Rapid earthquake damage and loss assessment is crucial both for insuring the safety of inhabitants in the immediate aftermath of an earthquake and for the recovery of the stricken communities in the long run. This paper investigates the potential of different machine learning methods for building a rapid earthquake loss assessment system intended for residential houses in a municipal area. The system is trained on a pre-earthquake selected representative set of residential houses, after observing their damage and loss states. Two representative sampling strategies and three machine learning algorithms are described and evaluated on the 2010 Kraljevo M5.4 earthquake data set. The proposed models showed satisfactory accuracy in predicting the total expected repair cost (less than 20% error with the representative sample size of 10% of the inventory). The approach is independent of geological and earthquake data and does not require local peak ground acceleration values.
Sampling and Machine Learning Methods for a Rapid Earthquake Loss Assessment System

M. Kovačević5, Z. Stojadinović6, D. Marinković7 and B. Stojadinović8

ABSTRACT

Rapid earthquake damage and loss assessment is crucial both for insuring the safety of inhabitants in the immediate aftermath of an earthquake and for the recovery of the stricken communities in the long run. This paper investigates the potential of different machine learning methods for building a rapid earthquake loss assessment system intended for residential houses in a municipal area. The system is trained on a pre-earthquake selected representative set of residential houses, after observing their damage and loss states. Two representative sampling strategies and three machine learning algorithms are described and evaluated on the 2010 Kraljevo M5.4 earthquake data set. The proposed models showed satisfactory accuracy in predicting the total expected repair cost (less than 20% error with the representative sample size of 10% of the inventory). The approach is independent of geological and earthquake data and does not require local peak ground acceleration values.

Introduction

Today, there are analytical, empirical and hybrid approaches to evaluate the earthquake-induced damage [1]. Analytical methods assume the existence of a model that relates to a well-defined type of a building. Nevertheless, 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, especially for numerous and diverse building stock at the municipal or regional level. Empirical approaches are based on classifying buildings into classes depending on the 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 the observed damage after previous earthquakes, and the outcomes are presented in the form of discrete damage probability matrices or continuous fragility curves fitted to the data [1]. The weak side of this

5 Associate professor, Faculty. of Civil Engineering, University of Belgrade, Serbia (email: milos@grf.bg.ac.rs)
6 Associate professor, Faculty. of Civil Engineering, University of Belgrade, Serbia (email: joka@grf.bg.ac.rs)
7 Assistant professor, Faculty. of Civil Engineering, University of Belgrade, Serbia (email: dejan@grf.bg.ac.rs)
8 Professor, Institute of Structural Engineering, ETH, Zurich (email: stojadinovic@ibk.baug.ethz.ch)

approach is the subjectivity in classifying buildings and in assigning building damage states, done typically during quick post-earthquake surveys, as well as the accurate estimation of the local ground motion intensity. Furthermore, there are few cases where the earthquake damage and repair cost data has been collected to include enough buildings for a reliable statistical analysis.

The research presented herein investigates the use of machine learning (ML) methods for rapid seismic loss assessment of housing inventory on a municipality level, based on the observed damage in a small representative sample of houses. A sufficient set of attributes describing residential family houses is proposed and evaluated. This attribute set includes spatial coordinates of the houses, but the assessment is independent of the peak ground acceleration (PGA) values. The proposed data driven rapid loss assessment system belongs to a more general class of hybrid building portfolio vulnerability models. It is intended for the regions where the seismological networks are sparse and the structural engineering data on the existing residential building stock is poor. The system distinguishes activities conducted in the pre-earthquake and the post-earthquake phase. The pre-earthquake phase includes identification of a small representative set of houses from the available municipality building inventory using a proposed sampling procedure – representative sampling. The procedure is designed to extract enough information from the inventory required to build a ML assessment model, independently of future earthquakes. After the earthquake, a procedure for rapid observation of the damage states in the representative set is conducted, enabling building the ML model and using it to predict damage in the whole inventory. The ML model outputs a damage state probability distribution for each house in the inventory and, by utilizing an expert-defined cost matrix, allows for repair cost and repair time estimates specific to the built inventory of the observed municipal area. In this study, repair costs are taken as a direct measure of earthquake-induced losses to the residential housing inventory of a municipality. No other direct or indirect seismic losses are considered.

