16th World Conference on Earthquake Engineering, 16WCEE 2017
Santiago Chile, January 9th to 13th 2017
Paper N° 4987
Registration Code: S-P1463213697

DATA-DRIVEN HOUSING DAMAGE AND REPAIR COST PREDICTION FRAMEWORK BASED ON THE 2010 KRALJEVO EARTHQUAKE DATA

Z. Stojadinovic[(1)], M. Kovacevic[(2)], D. Marinkovic[(3)], B. Stojadinovic[(4)]

[(1)] *Associate Professor, Faculty of Civil Engineering, University of Belgrade, joka@grf.rs*
[(2)] *Associate Professor, Faculty of Civil Engineering, University of Belgrade, milos@grf.rs*
[(3)] *Assistant Professor, Faculty of Civil Engineering, University of Belgrade, dejan@grf.rs*
[(4)] *Professor, Institute of Structural Engineering, Swiss Federal Institute of Technology (ETH) Zurich, stojadinovic@ibk.baug.ethz.ch*

Abstract

This paper presents an earthquake damage and repair cost prediction framework for individual residential buildings and portfolios of residential buildings in a municipal area in a region where the seismological networks are sparse and the structural engineering data on the existing residential building stock is poor.

The proposed data driven framework is based on the damage and reconstruction data from an actual earthquake, in this case, the M5.4 November 3, 2010 Kraljevo, Serbia, earthquake. It belongs to a more general class of hybrid building portfolio vulnerability models. The earthquake in the model is defined by its magnitude and epicenter location. The geographical distribution of the intensity of the earthquake at the location of the buildings is modeled using the 2013 Akkar-Sandikkaya-Bommer ground motion prediction model suitable for seismically active crustal regions in Europe, with the peak ground acceleration as the intensity measure. The data on the soil type distribution was collected form the municipality building department sources. The residential building stock was classified into six types by identifying typical architecture layouts, structural systems and elements. The residential building damage was surveyed after the 2010 Kraljevo earthquake by local engineers using a locally-developed survey form. The form contained the information about the individual damage, classified into four categories ranging from slight damage to collapse, varying amount of building-specific details, and addresses from which geographic locations of the buildings were derived.

A random forest machine-learning algorithm was used to derive a predictive model for residential building portfolio seismic damage and repair cost using a portion of the 2010 Kraljevo data as the learning dataset. The model outputs both the individual building fragility and the aggregate portfolio-level vulnerability data. The calculation of the expected repair cost for each building type was done using an expert-defined matrix that specifies average repair costs for each building type and damage category. The model is verified on a separate test portion of the 2010 Kraljevo dataset, yielding a satisfactory relative error when comparing total predicted to total actual repair costs.

The model is limited to regions with similar seismicity and similar building stock. However, there many regions in the Balkans that fit this constraint. The proposed framework is, however, more general. It can be applied to other regions with different seismicity and building stock using the data from a recent earthquake as its learning input dataset and an expert-defined repair cost matrix for analyzing the repair cost scenarios.

Keywords: earthquake, damage state, repair cost, machine learning

1. Introduction

Earthquake damage prediction has been in the focus of research for many years. The main problems for researchers on the structural engineering side, and the main reasons that the research is still ongoing, are: the lack of sufficiently good field data, and the diversity of the building stock. Different approaches to evaluate the earthquake-induced damage are analytical, empirical and hybrid [1]. There are other classifications (direct/indirect/conventional/hybrid, first/second/third level) but the logic is similar.

These approaches have their advantages and disadvantages. Analytical methods are capacity spectrum based, collapse mechanism based, or fully displacement based. The strength of these methods is the provable and quantifiable accuracy of damage prediction for a given, well-defined, type of building. The weak side is that real building stock is diverse and cannot be easily classified and separately investigated. As-built buildings often differ from the design documents, and their capacity diminishes due to aging or poor maintenance, making it more difficult to use the analytical methods. More important, it is hard to verify and validate the analytical models against real earthquake damage data. Thus, it is difficult to apply the analytical approach to numerous and diverse building stock at the municipal or regional level. Empirical approaches are based on classifying buildings into classes depending on materials, construction methods, structural elements and other factors influencing their seismic behavior. The evaluation of damage state probabilities for each building class is based on observed damage after previous earthquakes, and the outcomes are presented in terms of damage probability matrices or in terms of continuous fragility curves (conditional probability of exceeding a damage state, given the ground motion intensity) fitted to the data [1]. The weak side of this approach is the subjectivity in classifying buildings and in assigning building damage states, done typically during quick post-earthquake surveys, as well as accurate estimation of the local ground motion intensity. Furthermore, there is a very small number of examples where earthquake damage and repair cost data has been collected from a number of buildings large enough to allow the development of reliable building vulnerability statistics.

The proposed earthquake damage and repair cost prediction framework belongs to a more general class of hybrid building portfolio vulnerability models. The framework is intended for individual residential buildings and portfolios of residential buildings in a municipal area in a region where the seismological networks are sparse and the structural engineering data on the existing residential building stock is poor. The hybrid approach is implemented using the building damage data collected from the building damage surveys made after the M5.4 2010 Kraljevo, Serbia combined with estimates of ground motion intensity, classification of the building stock, and estimates of repair cost obtained from the scientific literature or by consulting the experts. A machine learning technique is used to process the available Kraljevo earthquake data in order to build a damage prediction model using ground motion intensity at building location and descriptive building attributes as its inputs. The model outputs a damage state probability distribution for each building and, by utilizing expert-defined cost matrix, allows for repair cost estimates specific to the residential housing inventory found in the city of Kraljevo. The framework is verified using standard machine learning procedures and validated using fragility curves. While the trained model is specific to the particular building stock, the proposed data driven approach makes it possible to investigate what-if scenarios involving different earthquake intensities, different building stock composition, and different construction cost environments.

