actual buildings to theoretical models or empirical data gathered elsewhere), and in the loss assessment (both in repair costs and replacement values). According to HAZUS, the total uncertainty is “possibly at best a factor of two or more”, motivating researchers to investigate new approaches to create better rapid earthquake loss assessment systems.

Although each step is logical and scientifically justified, the system as a whole is not accurate enough. The conclusion is that two problems hamper the traditional approach to loss assessment: analytical methods cannot be accurate because of too many assumptions and approximations, and empirical methods cannot work because of the lack of a sufficiently large set of reliable data.

Earthquake engineers are stuck – they don’t want to let go of analytical and don’t have a solution for the lack of big enough samples for empirical methods.

3. New approach to rapid earthquake loss assessment - based on representative sampling and an appropriate building representation

This case study explains how to overcome the weaknesses of traditional approaches to damage predictions. When faced with uncertainties and limiting small data sets, researchers need to look for multidisciplinary expert knowledge and make higher-level abstractions to enable ML techniques and eliminate unnecessary sources of uncertainty (which are not essential for loss assessment accuracy).

The general idea was to view an earthquake as an event that causes the distribution of various damage states to buildings across a territory. Regardless of the seismic nature of the cause, the hypothesis is that an ML technique could learn such a distribution from a small observed data set. That way, most of the uncertainties could be eliminated from the process. Since it will always be challenging to obtain large input data sets, the idea was to make an informative selection that enables even a small data set to teach an ML algorithm.

The first task was to prove that ML techniques can be used to predict damage states. The second was to prove that eliminating earthquake data from the input set does not affect accuracy. The third was to devise a method for creating a representative sample. The final task was to explore the relation between the size of the representative sample and corresponding accuracy and find the optimum tradeoff.

The case study relates to the M5.4 2010 Kraljevo earthquake. The earthquake caused two fatalities and just over one hundred medically treated injuries, but almost 6,000 structures sustained damage, a quarter of which were unsafe to occupy after the earthquake. Due to minor losses elsewhere, "loss assessment" in this research refers to predicting damage states and repair costs of residential buildings in the earthquake-affected area. Only the main findings are presented in this paper, while the complete research is in [9].

3.1 M5.4 2010 Kraljevo earthquake data modeling

The final dataset contained 1979 buildings located in three representative districts of Kraljevo, of which 652 were damaged. It took more than a year to establish the data set, working with different institutions, which illustrates the difficulty to obtain large data sets.

The buildings were classified into six building types representing the building stock in Serbia and the broader region of the West Balkans. These building types, regarding typical architecture layouts, structural systems and elements, are:

- BT1: Traditional, stone foundation, wooden superstructure buildings (constructed until the 1950s)
- BT2: Masonry structures with the old brick format (constructed until 1933)
- BT3: Masonry structures with the new brick format (constructed after 1933)
- BT4: Masonry structures with horizontally reinforced concrete ring beams (constructed 1963-1975)
- BT5: Masonry structures with horizontally and vertically reinforced concrete ring beams (constructed 1975-1990)
- BT6: Masonry structures with horizontally and vertically reinforced concrete ring beams (constructed after 1990)

The *Building* group of input attributes consists of a building type, year of construction, number of floors and the footprint area. The *Location* group of input attributes contains numeric GIS x and y coordinates of the building and the local soil type, a discrete attribute.

The damage state classification was: DS0 - no damage, DS1 - slight damage, DS2 - moderate damage, DS3 - heavy damage, and DS4 – collapse, consistent with the EMS-98 [10] damage classification with the provision of combining the EMS-98 very heavy and destruction damage state into DS.

Due to the sparsity of local measurements, the spatial distribution of the earthquake ground motion intensities at the locations of the buildings in the database was described by the earthquake magnitude and epicentre location and modelled using the Akkar-Bommer ground motion prediction equation (GMPE) suitable for seismically active regions in Europe [11], [12]. The *Earthquake* group of input attributes contains all elastic acceleration response spectrum values in the [0, 4 sec] interval (including the Peak Ground Acceleration), and the distance from the building to the epicentre.

For loss quantification construction management experts, familiar with the current market conditions, created repair cost matrices based on direct loss estimates per unit area. The matrices contain repair cost (replacement cost for collapse buildings) for each BT-DS combination.