The proposed ML methods and sampling procedures are compared and verified using the data collected from the M5.4 2010 Kraljevo, Serbia earthquake [2]. The verification against the real data showed that the proposed data driven rapid loss assessment system is capable of predicting the actual repair costs quite well.

Background

Research on damage prediction and loss estimation of municipal building stock is very diverse and includes: block-by-block based damage and loss distributions in Canada [3]; investigating the capabilities and efficiency of the seismic risk and loss assessment tool in Italy [4]; probabilistic assessment of structural damage in mid-America [5]; procedure for the seismic performance assessment of low to mid-rise RC buildings in Turkey [6]; discussion of methods of predicting earthquake damage to urban systems based on the earthquake damage in Japan [7]; analyzing the seismic risk of the buildings in Spain, by using a method based on the capacity spectrum [8]; using statistical data to derive damage matrices in Greece [9]; comparing two main regional damage estimation methodologies in Turkey [10]; examining the losses of building and infrastructure materials after an earthquake and tsunami in Japan [11]; proposing a system for estimating earthquake damage in the early post-disaster period in Turkey [12], and many more. A broad array of machine-learning and data classification methods have also been deployed: artificial neural networks [13]; fuzzy logic [14]; fuzzy sets [15]; expert systems [16]; and others. Machine learning has been used in the scope of probabilistic seismic hazard analysis [17] and earthquake damage
classification [18] only recently. The data-driven approach to repair cost assessment presented in this paper is unique in using the actual post-earthquake damage data, instead of simulations to form the training and verification datasets.

Systems for rapid earthquake loss assessment are formulated based on various approaches to damage prediction. Rapid loss assessment systems can be global or local depending on the size of area they cover. An excellent overview of existing systems was given by Erdik et al. [19], [20]. The ML methods presented in this paper are intended for a municipality-level housing inventory, and therefore, the rapid earthquake loss assessment system that uses the presented ML methods is local by its nature.

**Machine learning approach to earthquake loss assessment**

Machine learning (ML) is the area of computer science which studies intelligent programs that learn from the experience obtained from an environment [21]. Many ML techniques such as Decision Trees [22], Neural Networks [23] or Random Forests [24], can be used to build predictive models from the available data. In a data driven approach, a ML model is built in a training (learning) procedure in which a learning algorithm tries to find the best possible mapping between the presented inputs and the desired output. This supervised learning approach assumes the existence of a training set containing examples in which both inputs and outputs are known in advance. Depending on the nature of the output, supervised learning is used to predict continuous (regression), or discrete outputs (classification). In this research, the loss assessment is indirectly treated as a classification problem in which classes correspond to the observed damage states (DS), and suitable attributes describing houses represent model inputs. Knowing the probability distribution of damage states for a house, it is possible to calculate its expected repair cost.

**Proposed house representation**

In order to learn the DS boundaries, in the space of all possible houses, ML methods assume the existence of the convenient representation for each house in the inventory. A proposed representation maps a house into a six-dimensional space where the dimensions (attributes) are: building type, construction year, footprint area, number of floors, and x and y geographic coordinates describing the spatial location of the house. Building type (BT) identifies the architectural layout and the structural system and elements. Together with the footprint area and the number of floors, BT is a proxy for the static and dynamic behavior of the house. Construction year could be related to the condition of a house. Assuming that the classification model is trained immediately after the disaster, the geographic coordinates indirectly contain the information about the local earthquake intensity and the soil type. Furthermore, the geographic coordinates contain the unknown dependences between BT and DS in local neighborhoods.