2. Background

The proposed framework utilizes three types of inputs, the ground motion intensity estimates, the building typology, and the estimates of repair cost, obtained from scientific literature or by consulting the experts.

The seismic hazard environment of Serbia has recently been characterized within the EU SHARE project (http://www.share-eu.org) and has been studied by reinsurance companies [2]. The seismotectonic characteristics of the M5.4 2010 Kraljevo earthquake have also been investigated [3]. Based on these findings, a GMPE developed by Akkar, Sandikkaya and Bommer [4] using the pan-European ground motion databases and intended for crustal earthquake scenarios in Europe and Middle East was selected for this study to provide the

ground motion intensity estimates at individual building locations. The peak ground acceleration (PGA) intensity measure was used in this study.

A review of the building typology used in by the Global Earthquake Model, the GEM Building Taxonomy v2.0 [5], based on the NERA European Building Classification [6] derived from the analyses of the building stock in typical European cities has been conducted first. Using the data from 2010 Kraljevo post-earthquake damage surveys and observations from visits to the affected region, six types of building structures were identified. They correspond roughly to GEM MUR+ADO/LWAl (adobe), GEM MUR+CL99 (brick masonry with a variety of floors), GEM RM+CL99+RCB/LWAL (reinforced masonry with RC bands), and GEM CR/LFINF (reinforced concrete infilled frame) building classes, but an exact match could not be made. Given that the six custom classes identified in Kraljevo are representative of the building stock in Serbia and in the broader region of West Balkans, they were adopted in this study. Similarly, a four-state damage classification developed for the post-earthquake survey was derived from the post-earthquake survey forms used after the 2010 event and customized to the six custom building types, and follows the general principle of dividing the damage range from slight damage to collapse into categories associated with life safety and reparability.

The estimates of repair costs for each of the six building types and each damage state identified in the post-earthquake damage surveys were done by the authors based on their knowledge of local construction practices, labor and material costs, market fluctuations, and typical repair methods. This approach is similar to the one used to estimate the repair time and cost for seismic damage of typical California overpass bridges [7].

Research on damage prediction and loss estimation of municipal-level building stock is very diverse and includes: block-by-block based damage and loss distributions in Canada [8]; investigating the capabilities and efficiency of the seismic risk and loss assessment tool in Italy [9]; probabilistic assessment of structural damage in mid-America [10]; procedure for the seismic performance assessment of low to mid-rise RC buildings in Turkey [11], discussion of methods of predicting earthquake damage to urban systems based on the earthquake damage in Japan [12]; analyzing the seismic risk of the buildings in Spain, by using a method based on the capacity spectrum [13]; using statistical data to derive damage matrices in Greece [14]; comparing two main regional damage estimation methodologies in Turkey [15]; examining the losses of building and infrastructure materials after an earthquake and tsunami in Japan [16]; proposing a system for estimating earthquake damage in the early post-disaster period in Turkey [17], and many more. A broad array of machine-learning and data classification methods have also been deployed: artificial neural networks [18]; fuzzy logic [19]; fuzzy sets [20]; expert systems [21]; and others. Machine learning has been used in the scope of probabilistic seismic hazard analysis [22] and earthquake damage classification [23] only recently. The data-driven housing damage and repair cost prediction framework presented in this paper is unique in using actual post-earthquake damage data instead of simulations to form the training and verification datasets.

3. Damage and Repair Cost Prediction Framework

The proposed damage and repair cost prediction framework is divided into the damage prediction and the repair cost prediction modules. The damage prediction module (Figure 1) utilizes four databases. The first database contains the actual earthquake magnitude and epicenter location, as well as other data needed to utilize the selected GMPEs. The second database contains the soil type distribution in the observed region, including geo-location information. The third database contains the building stock data. Standard parameters for a building (element of the building stock) are: geo-location, year of construction, gross area, the number of floors, and footprint area (FA, equal to the gross area divided by the number of floors). The building type information is added later, as described below. Finally, the fourth database contains the data on the actual damage state (DS) of buildings derived from post-earthquake damage survey forms, including the location information. The damage states are classified into four categories according to the percentage-based locally-developed damage scale used to conduct the post-damage survey. Therefore, damage state classification is, despite the best efforts of the post-earthquake surveyors, somewhat subjective.

The layer below the databases in Figure 1 indicates the role of the experts in preparing the inputs for the machine learning engine from available data. Use of expert knowledge about the regional design practices and

the outcomes of detailed post-earthquake investigations of some buildings is needed to define the building type (BT). The building type represents the knowledge about the seismic behavior of the structure. The discrepancies between the actual building type and the building type that can be established on the basis of design documents and quick post-earthquake walk-by or drive-by surveys are often quite significant. For example, construction of one or two floors on top of existing buildings was common in the Kraljevo region during the 1990's. Such additions substantially alter the seismic behavior of the building and end up govern the resulting damage state, thus effectively changing the building type. Another aspect of expert knowledge is the use of GMPEs to evaluate the ground motion intensity at the locations of the buildings to compensate for the relatively sparse seismic instrumentation in the region. In addition, the experts can identify and correct the errors made during the post-earthquake survey (e.g. corrections in building geo-location, identification of the soil type), and possibly provide some data that may be missing in the survey. Finally, an expert analysis of the damage repair cost (for each DS of each BT) is performed at this stage and added to the dataset as an expert repair cost matrix. This data is used in the model verification and validation procedure described below.