3.2 Can ML be used for earthquake damage prediction?

The Random Forest algorithm was chosen to demonstrate that an ML model can learn the unknown mapping between the representation of a building and the observed damage states. Each building is represented as a vector containing all input attribute groups: *Earthquake* (distance to the epicenter, spectral acceleration values series), *Location* (x geo-coordinate, y

geo-coordinate, soil type), and *Building* (BT, construction year, footprint area, number of floors). 10-fold cross-validation was applied to the entire dataset (1979 buildings). The performance of the classification model was evaluated after accumulating the hits and misses from each test fold in a single confusion matrix. In this experiment, the model is validated on two different levels: by testing the accuracy of predicted damage states and by testing the accuracy of the predicted repair costs (using the expert-defined cost repair matrix, fully explained in [9]).

Results in Table 1 confirm the high accuracy (85%) of the Random Forest classification model. Precision and recall values for DS0 suggest that the model could learn the concept of the prevailing undamaged buildings. Damage states DS1 and DS3 were recognized moderately well. The small number of buildings in some damage states in the training set (127 in DS2, 64 in DS4) directly influenced the classification performance. An increased number of entities in problematic damage classes would likely improve the classification performance. This experiment shows that an ML algorithm can map damage states to buildings types, but it needs a larger data set or a data set with a better BT-DS distribution to be used directly.

Predicted / Actual	DS0	DS1	DS2	DS3	DS4	Total number of buildings:	**Recall**
DS0	1314	11	0	1	1	1327	0.99
DS1	27	234	38	18	11	328	0.71
DS2	6	64	37	13	7	127	0.29
DS3	7	31	17	69	9	133	0.52
DS4	3	18	14	10	19	64	0.30
total:	1357	358	106	111	47	**Accuracy** 1673/1979=0.85	
Precision	0.97	0.65	0.35	0.62	0.40		

Table 1. Confusion matrix with accuracy, precision and recall measures

3.2 Do we even need earthquake data to predict earthquake damage?

This intriguing hypothesis is very important for the operational versatility of the new approach and would open up various possibilities if proved to be true. Choosing the right combination of features in a building representation has two goals: to discover the minimum representation that returns acceptable prediction accuracy while using building features that can realistically be obtained. Accordingly, an experiment was conducted to assess the prediction accuracy with different combinations of building features.

In the previous experiment it was shown that an ML model can be used to predict the damage state of buildings in an area struck by an earthquake. Aiming to build an earthquake loss assessment framework usable in regions where seismic instrumentation is sparse or non-existent, it is important to show that an ML model can learn the spatial distribution of earthquake intensities at building locations without any information about the earthquake or the local soil conditions, using only the geo-coordinates and the structural characteristics of the buildings coupled with the observed damage states. An experiment was conducted to investigate this hypothesis. Four-building representations were used, namely: 1) *Earthquake* + *Location* + *Building* (all input attributes); 2) *Location* + *Building*; 3) *Earthquake* + *Building*; and 4) *XY* + *Building* (geo coordinates + descriptive structural characteristics of a building). Four different damage prediction models were trained using the Random Forest ML algorithm.

The hypothesis testing protocol comprised performing 10-fold cross-validation repeated 100 times. In each repetition, the building dataset is divided into 10 parts differently. The cross-validation results were averaged over all repetitions to perform a paired t-test [13] between the model, which used the Earthquake + Location + Building representation, and the models with other, smaller, representations. The objective was to find the representation that enables the best damage predictions (the t-test significance level was set to 0.05). The results

shown in Table 2 suggest that ML models based on all four representations performed similarly (the paired t-tests indicated that none of the building representations performed statistically better than others). However, representation *Location + Building* and *XY + Building* performed slightly better in terms of Cohen's Kappa statistics [14]. Kappa is often used as a measure of agreement between the classifier decisions and the actual class labels. It is considered a better indicator than accuracy since it takes into account the possibility of agreement occurring by chance. Models with Kappa greater than 0.7 are considered to indicate substantial agreement [15].