The proposed attributes are fairly easy to obtain from local housing inventory data and local structural engineering experience, which makes it possible to apply the proposed seismic loss assessment model in various countries, independent of the existence of local seismic networks. Once trained, the ML assessment model allows skipping the earthquake intensity assessment level (e.g. identification of the epicenter and application of attenuation relations), and makes it possible to go directly to the observable damage and losses, omitting the uncertainties of estimating local PGA values and using theoretical fragility and vulnerability curves.
Applied classification and clustering methods

A comparison of three specific ML methods is performed: Decision Trees [22], Neural Networks [23], and Random Forests [24]. The methods produce classifiers that output both class labels and the predicted probability distribution of classes (damage states). In addition, to improve the proposed sampling method for choosing the representative houses (training set), a K-Means clustering method is used.

Decision Tree (DT) is a non-parametric learning method which builds a tree from the available examples. In each node of a built tree, an example is tested against the value of the attribute associated with the node. Depending on the test result, the example is forwarded down the tree until it reaches the leaf node with the appropriate class label. The selection of the attributes when growing the tree is conducted using different possible criteria, which place more informative attributes near the top of the tree, and the most informative in the root. The attribute is informative if it splits the examples entering the associated node into subsets in which class labels are more uniform (ideally, each subset would contain the examples from only one class). The procedure is recursively repeated until all nodes are inserted into the tree. The built decision tree is interpretable, since it provides rules for classification in the form of a sequence of if/then clauses (each path from root to a leaf node is a conjunction of attributes tests). This method is selected because the interpretable model could reveal the reasons for occurrences of certain damage states.

Neural network (NN) consists of a set of interconnected processing units (neurons). Each neuron calculates a weighted sum of its real-valued inputs, and then uses an activation function to form the single output value. Input lines and the connections between neurons are associated with real-valued weights. Classification networks typically contain two layers of neurons: hidden layer neurons collect the network inputs (real-valued attributes forming the learning example), and are not connected between themselves; output layer neurons take as their inputs all the outputs from the hidden layer. The number of output neurons is equal to the number of classes. In the learning phase, examples are fed into the network and the outputs are compared with the desired ones (i.e. binary output vector with zeros and a one representing the class). All weights in the network are updated during training to minimize the error function (difference between output and desired values), which forms a high-dimension surface in the space of weights. A neural network is not guaranteed to find the global minimum of the error function, but is commonly used due to its ability to learn very complex class boundaries (expressiveness). The popularity and the expressiveness of this approach are the main motives to use neural networks in this research.

Random Forest (RF) belongs to the family of ensemble-based methods. The idea behind these methods is that a group of “weak learners”, such as Decision Trees 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 overlapped portion of a training set. In addition, the 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 to more general predictive models. Apart from good generalization, Random Forest is capable to process very large input sets with many input attributes, and can handle missing values well. Therefore, it is convenient for the datasets that appear in the process of earthquake loss assessment.
K-Means clustering is an unsupervised learning method in which class labels are not known during the training phase. The aim of the learning process is to find the clusters of similar examples based on the values of their attributes. The input parameter K represents the desired number of clusters produced by the learning algorithm. The algorithm starts by randomly choosing K points representing the centroids of initial clusters. An example is associated with a cluster if its distances to centroids of other clusters are greater than the distance to the centroid of the associated cluster. After assigning all examples to the clusters, the algorithm re-computes the centroids and repeats the same procedure a predefined number of times, or until the new centroids do not move from iteration to iteration. The algorithm is convenient if the number of clusters is known in advance, which holds in the proposed sampling procedure described in the next section.

**Representative sampling**

The representative sampling is used to form a small representative set of houses from the municipality inventory before the occurrence of an earthquake. Since the representative set will be used to train the classification model immediately after the earthquake (after observing its damage states), the sampling algorithm should be able to capture all the variations in the inventory, from the point of both building characteristics and their spatial distribution. In addition, for the assessment system to be rapid, the set must be small enough to be rapidly observable by the available number of local accessors.