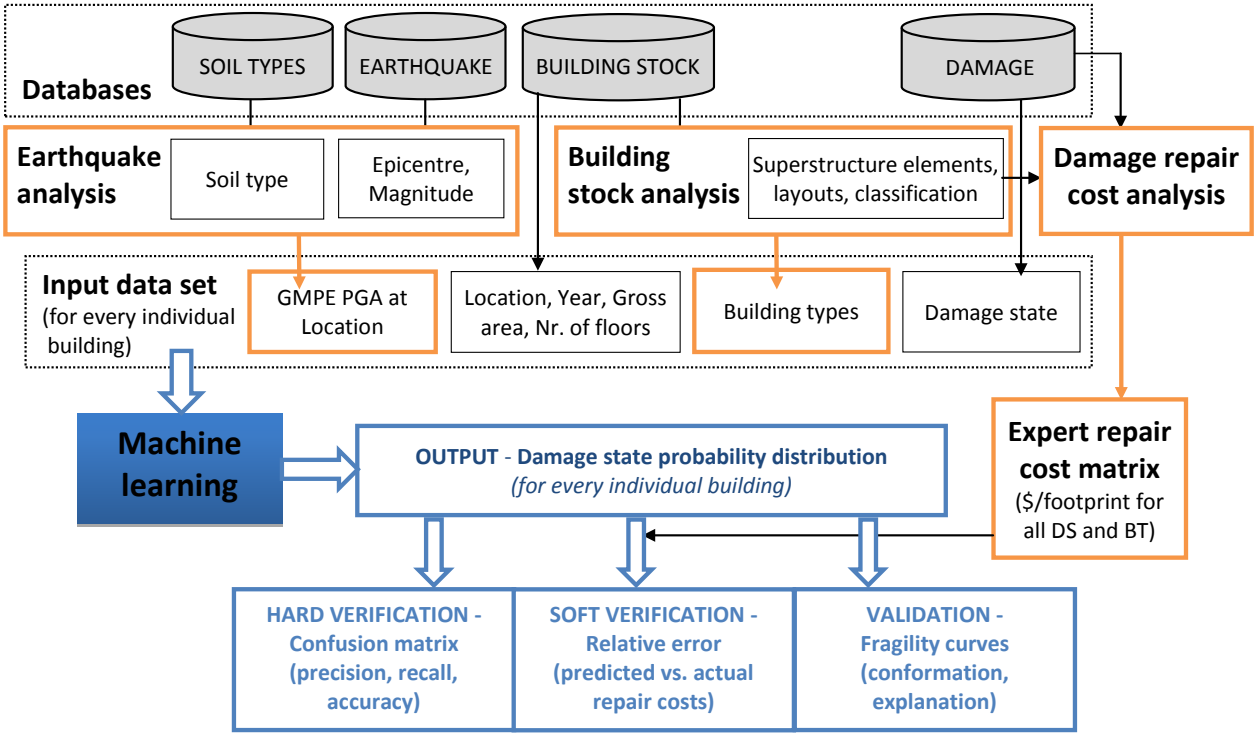

Fig. 1 – Damage prediction module of the proposed framework

The final input dataset for machine learning contains the following attributes about each building in the database: ground motion intensity (PGA in this study), building type, footprint area, building construction year, and the observed building damage state. The amount of features per building is relatively small and fairly easy to obtain from post-earthquake damage surveys. To achieve better prediction accuracy, the input dataset can be improved by including more data from GMPEs (e.g. the response spectrum information), refining the building description (e.g. predicting the fundamental vibration period of the structure to include more information about the building dynamic response), and refining the damage state classification. This effort/size/accuracy tradeoff is currently being investigated.

The central part of the damage prediction module is the machine learning algorithm and its output verification and validation procedure. The objective of the machine learning process is to establish the relation between the input building attributes and the target attribute – damage state. Thus, the obtained machine-learned relation can be used, given the input data on the ground motion intensity and building typology, to predict the earthquake damage probability distribution in the municipal building stock.

The open-source Weka tool [24] was used to perform the machine learning task. Among several possible machine learning algorithms, the Random Forest algorithm [25] was used in this study. This classification algorithm belongs to ensemble-based classifiers. The idea behind ensemble based methods is that a group of "weak learners", such as Decision Trees [26] used in a Random Forest, can be combined together to form a "strong" one. The applied combination of decision trees assumes voting between their individual decisions in order to form the output of the forest. However, each tree in the forest is trained on a different, possibly overlapping, portion of a training set. In addition, a random subset of input attributes is chosen at every node when growing each tree in the forest. Such training approach avoids overfitting the data, thus yielding more general predictive models. Apart from this generalization property, Random Forests were chosen in the framework because they can process very large input sets with many input attributes, and can handle missing values well.

To enable verification of the learning process, the input dataset is divided into the training dataset (roughly 50% of the buildings, used to "learn" the relation between the input and the output attributes) and a separate test dataset (the remaining data, used to assess the quality of the "learned" relation). This information is crucial to evaluate if the obtained model can be used to correctly predict the output for new input datasets, i.e. if the proposed framework can be used to predict the damage and repair cost of the municipal building stock in other earthquake scenarios. The sampling procedure is performed such that both the training and the test datasets credibly represent the survey earthquake outcome in terms of the distributions of the ground motion intensity, soil types, and building typology in the affected region. Verification of the prediction ability and the quality of the predictions consists of two procedures: hard and soft verification.