	Earthquake + Location + Building	*Earthquake + Building*	*Location + Building*	*XY + Building*
Accuracy (%)	85.4	83.6	85.4	85.4
Kappa	0.69	0.68	0.71	0.71

Table 2. Damage classification performance measures for different building representations.

The results shown in Table 2 indicate that a minimal building representation, consisting of the building geo-coordinates (x and y), its structural characteristics (building type, number of floors, construction year, and footprint area) and its damage state, provides a sufficiently good earthquake damage prediction for buildings in the affected area. It may be surprising to find that the information about the earthquake ground motion (location of the epicenter, magnitude of the earthquake, and the local soil type) can be omitted, and that resulting building damage prediction accuracy can still be equally good (or even slightly better!). An explanation is that the observed building damage indirectly contains the information about the local earthquake ground motion intensity. In fact, each building serves as a rudimentary seismometer. Furthermore, each building type captures a particular aspect of seismic behaviour that occurs during the earthquake, while the number of floors and the building footprint area capture the dynamics of the building. Moreover, viewed together, building damage and location information enables detecting the dependences between building types and damage states in local neighbourhoods. Based on these findings, the proposed rapid earthquake loss assessment (RELA) framework is built using the smallest building representation.

3.3 The proposed RELA Framework

Clearly, from the findings in Chapter 2, a new approach to rapid earthquake loss assessment is needed. Traditional systems assume making vulnerability relations before an earthquake which inevitably includes too many assumptions which compromise the accuracy. The new idea is to explore the possibility to form the BT-DS relations after the earthquake but quickly enough to still be considered rapid. The idea is to observe damage states on a small sample and use ML algorithms to predict damage states on the rest of the portfolio.

Two methods exist for gathering damage data after the earthquake: remote sensing and on-the-ground surveys. Remote sensing uses different methods for damage observation such as aerial or satellite images [16-20], or more recently, synthetic aperture radar data [21]. Nowadays, ML methods combined with different image processing techniques gain in popularity in remote damage assessment, both at the city block and the individual building levels [22-25]. However, in most of these studies, buildings were classified as damaged or intact, which is not detailed enough for accurate loss assessment. Findings in [17] state that "there is a general tendency in remote sensing to underestimate damage" and that "there is a need to carry out more detailed 'ground-truthing' exercises to establish how typical the degree of error found was", support on-the-ground post-earthquake damage inspection. Since on-the-ground surveys would be easier to implement and must be conducted at least once during recovery, the authors chose to explore such an approach.

As with traditional systems, the proposed RELA framework also consists of two phases:

Pre-earthquake phase - preparing the system before an earthquake. Form a database of buildings with minimum representation. Form a sample of buildings that represents the

portfolio well – a representative set. Train a group of assessors to detect the damage states of buildings. Experts prepare and regularly update a cost matrix with repair costs for each BT-DS combination.

Co-earthquake phase - activating the system immediately after an earthquake. In the co-earthquake phase, trained assessors visit the representative set, detect the damage state and upload the results. The ML Random Forest algorithm immediately predicts damage states for the rest of the portfolio. Each day, the assessors continue to visit more buildings. As the observed number of buildings increases, the prediction accuracy for the remaining part of the portfolio increases as well. After a couple of cycles, the overall prediction process is accurate enough. Since observing the representative set takes a couple of days, the prediction process can be categorized as rapid.

The RELA framework for ML-based loss assessment is shown in Fig. 2.