![Diagram](image)

**Figure 1.** Representative sampling: a municipality is divided into subsets representing each possible \{BT, number of floors\} combination. A subset contains houses with probably similar seismic behavior. Here, it holds that \(n = n_1 + n_2 + n_3\).

In the first step, the inventory is divided into separate sets of houses representing settlements in a municipality area. Applied on the municipality level, the algorithm chooses a proportional number of samples from each settlement \((m)\), to reflect the distribution of number of houses across the settlements. Each settlement is associated with the pool of local accessors living on that territory. They should be trained to recognize damage states for the building types in that settlement. Let the settlement contains \(n\) houses (Fig. 1). In the second step, it is divided into subsets containing only houses from the specific \{BT, number of floors\} combination. If one needs to select \(m\) houses from the settlement, the algorithm choses the proportional number of houses
from each subset (in Fig. 1, \( \frac{n}{a} m \), \( \frac{n}{b} m \) and \( \frac{n}{c} m \)). The houses representing the subsets could be selected randomly: thus, this version of the sampling algorithm will be denoted as a random representative sampling.

In order to achieve the proper spatial distribution of samples in a subset, the houses are clustered with respect to their coordinates by using K-Means method. Parameter K is set to the number of houses needed from the subset (i.e. for the first subset from Fig. 1, K= \( \frac{n}{a} m \)). The samples are selected by picking the nearest-to-the-cluster-center houses from each cluster. The motivation for picking only the central house in a cluster is to avoid injecting a possibly redundant information into a small-size representative set. This version of the sampling algorithm will be denoted as a clustered representative sampling.

Model evaluation

The loss assessment models developed in this research are differentiated by the type of representative sampling algorithm (random or clustered), and by the learning method applied on the representative set (DT, NN or DF). A brief description of the dataset and the evaluation metrics is given before the explanation of evaluation experiments.

Dataset

The 2010 Kraljevo (Serbia) earthquake (M5.4) could be classified as a typical recent earthquake disaster. Almost 6,000 structures sustained damage, a quarter of which were found to be unsafe to occupy. The immediate recovery process was well organized and documented by the local government. This was followed by a well-documented long-term housing reconstruction process, with the City of Kraljevo keeping track of the damage inspection reports, reconstruction funding, repair permit applications, repair costs and durations, and the repair work outcomes in terms of the return of the inhabitants to their pre-disaster homesteads.

This research started with the considerable effort to establish a database of damaged residential houses, since the data were not centralized, and some of the information was only in the paper form. Various local agencies were contacted and finally a database was formed containing 652 damaged houses in three representative settlements. For the purpose of proper damage predictions, a database of undamaged houses was obtained from the tax department, resulting in the final dataset which contained 1979 houses. The houses were classified into six building types by identifying typical architecture layouts, structural systems and elements:

- BT1: Traditional, stone foundation, wooden superstructure buildings (pre-1950's)
- BT2: Masonry structures, the old brick format (pre-1933)
- BT3: Masonry structures, the new brick format (post-1933)
- BT4: Masonry structures, horizontally reinforced concrete ring beams (1963-1975)
- BT5: Masonry str., horizontally and vertically reinforced concrete beams (1975-1990)
- BT6: Masonry str., horizontally and vertically reinforced concrete beams (1990-today)

The residential building damage was surveyed after the earthquake by local engineers using a locally-developed damage survey form. The form contained the information about the particular
building damage classified into six categories ranging from slight damage to collapse (10%, 20%, 30%, 50%, 70% and 100%), varying amount of building-specific details, and an address from which the geographic locations were obtained. However, two pairs of damage states were merged (10% with 20%, and 50% with 70%) because of their similarity and to better match the damage classifications found elsewhere. Hence, the final damage state (DS) classification used in the research is: DS0 - no damage, DS1 - slight damage, DS2 - moderate damage, DS3 - heavy damage, and DS4 - collapse. The distribution of damage states for building types in the data set is shown in Table 1.