Hard verification: This is a test of the damage state classification accuracy for each building type. A confusion matrix $\mathbf{D}_{nxn}$ for n different damage states is formed by counting the number of correct and incorrect predictions for buildings from the test dataset. Each element d_{ik} in the confusion matrix represents the number of buildings with actual damage state i predicted as being in damage state k. The model performance is assessed using total accuracy ($\frac{\sum_{i=1}^{n} d_{ii}}{\sum_{i=1,\,j=1}^{n} d_{ij}}$), as well as damage state (DS_i) precision ($\frac{d_{ii}}{\sum_{j=1}^{n} d_{ji}}$) and recall ($\frac{d_{ii}}{\sum_{j=1}^{n} d_{ij}}$) for each damage state and building type. Since the Random Forrest algorithm outputs a damage state probability distribution for each building, hard verification assumes that the final predicted class is the most probable one.

Soft verification: This is a test of the aggregate repair cost prediction accuracy for each building type. An expert repair cost matrix is needed to compare the total actual and the total predicted repair costs for each building type, summed up over the dataset. The damage state probability distribution is expressed by probability p_i of a building being in a damage state DS_i derived from the Random Forrest algorithm output. The repair cost matrix specifies the repair cost per footprint area (€/m_2) c_{ij} for BT_j in DS_i. Let n be the number of damage states, m be the number of building types, and K(j) be the number of buildings of BT_j. Total predicted repair cost (PRC) is the sum of repair costs multiplied by respective damage state probabilities and footprint areas for all buildings:

$$PRC = \sum_{j=1}^{m} \sum_{k=1}^{K(j)} \left(FA_k \sum_{i=1}^{n} \left(p_i * c_{ij} \right) \right) \tag{1}$$

The total actual repair cost (ARC) is the sum of repair costs for the actual damage state. Equation (1) can be used to compute ARC with p_i=1 for the actual state and p_i=0 for other damage states. The final result of soft verification is the relative error between PRC and ARC defined as (PRC-ARC)/ARC.

The validation procedure consists of plotting and analyzing the seismic fragility curves for different building types, constructed from the predicted values the machine learning algorithm provides for the test

dataset. The aim is to confirm that the fragility curves have the expected shape and values for certain known building types and seismic regions and to explain the potential deviations in a scientific manner.

The repair cost prediction module of the proposed framework is shown in Figure 2. The expert repair cost matrix and the custom repair cost matrix are the principal inputs to this module. Both matrices contain repair costs per footprint area (€/m_2) c_{ij} for BT_j in DS_i. The main problem with establishing the repair costs is the qualitative definition of damage states used in the post-earthquake damage survey and the inevitable subjectivity of each surveyor. For example, the terms used in damage state definitions, such as "significant portion of the roof surface" or "extensive damage to the pillars - numerous cracks", could be interpreted in different ways in terms, yielding very different repair work estimates. Furthermore, certain repair work item quantities and durations can vary significantly from one building to another even if the buildings are of the same type, particularly for the non-structural building components. As a consequence, cost of repairing a building in a lower damage state may exceed the repair costs for building in a higher damage.

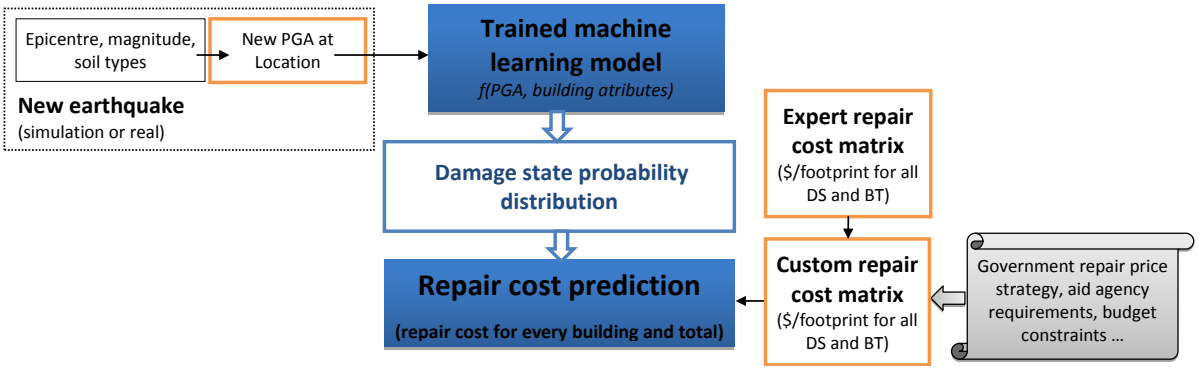

Fig. 2 – Repair cost prediction module of the proposed framework

The repair costs can be determined by statistical analysis of the past cases, or by conducting an analysis of the likely repair methods and deriving cost and time estimates using standard construction work cost data. The statistical approach is recommended for lower damage states (e.g. slight, lite) because of the mostly random nature of repair work quantities for these damage states. For higher damage states (e.g. heavy, severe) the statistical approach is applicable, but an analysis of the repair methods is recommended, since significant cost differences can occur as described above. A construction management analysis of a repair method consists of determining the work items, estimating the manpower and material quantities, and calculating the unit prices. The most challenging part is in estimating the quantities. One way to approach this problem is to develop a set of typical damage and repair scenarios for a certain BT and DS combination, evaluate the manpower and material quantities and estimate their likely variation, and associate probabilities of occurrence with each scenario. Based on this data, "mean repair quantities" can be associated with each DS for each BT. Calculating unit prices should be done using the norms (typically annual construction labor and material standards) which determine the usage of materials and labor (assuming that the work items in question do not require any special construction equipment). Defining the unit cost separately for material and labor is important: this makes the proposed framework transferable to other regions and applicable through time as the costs of material and labor change over years. In both cases, costs should be estimated using the mean repair costs and the repair cost intervals provided in the norms.