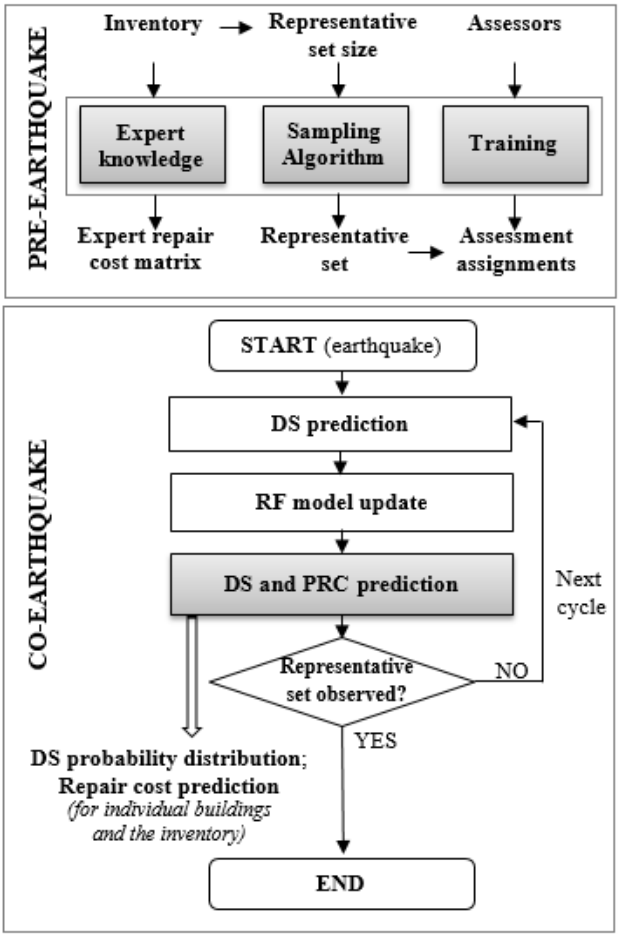

Fig. 2. Proposed Rapid Earthquake Loss Assessment (RELA) framework

3.4 Representative sampling and the optimum size of the representative set

In the previous chapter, the RELA framework was presented in general terms. Here, the main aspects, which make RELA fully operational, are explained in more detail. Specifically:

1. How to perform representative sampling to determine a representative set?
2. What is the minimum size of the representative set enough for acceptable accuracy?

A sampling algorithm used in the pre-earthquake phase to create a (relatively small) representative set of buildings from the building portfolio is illustrated in Figure 3. For the sampling procedure, <BT, number of floors> combination was chosen, sufficient to capture seismic behaviour. The portfolio is represented as a set S of n buildings. The sampling

algorithm should capture the variations in both the spatial distribution and the characteristics of the buildings in S.

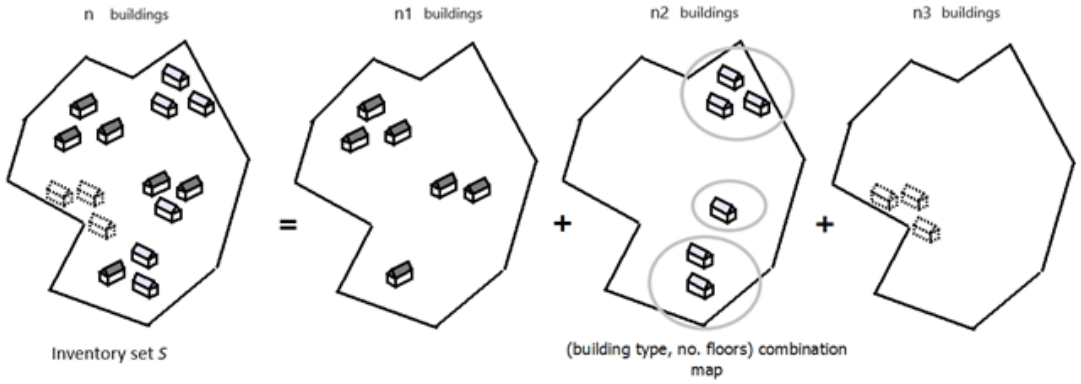

Fig. 3. Representative sampling algorithm divides the buildings in the portfolio into separate sets representing each <BT, number of floors> combination. Circles represent discovered spatial clusters.

The algorithm divides S into subsets containing only buildings from the specific <BT, number of floors> combination. These subsets contain buildings whose seismic behavior is expected to be similar. If m<<n, buildings are selected to represent the portfolio, and the algorithm chooses a proportional number of buildings from each subset. In the example shown in Figure 3, $\frac{n1}{n}m$, $\frac{n2}{n}m$, and $\frac{n3}{n}m$ buildings are selected from the corresponding subsets (n=n1+n2+n3).