Table 1. Distribution of damage states for different building types.

<table>
<thead>
<tr>
<th>DS</th>
<th>BT1</th>
<th>BT2</th>
<th>BT3</th>
<th>BT4</th>
<th>BT5</th>
<th>BT6</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS4</td>
<td>7</td>
<td>14</td>
<td>24</td>
<td>16</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>DS3</td>
<td>10</td>
<td>46</td>
<td>31</td>
<td>23</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>DS2</td>
<td>7</td>
<td>24</td>
<td>25</td>
<td>30</td>
<td>29</td>
<td>12</td>
</tr>
<tr>
<td>DS1</td>
<td>6</td>
<td>38</td>
<td>42</td>
<td>62</td>
<td>131</td>
<td>49</td>
</tr>
<tr>
<td>DS0</td>
<td>0</td>
<td>116</td>
<td>305</td>
<td>432</td>
<td>276</td>
<td>198</td>
</tr>
</tbody>
</table>

Evaluation metrics

Let \( h \in H \) be a house, and \( d \in D \), a damage state from the sets of all houses and damage states (other than DS0) in the data set. In order to verify the ability of the model to estimate the repair costs, an expert-defined repair cost matrix \( C=[c(d, t)] \) should be developed for the inventory under consideration (\( c(d, t) \) – repair cost per footprint area for \( DS = d \), and \( BT = t \)). The matrix \( C \), shown in Table 2, was used to compute the expected repair cost for each individual house and the total predicted repair cost (\( PRC \)) for all houses in the dataset. This value was compared to the actual repair costs (\( ARC \)) obtained from the City of Kraljevo database to evaluate the ML assessment models developed in this research project.

Table 2. Expert-defined repair cost matrix for the City of Kraljevo, year 2010, (in €/m\(^2\)).

<table>
<thead>
<tr>
<th>€/m(^2)</th>
<th>BT1</th>
<th>BT2</th>
<th>BT3</th>
<th>BT4</th>
<th>BT5</th>
<th>BT6</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>3,43</td>
<td>12,04</td>
<td>8,36</td>
<td>10,43</td>
<td>9,14</td>
<td>9,14</td>
</tr>
<tr>
<td>DS2</td>
<td>13,40</td>
<td>16,04</td>
<td>15,14</td>
<td>18,86</td>
<td>17,54</td>
<td>18,64</td>
</tr>
<tr>
<td>DS3</td>
<td>55,69</td>
<td>46,30</td>
<td>44,15</td>
<td>38,87</td>
<td>29,72</td>
<td>32,75</td>
</tr>
<tr>
<td>DS4</td>
<td>350,00</td>
<td>350,00</td>
<td>350,00</td>
<td>350,00</td>
<td>350,00</td>
<td>350,00</td>
</tr>
</tbody>
</table>

Let \( p_d(h) \) be the probability of \( h \) belonging to a damage state \( d \), computed by the ML classifier. If \( f(h) \) denotes the footprint area of \( h \), and \( t(h) \) denotes the type of \( h \), then:

\[
PRC = \sum_{h \in H} \sum_{d \in D} p_d(h)f_d(h)c(d, t(h))
\]  

There is a slight modification in Eq. 1 related to the damage state DS4, since the related repair cost from \( C \) is given in euros per gross area, instead of euros per footprint area. In this case, \( f(h) \) should be multiplied by the number of floors. The total actual repair cost is defined to be the
sum of repair costs for the actual damage states. It is calculated using Eq. (1), where \( p_d(h) \) is set to one for the actual state, otherwise zero. Finally, in the experiments described in the next section, the percent error of the total predicted cost \( \delta(\text{PRC}) \) is used:

\[
\delta(\text{PRC}) = |\text{PRC} - \text{ARC}| / \text{ARC}
\]  

(2)