The state of the construction material and labor market changes dramatically after an earthquake: typically, the demand for materials and labor increases to satisfy the post-earthquake recovery construction needs. Thus, the costs of material and labor estimated for normal market conditions (assumed in the norms) need to be modified to match the increase in demand and the possible shortage in supply, as well as the budget constraints, municipality recovery strategies and priorities, financial and material aid dynamics, and aid agency requirements. The effect of these constraints is expressed by modifying the expert repair cost matrix (derived assuming normal market conditions) to derive a custom repair cost matrix, which is used in actual repair cost predictions. Special attention should be paid to the collapse damage state. The economic loss is 100% of the building cost but the actual compensation received by the owner(s) can vary significantly from case to case.

The repair cost prediction module can be used to explore the repair costs for different earthquake scenarios, and obtain the statistical predictions for the repair quantities and costs for individual buildings of interest or, more appropriately, for the municipality in aggregate. These scenarios use the building fragilities "learned" in the damage prediction module, and thus are representative of the building stock in the region affected by the actual earthquake used to generate the training dataset. However, different scenarios can be generated by varying the location of the epicenter and the magnitude of the earthquake for the same municipality, or by changing the building type quantities to adapt the model for a different municipality with the same building typology and a similar seismic hazard setting. Such "what-if" analyses, customized to the specific building stock typology and seismic hazard setting, can be used for future municipality-level post-earthquake recovery planning.

4. Framework implementation

The proposed framework was implemented using the data from the M5.4 2010 Kraljevo, Serbia, earthquake [27]. This earthquake resulted in only two fatalities and just over one hundred medically treated injuries, but almost 6,000 structures sustained damage, a quarter of which were found to be unsafe to occupy. The economy, education and public services of the city were affected significantly. The immediate recovery process was well organized and documented by the SEESAC [28] and local government [29]. The efforts of the local government resulted in a well-organized long-term housing reconstruction process, with the City of Kraljevo keeping track of the damage inspection reports, reconstruction funding, repair permit applications, and the resulting repair work outcomes in terms of the return of the inhabitants to their pre-disaster homesteads.

A considerable effort was needed to establish a usable database of damaged buildings, since post-earthquake survey and recovery data was not centralized, some of the information was not in electronic form, and the formats and the amount of available data varied significantly from one building to the other. This involved several site visits and repeated contacts with various local agencies. A database containing 649 damaged buildings was used in this study, while the effort to clean up the acquired data and collect more data is still ongoing.

Ground motion intensity model

The geographic distribution of the ground motion intensity for the 2010 Kraljevo earthquake was estimated using the GMPE developed by Akkar, Sandikkaya and Bommer [4]. The data on the earthquake epicenter and magnitude was obtained from [3]. The data on the soil type distribution was collected form the municipality building department sources. Using Eurocode 8 soil classification, the soil types in the Kraljevo region are mostly B and C, with some locations having soil type A and some locations near the banks of Zapadna Morava and Ibar rivers having soil type E.

Building stock model

The residential building stock was classified into six types by identifying typical architecture layouts, structural systems and elements:

BT1 – "Chatmara" buildings are the oldest residential houses in the region, with a wooden superstructure placed on stone foundation and walls made of interwoven hazel rods (chatma). They are typically single-story buildings. The roof was made of densely compacted straw. The foundations are made out of stone and clay mortar. The foundation depth depends on the soil type, and usually is around 80cm. Cellar walls are made of stone. The ground floor structure is made out of wooden beams. The space between beams is filled with mud and floor boards are put on top. The roof is typically covered with clay roof tiles.

BT2 – Unreinforced masonry structures with old brick format were built until 1933, when the new code and format for bricks was introduced. These are typically single-story buildings. The foundations, depending on the soil type, are usually 80cm deep. The exterior walls are made of brick, usually 35-40cm thick. The interior walls are usually 15cm thick. All walls are plastered. The floor construction is similar to BT1, involving wood beams. The roof is made of wood beams and lathing, covered using clay roof tiles.

BT3 – Unreinforced masonry structures with new brick format were built after 1933, when the new code and format for bricks was introduced. These kinds of houses were built until the 1960's, and may be one- or two-story tall. BT3 is similar to BT2, only the type of bricks changed and the floor heights were reduced from about 3.5m to about 2.7m. The attics were still not used as living spaces. The foundations are made of stone and are about 80cm deep, depending on the soil type. Exterior walls are 25cm or 38cm thick, depending on the number of floors. Interior walls are 12cm thick. The roofs are made using wood beams and lathing and are covered using clay tiles. There are no reinforced concrete elements in the superstructure.

BT4 – Masonry structures with horizontal reinforced concrete ring beams were built during the 1960's and 1970's. They are similar to BT3, with the main difference being the reinforced concrete horizontal ring beams placed at floor levels. Typically, these are one- and two-story structures, with the floor heights varying between 2.4m and 2.7m. The attic floor is sometimes made of concrete elements, and is typically used as living space. The foundations are sometimes made of concrete. The foundation depths are the same, so is the wall thickness, roof, etc.

BT5 - Masonry structures with horizontal and vertical reinforced concrete elements were built between 1975 and 1990. These houses were built according to approved design and have licenses. BT5 houses were built with horizontal ring beams and vertical reinforced concrete columns, but the joints typically lack the detailing to allow the reinforced concrete elements to work as a moment-resisting frame under seismic loads. These are still mostly one-story houses. Buildings of this type with more than two stories are rare. The foundations are made of reinforced concrete with the same depths as in previous types. Exterior walls made of brick are 25cm thick. Interior brick walls are 12cm thick. In this period, the use hollow bricks instead of solid bricks came into practice. All walls are plastered. The minimum floor height is 2.6m. Basement walls are made of reinforced concrete, and so is the ground floor slab. The attic floor is made mostly semi-prefabricated TM block system, a ribbed reinforced concrete ceiling made of hollow-blocks which remain built in after concreting. The roof is made of wooden beams and lathing, covered with clay roof tiles.

BT6 - Masonry structures with horizontal and vertical reinforced concrete elements built 1990 are similar to BT5 houses. The main difference is the floor construction, where the TM block system is replaced with the FERT system which uses hollow masonry elements. Exterior and interior walls are now made using hollow masonry, and the facades are plastered and more thermally insulated then before.

Damage data

The residential building damage after the 2010 Kraljevo earthquake was surveyed by local engineers using a locally-developed survey form. The form contained the information about the individual building damage, classified into six categories ranging from slight damage to collapse, denoted as 10%, 20%, 30%, 50%, 70% and 100%, varying amount of building-specific details, and addresses from which the geographic locations of the buildings were derived. The damage scale used in the 2010 Kraljevo earthquake survey resembles other damage scales that can be found in literature, but was re-classified into four damage states for this study. The first two damage states (10% and 20%) were merged into DS1 because of their similarity to reduce subjectivity. Damage state DS2 was assigned to buildings assessed to have 30% damage. The 50% and 70% damage state categories were merged into DS3 due to an insufficient number of buildings in the dataset used in this study, but will be decomposed in further work once the database population is completed. Collapsed buildings (100% damage) were classified into DS4. The distribution of damage states for the six building types is shown in Figure 3.

Repair cost matrix

The repair cost analysis was conducted separately for different damage states. For DS1 and DS2, a statistical analysis was done to determine the repair cost per footprint area for different building types and damage states. The repair cost statistics (mean and interval values) showed huge variations, which confirms the random nature of the repair costs (as well as the subjectivity of damage assessment and large variations in repair methods and extent) for low damage states. The mean damage repair costs were calculated. For DS3 an expert analysis was conducted using several damage and repair scenarios for each building type. Since different design and contractor firms were engaged in the recovery process, it was inevitable that different repair methods were proposed and used. This lead to very different repair costs per footprint area, resulting in large cost intervals for

DS3 across all building types. For the purpose of this study, average costs for the most appropriate methods were calculated. In order to establish the replacement cost for the collapsed (DS4) structures, data collected after recent flood disasters in Serbia were analyzed. In most cases, the government provided prefabricated wooden frame houses as replacement for collapsed buildings. The value of this type of building is typically 350€/m^2. The resulting repair cost matrix is presented in Table 1. This cost matrix will be used later on for machine learning accuracy verification and for repair cost predictions.

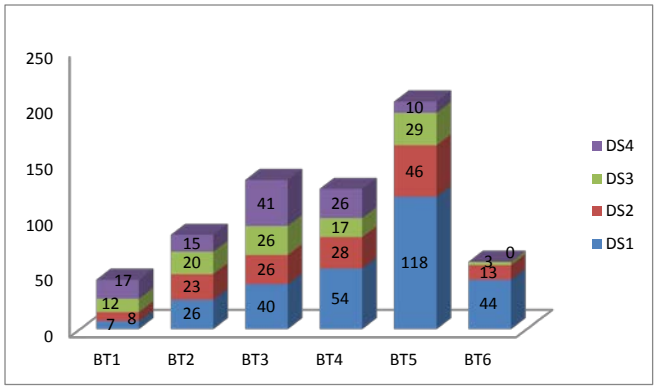

Fig. 3 – Distribution of damage states for building types

Table 1 – Repair cost matrix (€/m^2)

	BT1	BT2	BT3	BT4	BT5	BT6
DS1	3,43	12,04	8,36	10,43	9,14	9,14
DS2	13,40	16,04	15,14	18,86	17,54	18,64
DS3	55,69	46,30	44,15	38,87	29,72	32,75
DS4	350,00	350,00	350,00	350,00	350,00	350,00

Machine learning verification and validation

In order to verify the proposed model, the collected and cleaned-up 649-building database was separated into the training and the testing dataset that, respectively, contained 324 and 325 buildings uniformly distributed over the input attributes. The Random Forest machine learning method was applied to the training set. The resulting model produced damage state probability distribution for each building from the testing set.

The confusion matrix, used for hard verification, is shown in Table 2. Before commenting on these values, the confusion matrix is explained in more detail. A total of 39% of buildings were misclassified in terms of predicting the damage state: the correctly classified buildings are on the main diagonal. About half of the misclassified buildings are next to the main diagonal, particularly for DS1 and DS2. These mistakes are due, in part, to mistakes in survey field work, invisible or unidentified renovations, unrecognized pre-earthquake damage, poor quality of construction, or just building-specific conditions. For example, two almost identical houses on the same lot were in different damage states, most likely due to the proximity of a septic tank to the more damage one. The confusion matrix suggests that it was difficult for field teams to make a clear difference between DS1 and DS2 (i.e. 10%, 20% and 30% damage). Hence, the definition of lower damage states should be improved in future post-earthquake survey forms.

The most interesting result is that a significant number of buildings were severely misclassified: 40 buildings (12%) were predicted to be slightly damaged, and yet they collapsed or were heavily damaged. This outcome heavily influenced the overall prediction accuracy and the DS precisions and recalls for individual BTs. Since this was an unexpected result, these buildings were visited and investigated individually. It turned out that

many of these structures were built with serious flows, mostly missing or very poor foundations, and thus suffered serious water-induced soil-movement degradation. The buildings in question were typically in a state of semi-collapse even before the earthquake struck.

Table 2 – Confusion matrix used for hard verification

Predicted / Actual	DS1	DS2	DS3	DS4	total:	Recall
DS1	119	9	9	6	143	0.83
DS2	24	33	7	8	72	0.46
DS3	20	3	22	7	52	0.42
DS4	20	4	7	23	54	0.43
total:	183	49	45	44	Accuracy 197/321=0.61	
Precision	0.65	0.67	0.49	0.52		

The hard verification process suggested that the Random Forest model has satisfactory accuracy and reflects the reality reasonably well. The accuracy would be better if the information about the actual pre-earthquake condition of the buildings was available as an input attribute to the model. Soft verification was carried out according to equation (1) using the repair cost matrix from Table 1. The estimated repair cost was 1.964.808€ and the actual repair cost, obtained from the City of Kraljevo construction administration data for these residential buildings, was 1.991.153€. The relative error is only 1,3%, which implies that the framework could be directly used for repair cost prediction for municipalities with similar building stock and similar seismic hazard exposure. This exceptionally good result can be partially explained by analyzing the confusion matrix. It is evident that misclassifications occur on both sides of the main diagonal, which means that repair cost are almost equally overestimated and underestimated such that these errors cancel each other out to some extent.

For the purpose of validation, seismic damage fragility curves for BT3, BT4 and BT5 were computed by fitting a log-normal distribution to the data points predicted using the Random Forest model on the training dataset. Unfortunately, at this point in time the database of undamaged buildings is not complete. In order to compute the fragility data, it was assumed that the number of undamaged buildings (for each BT) was four times the number of damaged buildings in the training dataset (for each BT). The PGA values in the testing dataset vary between 0.2g and 0.4g. The fragility curves were fitted to this data, and then extrapolated and plotted in Figure 4 up to a PGA of 0.6g. The probability of being in DS1 is biggest and similar for all three building types, which could be expected for this PGA range. The biggest difference is in the probability of collapse (DS4) where it is clear that BT5 buildings, which have (almost) reinforced concrete frame masonry infill superstructures, are not likely to collapse, while unreinforced masonry BT3 and semi-reinforced masonry BT4 are more likely to collapse in this PGA range. BT3 buildings have a higher probability of being in DS2 and DS3 than other BTs, which is also logical, having in mind their superstructure and age. The selected building types represent the unreinforced and reinforced masonry structures typical for the region. These structures are comparable to the unreinforced masonry structures [30] and reinforced concrete frame structures [31] typical for Italy and Greece, but are not the same. Furthermore, the damage state classification is somewhat different. Nevertheless, the fragility data is comparable. Thus, the machine learning output can be considered as validated.

5. Conclusion

This paper presents an earthquake damage and repair cost prediction framework for individual residential buildings and portfolios of residential buildings in a municipal area in a region where the seismological networks are sparse and the structural engineering data on the existing residential building stock is poor. This framework is based on a post-earthquake survey data model coupled with the expert knowledge used pre-process the relevant information. The unknown relation between inputs (ground motion intensity and building typology and geometry) and the outputs (building damage state probability distribution) is discovered from the training data

using the Random Forest machine learning method. In order to ensure that the output is accurate and the model is usable for predictions, a hard and soft verification procedure is established, and a validation using fragility curve analysis is proposed. The framework was used to model the outcomes of the M5.4 2010 Kraljevo, Serbia earthquake. The model outputs both the individual building fragility and the aggregate portfolio-level repair cost data. The calculation of the expected repair cost for each building type was done using an expert-defined matrix that specifies average repair costs for each building type and damage category. The model was verified on a separate test portion of the 2010 Kraljevo dataset, yielding satisfactory confusion matrix parameter values and an excellent relative error when comparing total predicted to total actual repair costs.

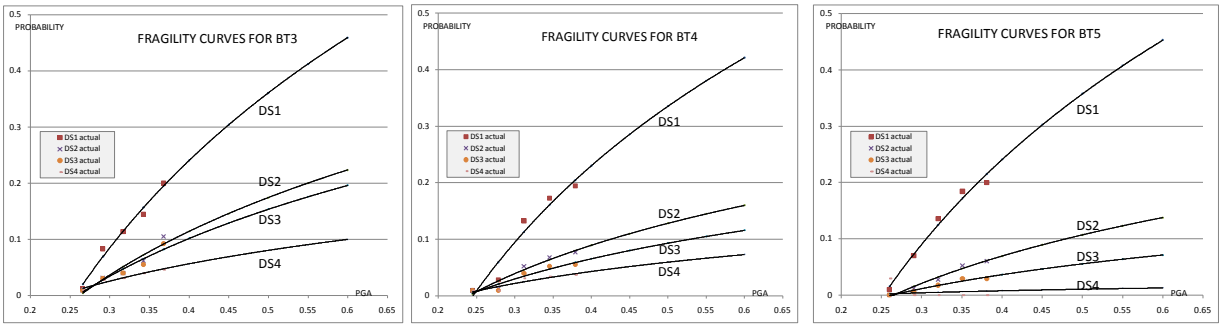

Fig. 4 – Fragility curves for BT3, BT4 and BT5 obtained for the test dataset

Acknowledgement

This work was made possible by the Swiss National Science Foundation SCOPES 2013-2016 grant number IZ73Z0-152522. The authors gratefully acknowledge this financial support, as well as the support of the Faculty of Civil Engineering of the University of Belgrade, Serbia, and the Swiss Federal Institute of Technology (ETH) Zurich, Switzerland. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the supporting institutions.

6. References

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

[2] Galasso C., Gomez I., Gupta A., Shen-Tu B. (2013): Probabilistic Seismic Hazard Assessment of Albania, Macedonia and Serbia. Proceedings of 50 SE-EEE 1963-2013 Int. Conf. on Earthquake Engineering, Skopje.

[3] Knezevic-Antonijevic S., Arroucau P., Vlahovic G. (2013): Seismotectonic Model of the Kraljevo 3 November 2010 Mw 5.4 Earthquake Sequence. *Seismological Research Letters*, vol. 84, no 4.

[4] Akkar S., Sandıkkaya M. A., Bommer J. J. (2013): Empirical ground-motion models for point- and extended-source crustal earthquake scenarios in Europe and the Middle East. *Bulletin of Earthquake Engineering*, *12*(1), 359–387.

[5] https://www.nexus.globalquakemodel.org/gem-building-taxonomy/overview (accessed 31.08.2016).

[6] Crawley H., Colombi M., Ozcebe S. (2014): European Building Classification. NERA Report D7.3.

[7] Mackie K. R., Wong J.-M., Stojadinovic B. (2011): Bridge Damage and Loss Scenarios Calibrated by Schematic Design and Cost Estimation of Repairs", *Earthquake Spectra*, Vol.27, No. 4, pp. 1127-1145.

[8] Onur T., Ventura C.E., Finn W.D.L. (2005): Regional seismic risk in British Columbia - damage and loss distribution in Victoria and Vancouver. *Can. J. Civ. Eng*, 32: 361-371

[9] Lang D. H., Molina-Palacios S., Lindholm C. D. (2008): Towards Near-Real-Time Damage Estimation Using a CSM-Based Tool for Seismic Risk Assessment. *Journal of Earthquake Engineering*, 12(S2): 199-210

[10] Bai J.W., Hueste M.B.D, Gardoni P. (2009): Probabilistic Assessment of Structural Damage due to Earthquakes for Buildings in Mid-America. *Journal of Structural Engineering*, 10: 1155-1163

[11] Yakut A., Ozcebe G., Yucemen S. (2006): Seismic vulnerability assessment using regional empirical data. *Earthquake Engineering and Structural Dynamics*, 35: 1187-1202

[12] Shibata A. (2006): Estimation of earthquake damage to urban systems. *Structural Control And Health Monitoring*, 13: 454-471

[13] Barbat A.H., Pujades L.G., Lantada N. (2006): Performance of Buildings under Earthquakes in Barcelona, Spain. *Computer-Aided Civil and Infrastructure Engineering*, 21: 573-593

[14] Eleftheriadou A.K., Karabinis A.I. (2008): Damage probability matrices derived from earthquake statistical data, *Proceedings of the 14th World Conference on Earthquake Engineering*, Beijing, China

[15] Onur T. Ventura C.E. Finn W.D.L. (2006): A comparison of two regional seismic damage estimation methodologies. *Can. J. Civ. Eng*, 33: 1401-1409

[16] Tanikawa H., Managi S., Lwin C. (2014): Estimates of Lost Material Stock of Buildings and Roads Due to the Great East Japan Earthquake and Tsunami. *Journal of Industrial Ecology*, Volume 18, Number 3: 421-431

[17] Aydin C., Tecim V. (2014): Description logic based earthquake damage estimation for disaster management. *J. of Sci. and Technology*, 15(2): 93-103

[18] Molas G.L., Yamazaki F. (1995): Neural Network for Quick Earthquake Damage Estimation, *Earthquake Engineering and Structural Dynamics*, vol 24, pp. 505-516.

[19] Sanchez-Silva M., Garcia L. (2001) Earthquake Damage Assessment Based on Fuzzy Logic and Neural Networks. *Earthquake Spectra*, vol. 17, no. 1, pp. 89-112.

[20] Juang C.H., Elton D.J. (1986): Fuzzy logic for estimation of earthquake intensity based on building damage records, *Civil Engineering Systems*, vol. 3, no. 4, pp. 187-191

[21] Carreno M.L., Cardona O.D., Barbat A.H. (2004): Expert System for Building Damage Evaluation in Case of Earthquake, *Proceedings of the 13th World Conference on Earthquake Engineering*, paper No. 3047, Vancouver, Canada.

[22] Alimoradi A., Beck, J. (2014): Machine-Learning Methods for Earthquake Ground Motion Analysis and Simulation. *ASCE J. Eng. Mech.*, vol. 141, no. 4.

[23] Ferguson M., Martin A. (2015): Earthquake-Induced Structural Damage Classificaiton Algorithm, CS 229 Machine Learning – Final Report, Dept. of Civil and Env. Engineering, Stanford University.

[24] Hall M., Eibe F., Holmes G., Pfahringer B., Reutemann P., Witten I. (2009): The WEKA Data Mining Software: An Update; *SIGKDD Explorations*. Volume 11, Issue 1.

[25] Breiman L. (2001): Random Forests. *Machine Learning*: 45 (1) 5-32.

[26] Quinlan R. (1983): Learning efficient classification procedures. *Machine Learning: an artificial intelligence approach*, Michalski, Carbonell & Mitchell (eds.), Morgan Kaufmann, 1983, p. 463-482

[27] http://en.wikipedia.org/wiki/2010_Serbia_earthquake (accessed 31.08.2016)

[28] http://www.seesac.org/project.php?l1=140&l2=151 (accessed 31.08.2016)

[29] City of Kraljevo (2010) "Action Plan for Urgent Recovery of the Consequences from the Kraljevo 2010. Earthquake" (in Serbian).

[30] Rota M., Penna A., Maganes G. (2010): A methodology for deriving analytical fragility curves for masonry buildings based on stochastic nonlinear analyses, *Engineering Structures*, vol. 32, p. 1312-1323

[31] Kappos A. (2016): An overview of the development of the hybrid method for seismic vulnerability assessment of buildings. *Structure and Infrastructure Engineering*, DOI: 10.1080/15732479.2016.1151448