Buildings from each subset are selected using the K-means clustering algorithm [26], which finds building clusters according to their x and y geo-coordinates. The clustering algorithm starts by randomly choosing k buildings (i.e., $k = \frac{n1}{n}m$ for the first subset) and assumes they are the centroids of the initial spatial clusters. A building is associated with a cluster if its distances to centroids of other clusters are greater than the distance to the centroid of the currently associated cluster. After assigning all buildings to the clusters, the algorithm re-computes the centroids and repeats the same procedure a predefined number of times, or until the updated centroids do not move from iteration to iteration. Finally, each <BT, number of floors> subset is represented with one building per each discovered spatial cluster. The proposed sampling method selects the building which is closest to the centroid of the cluster it belongs to.

Representative set size with the corresponding accuracy is a critical framework feature because it addresses the problem of small samples. The goal is to determine the minimum set size which returns satisfactory prediction accuracy. If the set is not small, it takes too much time for assessors to observe damage states, and the whole system is not rapid.

Representative set size is tested in two experiments. Representative sets, consisting of 5%, 10%, 15%, and 20% of the case study building portfolio, were created using the K-means algorithm described above. Representative sampling was performed without the knowledge of the building damage states, simulating the RELA framework pre-earthquake phase. After a representative set was created, the damage states were assigned to the buildings in it using the 2010 Kraljevo earthquake damage data, simulating the damage inspection effort of the local assessors. Kraljevo operates a Civil Protection organization numbering 57 Civil Protection Deputies, distributed in all 16 districts, which were trained and are responsible for certain duties in an event of an earthquake. These deputes are assumed to be the local assessor network in this case study. Assuming the speed of assessment of 30-40 buildings per assessor per day, nearly 2,000 buildings could be assessed in one inspection cycle (day).

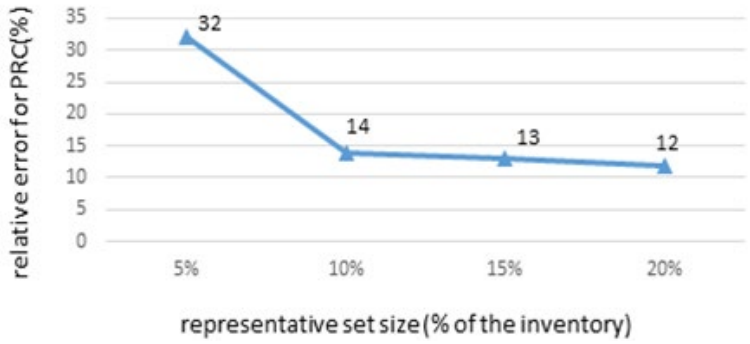

Fig. 4. Representative sampling algorithm divides the buildings in the portfolio into separate sets

In a real-life application of the proposed RELA framework, only one representative set will be selected among many possible samples that could emerge from the proposed K-means sampling procedure. What is the chance that the selected representative set will produce a Random Forest model that estimates the damage, and thus the Predicted Repair Cost (PRC), poorly? To answer this question, the probability that a Random Forest damage prediction model built on a representative set containing n% of the portfolio predicts the PRC with a relative error less than 10%, 20% and 30%, was estimated by counting the models that achieved the desired performance. This data is shown in Figure 4. Evidently, increasing the representative set size increases the probability that the proposed framework generates good damage prediction models. In this case study, a 10% representative set size delivered PRCs whose relative error was less or equal 30% with a probability of 0.85. A relative PRC error of less than 20% was observed in 7 out of 10 representative sets. In the whole city of Kraljevo, such representative set would comprise about 4,000 buildings. Given a network of 50 to 60 trained local damage assessors, the on-the-ground representative set damage assessment task could be completed in two days, allowing for one update cycle (Figure 2). This example shows that the proposed RELA framework enables rapid and fairly accurate earthquake loss assessment.

The data in Figure 5 also shows that predicting repair costs with a relative error smaller than 10% is difficult. This is because the costs of the repairs of individual buildings are quite random, characterized by wide repair cost intervals surrounding the mean values used in the expert-derived repair cost matrix (Table 3). In Kraljevo, such repair cost uncertainties were large enough to render increasing the representative set size (and prolonging the damage assessment period) ineffective in terms of improving the accuracy of PRC below the 20% relative error. This shows that the accuracy of the direct losses assessed using the RELA framework can be estimated, and that it needs to be reported to the decision makers together with the obtained loss assessments.