**Experiments**

Both versions of representative sampling are non-deterministic (recall that K-Means clustering, applied in the clustered version, randomly chooses the starting centroids). Therefore, when testing a loss assessment model, it is necessary to generate many \{representative set, test set\} splits for each applied ML method (DT, NN and RF). We generated 100 equally sized representative sets containing \( p \)% of the whole inventory (1979 houses), and the corresponding \((100 - p)\)% sized test sets. Since the approach assumes rapid observation of damage states in a representative set, \( p \) took values from \{5, 10, 15, 20\}. The models built on representative sets were evaluated on test sets using the Eq. 2. In order to remove the effects of extreme values, the median percent error (\( \delta(\text{PRC}) \)) was calculated and shown in Table 3.

Table 3. Percent error for the total \( \text{PRC} \), with respect to the total \( \text{ARC} \). Shaded cells represent the best results concerning the (relatively) small number of houses (5% and 10% of the inventory) that should be inspected after the earthquake.

<table>
<thead>
<tr>
<th></th>
<th>Random representative set</th>
<th></th>
<th></th>
<th>Clustersed representative set</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DT</td>
<td>NN</td>
<td>RF</td>
<td>DT</td>
<td>NN</td>
<td>RF</td>
</tr>
<tr>
<td>5%</td>
<td>30</td>
<td>28</td>
<td>22</td>
<td>29</td>
<td>25</td>
<td>32</td>
</tr>
<tr>
<td>10%</td>
<td>23</td>
<td>19</td>
<td>17</td>
<td>18</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>15%</td>
<td>20</td>
<td>21</td>
<td>17</td>
<td>18</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>20%</td>
<td>16</td>
<td>18</td>
<td>15</td>
<td>17</td>
<td>16</td>
<td>12</td>
</tr>
</tbody>
</table>

All median models achieved very good performance in predicting the total repair cost (percent error in the worst case is 32%). Clustered sampling was better than random sampling for all ML methods, while RF is the best learning method. Results from Table 3 suggest that the 10% sample is the optimal choice from the time-to-observe – performance point of view for the rapid damage and loss assessment system. In addition, we believe that the 5% results would be better if our dataset was larger because the 5% would also be larger and more diverse.

What is the chance that a selected representative set produces a bad assessment model? The probabilities that the best performing RF method developed using a clustered representative set to predict the total repair cost with percent error less than 10, 20 or 30 percent were estimated (Fig. 2). The probability estimates were calculated by counting the models that achieved the desired performance out of all 100 models, for all representative set sizes on the abscissa. Curves from Fig. 2 suggest that the optimal choice would be the 15% representative set, since it produces models with percent error less than 30% in 93% of cases, and less than 20% in 76% of cases. Furthermore, models built using a 20% representative set improve negligibly. Models built using a 10% representative set have a percent error less than 20% in 71% of cases. Errors smaller than 10% require larger representative sets, necessitating longer observations times.
Conclusions

The paper presents how to build an accurate machine learning model, intended for the rapid earthquake loss assessment of residential houses on a municipal level. After observing damage states on a pre-earthquake selected representative set of houses (e.g. 10% of the total inventory), it is possible to estimate the total repair cost with less than 20% error, with a probability of 0.71 for the 2010 M5.4 Kraljevo, Serbia earthquake. The model assumes that the houses are represented with their spatial coordinates, building type, construction year, number of floors and footprint area, but it does not require any geological data, or even data about the earthquake! The geology and earthquake intensity related information is implicitly included in the geolocation, thus freeing the model from PGA evaluation. Based on a carefully selected representative set, the model learns from the actual earthquake by observing damage states, and copes better with the empirical uncertainties, such as the unexpected seismic behavior of particular houses. The ML-based rapid earthquake loss assessment model is applicable to any municipality, providing there is a housing inventory with the specified attributes, a pool of local assessors, and an expert-defined repair cost matrix.

Acknowledgments

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.

References


3. Onur T, Ventura CE, Finn WDL. Regional seismic risk in British Columbia - damage and loss distribution in