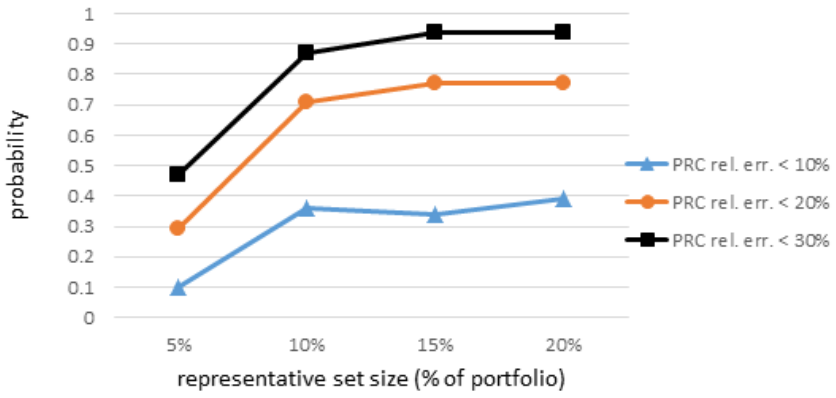

Figure 5. The probability that a model built on the representative set containing 5%, 10%, 15%, and 20% of the building portfolio produces the relative PRC error smaller than 10%, 20%, and 30%.

The case study shows how to solve the problems of traditional damage prediction by combining different kinds of expert knowledge with ML: earthquake engineers for building representation, project management experts for RELA framework and IT experts for representative sampling and testing the viability of expert's solutions. The proposed RELA framework can be considered a breakthrough in rapid loss assessment systems since it is more accurate than traditional approaches and far more implementable since it needs no infrastructure other than a group of trained assessors.

3. Conclusions

There are obstacles preventing the broader use of AI in construction, such as obtaining big data sets to employ AI techniques and adhering to analytical methods in fields more suitable for AI solutions. This case study provides ways to move forward.

The case study refers to rapid earthquake loss assessment systems. The problem with traditional systems is their low accuracy, making them unreliable and unusable in the recovery process, which is the purpose of loss assessment systems. Low accuracy is caused by too much uncertainty in analytical and insufficient data sets to create vulnerability curves in empirical methods. The contribution of this research is designing a new kind of rapid earthquake loss assessment system using multidisciplinary expert knowledge and AI methods. The problem of small data sets was solved using the procedure of representative sampling, which makes a small sample informative and sufficient to use. The low accuracy of analytical methods is caused by assuming *theoretical* vulnerability relations *before* an earthquake. The new approach uses trained assessors to perform on-the-ground observation of *actual* damage on the representative sample *after* an earthquake. AI methods are then used to predict damage to the remaining building portfolio, which is more accurate and still rapid enough. Another contribution is using a building representation without earthquake data which eliminates the need for analytical methods, shake maps and robust ground motion sensor networks, making the proposed framework applicable in any region.

The RELA framework is a result of multidisciplinary expert knowledge. A combination of representative sampling, lean building representation, machine learning for rapid damage classification based on on-the-ground inspection, and a repair cost matrix defined and updated by local experts is unique compared to existing rapid earthquake loss assessment systems. Furthermore, using buildings as damage sensors opens up the possibility for implementation in disaster scenarios other than earthquakes, using the same representative sample.

References

[1] Wald, D., Worden, B., Quitoriano, V., Pankow, K., 2005. ShakeMap manual: technical manual, user's guide, and software guide, Reston: USGS.

[2] Gehl, P., Douglas, J., D'Ayala, D., 2017. Inferring Earthquake Ground-Motion Fields with Bayesian Networks. Bulletin of the Seismological Society of America, 107(6), p. 2792–2808.

[3] Maio, R., Tsionis, G., 2015. Seismic fragility curves for the European building stock: Review and evaluation of analytical fragility curves, s.l.: JRC Technical Report EUR 27635 EN

[4] Erdik, M., Sesetyan, K., Demircioglu, M.B., Hancılar, U., Zulfikar, C., 2011. Rapid Earthquake Loss Assessment After Damaging Earthquakes. Soil Dynamics and Earthquake Engineering, Volume 31, p. 247–266

[5] Eleftheriadou, A., Karabinis, A.I., 2008. Damage probability matrices derived from earthquake statistical data. Beijing, 14th World Conference on Earthquake Engineering

[6] Calvi, G., Pinho, R., Magenes, G., Bommer, J., Restrepo-Vélez, L., Crowley, H., 2006. Development of seismic vulnerability assessment methodologies over the past 30 years. Journal of Earthquake Technology, Volume 43, pp. 75-104

[7] FEMA, 2020. HAZUS Earthquake Model Technical Manual, Washington, D.C.: Federal Emergency Management Agency
[8] Guerin-Marthe, S., Gehl, P., Negulescu, C., Auclair, S., Fayjaloun, R., 2021. Rapid earthquake response: The state-of-the art and recommendations witha focus on European systems. International Journal of Disaster Risk Reduction, Volume 52
[9] Stojadinovic, Z., Kovacevic, M., Marinkovic, D., Stojadinovic, B. 2022. „Rapid earthquake loss assessment based on machine learning and representative sampling“, Earthquake Spectra, 38, 1, 152-177
[10] Grunthal, G., 1998. European Macroseismic Scale, Luxembourg: Chaiers du Centre Européen de Géodynamique et de Séismologie, vol. 15
[11] Akkar, S., Bommer, J.J., 2010. Empirical Equations for the Prediction of PGA, PGV, and Spectral Accelerations in Europe, the Mediterranean Region, and the Middle East. Seismological Research Letters, 81(2), p. 195–206
[12] Akkar, S., Sandıkkaya, M.A., Bommer, J.J., 2014. Empirical ground-motion models for point- and extended-source crustal earthquake scenarios in Europe and the Middle East. Bulletin of Earthquake Engineering, 12(1), p. 359–387
[13] Montgomery, D., Runger, C., 2014. Applied Statistics and Probability for Engineers. 6th ed. s.l.:Wiley
[14] Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), p. 37–46
[15] Landis, J., Koch, G., 1977. The measurement of observer agreement for categorical data. Biometrics, 33(1), pp. 159-174
[16] Vu, T., Ban, Y., 2010. Context-based mapping of damaged buildings from high-resolution optical satellite images. International Journal of Remote Sensing, Volume 31, p. 3411–3425
[17] Booth, E., Saito, K., Spence, R., Madabhushi, G., Eguchi, R.T., 2011. Validating assessments of seismic damage made from remote sensing. Earthquake Spectra, Volume 27, p. 157–S177
[18] Chen, Z., Hutchinson, T., 2011. Structural damage detection using bi-temporal optical satellite images. International Journal of Remote Sensing, Volume 32, p. 4973–4997
[19] Klonus, S., Tomowski, D., Ehlers, M., Reinartz, P., Michel, U., 2012. Combined Edge Segment Texture Analysis for the Detection of Damaged Buildings in Crisis Areas.. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Volume 5, p. 1118–1128
[20] Tian, T., Nielsen, A., Reinartz, P., 2015. Building Damage Assessment after the Earthquake in Haiti using two Post-Event Satellite Stereo imagery and DSMs. International Journal of Image and Data Fusion, Volume 6, p. 155–169
[21] Plank, S., 2014. Rapid Damage Assessment by Means of Multi-Temporal SAR - A Comprehensive Review and Outlook to Sentinel-1. Remote Sensing, Volume 6, p. 4870–4906
[22] Li, P., Xu, H., Guo, J., 2010. Urban building damage detection from very high resolution imagery using OCSVM and spatial features. International Journal of Remote Sensing, Volume 31, p. 3393–3409
[23] Yu, H., Cheng, G., Ge, X., 2010. Earthquake-collapsed building extraction from LiDAR and aerophotograph based on OBIA. Hangzhou, 2nd International Conference on Information Science and Engineering, p. 2034–2037
[24] Cooner, A., Shao, Y., Campbell, J., 2016. Detection of Urban Damage Using Remote Sensing and Machine Learning Algorithms: Revisiting the 2010 Haiti Earthquake. Remote Sensing, Volume 8, p. 868
[25] Duarte, D., Nex, F., Kerle, N., Vosselman, G., 2018. Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks. Remote Sensing, Volume 10, p. 1636

[26] MacQueen, J. B., 1967. Some Methods for classification and Analysis of Multivariate Observations. Berkeley, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability