Will I Feel It?
Using Performance Based Earthquake Engineering to predict the extent of earthquake damage in California homes*

Ahmad Wani,† Nicole Hu,‡ and Yawar Aziz§
Stanford University
(Dated: December 12, 2014)

Predicting the scale and scope of damage as quickly as possible following an earthquake is beneficial in coordinating local emergency response efforts; implementing shelter, food, and medical plans; and requesting assistance from the state and federal levels. Additionally, estimating the damage state and economic losses of individual homes is important in assessing household risk and establishing insurance rates. This project responded to both of these needs by applying machine learning to predict earthquake damage and estimate losses. This is the first time machine learning techniques have been allied with Performance Based Earthquake Engineering to predict damage. Using features known to influence how earthquakes affect structures (e.g. seismic, soil, and structural parameters), extensive data was collected from multiple sources, and substantial pre-processing techniques were implemented. Precalculated damage states from thousands of homes from many past earthquakes served as a training set and learning techniques (svm, random forest and neural networks) were used to develop a web application that can predict damage and estimate losses to single family homes in the state of California.

Keywords: svm, random forest, neural networks, earthquake, damage prediction, usgs, loss, dyfi

I. INTRODUCTION

United States Geological Survey (USGS) website has an online post-earthquake survey form called "Did You Feel It?" (DYFI) where respondents report about what they felt and saw during an earthquake. Figure 1 shows a sample of the questions asked to respondents on DYFI. A complete list of survey questions can be found on `http://earthquake.usgs.gov/earthquakes/dyfi/`. The USGS computes a CDI value for each survey response using the Dewey and Dengler (1998) procedures, aggregates the data, and ultimately reports the aggregate CDI value for each zip code. For this project, the CDI values computed for each response are considered to be ground truth for machine learning. The scope was limited to California based on time constraints and in order to address one of the most seismically active areas of the United States. The scope was also restricted to predicting damage to single family homes, as this represents the largest single group of structures by type (`www.census.gov`). The authors sought to respond to this research question from the perspective of two separate audiences. Firstly, the individual homeowner, with limited knowledge of earthquake engineering, should be able to determine the damage state across a range of seismic hazard levels as well as calculate expected losses from each hazard level and calculate expected annual loss useful for making informed decisions regarding household financial planning. Secondly, immediately following an earthquake, the disaster response center within a community should be able to estimate the extent and severity of the damage to determine whether or not homes in their community are affected, and subsequently tailor response and recovery efforts.

The performance based earthquake engineering (PBEE) methodology developed by the Pacific Earthquake Engineering Research (PEER) Center follows a logical, stepwise approach to performance assessment and subsequent damage and loss estimates of a structure due to an earthquake. The framework is rigorous, probabilistic, and requires inputs from disciplines such as seismology, structural engineering, loss modeling, and

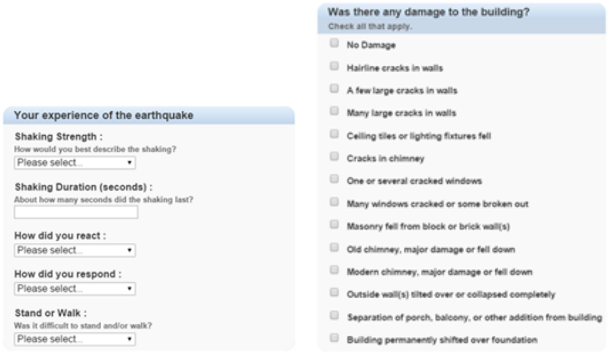

FIG. 1: Selected Did You Feel It Question Excerpts

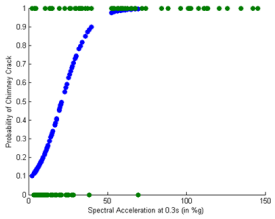

FIG. 2: Chimney Fragility Curve

risk management to ultimately inform stakeholders and decision makers of seismic consequences. More details about the PBEE framework can be found in several publications including Deierlein (2004) and Krawinkler and Miranda (2004).

II. METHODS

A. Data Collection

The project team requested DYFI data for all past California earthquakes with at least 10,000 responses from 36 seismic events with a bias towards more recent events, those centered near more populated areas, and events of larger magnitudes. The supplied data spanned from magnitude 3.4 (San Francisco Bay area, April 2011) to 7.2 (Baja, April 2010).

Many additional features were appended to each response in the DYFI dataset. Table 1 lists the features collected as well as the source for the data. Primary features are raw data and derived features are those calculated given the knowledge of primary features and other known information.

* `http://www.stanford.edu/~ahmadw/cgi-bin/index.php`
† ahmadw@stanford.edu
‡ nicolehu@stanford.edu
§ yawar@stanford.edu

Feature	Source
Primary	
House location	DYFI[a]
Damage state (CDI)	DYFI
Earthquake Magnitude	USGS[b]
Duration of shaking	USGS
Spectral acceleration at T=0.3s (ShakeMap)	USGS
Description of home damage	DYFI
Distance to epicenter	USGS
Soil type (Vs30)	USGS
Elevation	USGS
Spectral acceleration at various return periods	Google[c]
House size	USGS
House age	Zillow[d]
House price	Zillow
Derived	
Spectral displacement	
Probability of no damage	Hazus[e]
Probability of being in each of four damage states	Hazus
Probability of chimney cracking	
Distance to Fault	

[a] Data emailed by the managers of USGSs Did You Feel It? web application directly to the project team
[b] Various sources within the United States Geological Survey website.
[c] Google API
[d] Zillow real estate website API
[e] HAZUS technical manual

TABLE I: Features Collected and Their Source

Vs30 is a parameter describing soil conditions. Sa is a ground motion intensity parameter is based on the earthquake and needs to be scenario-specific inputs to the model in real time. Sd is another ground motion parameter calculated from Sa (equation 1) where T is the assumed structural period, either 0.35s or 0.4s, following HAZUS guidelines depending on the size of the home.

$$Sd = Sa(T/2pi)^2 \quad (1)$$

The fragility curve parameters depend on the structural type (construction material), size, seismic zone, and seismic design code used (which is a function of location and age of the structure) (UBC, 1997). The damage state labels are S: slight, M: moderate, E: extensive, and C: complete. P(no damage) and P(slight damage) only require Sd as an input along with stored fragility parameters. The probability of no damage and being in each of four damage states were computed using the HAZUS fragility curve parameters (HAZUS Technical Manual) assuming a wooden structure. The probable damage states for structural, non-structural drift-sensitive, and non-structural acceleration-sensitive components were computed separately.

DYFI data includes information about observed damage to walls, chimneys, etc. too. The probability of chimney cracking was computed by sorting DFYI responses into two categories: whether any type of chimney damage was reported or not. A sigmoid function was then fit through logistic regression such that the independent variable is spectral acceleration at a structural period of 0.3 seconds. Figure 2 shows the chimney fragility curve. Probability of 1 corresponds to Sa values that drove chimney damage. The sigmoid curve is fairly steep indicating there is a fairly abrupt transition from no damage to some damage for values of spectral acceleration. Thus, an empirical fragility curve was derived; the equation is shown in (2).

$$Pchim = 1/(1 + exp(3.1269 - 0.1165 * Sa)) \quad (2)$$

B. Data Pre-processing

Significant pre-processing of data was needed for many reasons: Initially, to fit within the single family home scope, all

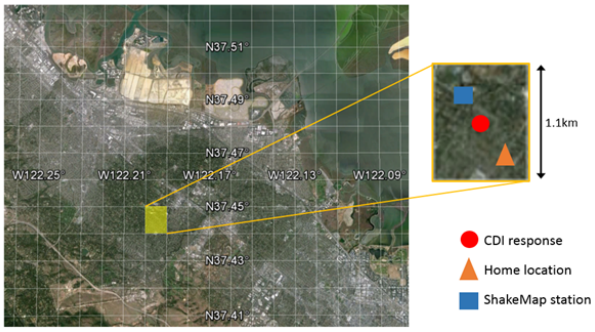

FIG. 3: Example of nearest neighbor function applied to ShakeMap and Zillow data

DYFI responses which did not list the location during the earthquake to be a single family home were removed. Next, all responses that were not geo-located by USGS were removed. Of the data from 36 earthquakes provided to the project team, data from the top 10 earthquakes (by magnitude), with at least 1000 responses remaining, were used for the training set. For privacy constraints, USGS supplied the project team DYFI data with two-digit latitude and longitude accuracy, meaning the geolocated point could be up to about 0.6 km away from the true location. Still, the location of these responses did not exactly align with the data from the other sources. Spectral acceleration information from USGSs ShakeMap website was gathered for the earthquakes. These ShakeMap files include not only data from strong motion stations throughout the state, but also the interpolated spectral ordinates using weighted contributions from three attenuation functions at regular, closely-spaced intervals. Using a nearest neighbor function, the nearest value of spectral acceleration was assigned to each DYFI response (Brown, 2007). If there was no ShakeMap data point within 1 km of a DYFI response, the DYFI response was excluded from the training set. Similarly, when appropriating housing data to a DYFI response, the nearest neighbor function was used. Figure 3 shows an example of how nearest neighbor was used to aggregate data from multiple sources to the same locations.

The final bit of data pre-processing is due to eliminate the skewness of the data towards lower to mid-level CDIs (below 8). Approximately equal number of data points pertaining to each damage state makes learning more productive and effective in future predictions. Monte Carlo simulation was used in order to increase the amount of data points for higher CDIs (above 8). The data was then randomized and features were scaled between 0 and 1. This scaling allowed the algorithm to treat each feature equally and avoided the possibility of a skewed dataset. At the conclusion of the pre-processing phase, only the most accurate data spanning the entire range of CDIs remained. This became the training dataset.

C. Models

Random forest (RF), support vector machine (SVM), and neural networks (NN) were considered for this earthquake damage estimation problem. RF was considered because it is robust in dealing with outliers, such as variation in damage states of nearby points, at the expense of relatively less predictive power. Moreover, RF is good at ignoring irrelevant data. SVM was considered because of its higher accuracy potential and theoretical guarantee against over-fitting. NN was considered because it produces an equation relating damage with the features. This equation could then be used in getting empirical relationships between damage and features.

After implementing RF, SVM and NN algorithms, damage

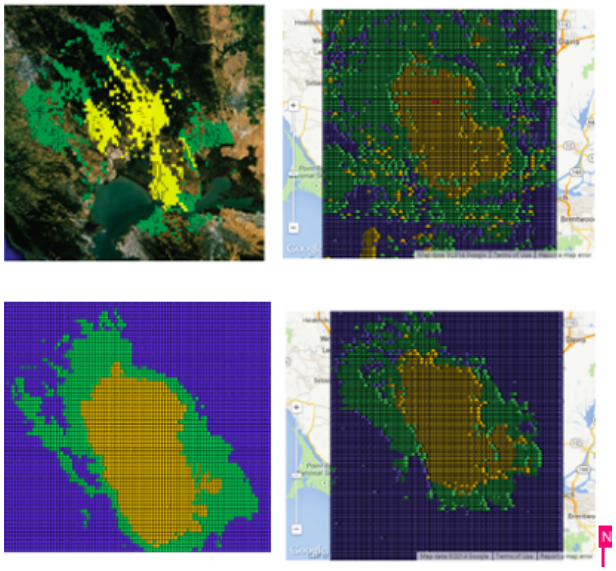

FIG. 4: Comparison of a) DYFI data to b) RF to c) NN BDI (4 hidden layers) to d) SVM BDI damage prediction results of the American Canyon 2014 earthquake

predictions for one earthquake were compared to the DYFI data. Figure 4 shows this comparison for damage predictions on a scale of 1-4 (1 being the lowest and 4 highest) for the August 2014 (Napa) earthquake. The distribution of the damage states compares well with the actual DYFI data distribution. Also, the algorithms were robust and calculated damage states for regions where no DYFI response was recorded. This can be helpful in areas where the community is not able to access DYFI quickly after an earthquake due to lack of connectivity or significant damage. It is noteworthy that the boundary between the lower two damage states is much more refined in SVM as compared to RF due to its resistance to over-fitting. Hence, SVM was considered to be the optimal machine learning model for this problem.

With the large variation that can be expected in observed damage states from an earthquake, it was decided to classify damage into one of four damage states, where each batch of damage states was given a Block Damage Index (BDI) label in lieu of a CDI label. This was reasonable based on the exclusivity and differentiability of each of the four damage states. BDI labels from existing earthquakes can be computed by equation 3. It is reasonable to assume that the general scope of damage and loss is fairly similar within the same damage state. A similar assumption is made in the PBEE approach, and structures are said to be in the same damage state if they would undergo the same degree of retrofit measures.

$$\begin{aligned} BDI = 1 \; for \; CDI <= 4 \\ BDI = 2 \; for \; CDI <= 7 \\ BDI = 3 \; for \; 7 < CDI <= 9 \\ BDI = 4 \; for \; CDI > 9 \end{aligned} \tag{3}$$

The tuning parameters for SVM, C (penalty) and (margin) were also determined. Figure 5 shows a cross-validation contour plot for a preliminary dataset. The best accuracy on the plot is 70.92%, occurring when C=5.8 and g=10.4. A Gaussian kernel was chosen as the best fit after experimenting with linear, polynomial and other RBF kernel options.

Forward and backward search methods were used to determine which features contribute to accurate damage prediction more than others. Ultimately, the parameters Vs30, Sa, Sd, P(no damage), P(slight damage), and P(chimney damage) were used.

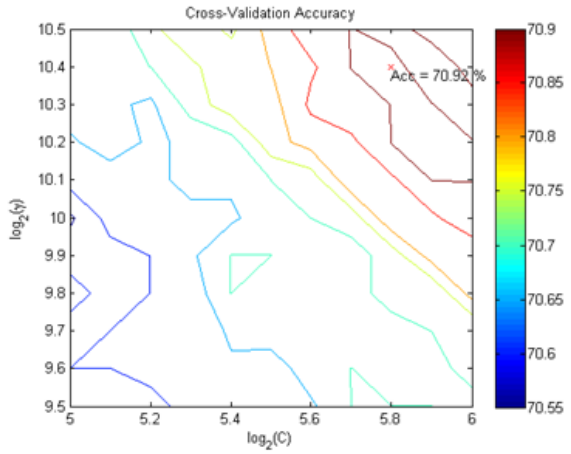

FIG. 5: Accuracy Contour Plot Example in the Cross-Validation Process

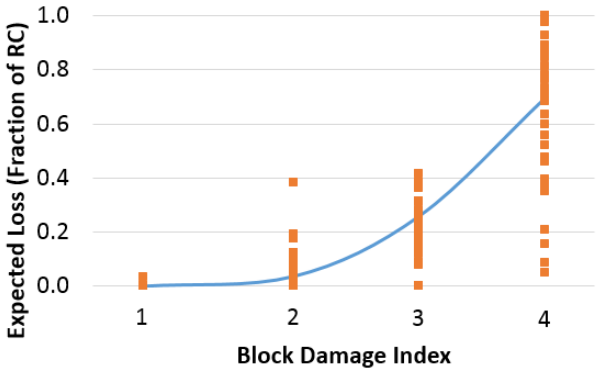

FIG. 6: Expected Loss of the Home Curve

D. Data Post-processing

In addition to the predicted BDI, expected values of economic loss and recovery time were calculated. Using the entire training set, repair cost ratios from HAZUS were used for the calculations. To calculate the expected loss, a weighted sum of the loss given damage state and the probability of being in each HAZUS damage state was done through a weighted sum technique. Structural, non-structural drift-sensitive, non-structural acceleration-sensitive, and contents were considered separately. The conditional loss parameters have been adopted from the Hazus technical manual.

The expected loss of the home is defined as the sum of expected losses for structural and non-structural elements, not including contents. A similar plot was developed for expected loss of contents. Expected annual loss (EAL) for both home and contents were calculated by numerical integration across the hazard curve from 0.01g to 5.0g using a step size of 0.01g. Recovery time was computed in a similar fashion as expected losses. Recovery parameters can be obtained from the Hazus technical manual, and include not only construction time, but also time to procure financing, design, and decision making. The mean and standard deviation of loss and recovery time at each BDI were determined and applied to each respective BDI prediction.

III. RESULTS AND DISCUSSION

The results of this study involve how well the machine learning model can predict damage states. Figure 7 shows the confusion matrix for predictions of damage for the 512 testing points. Of the 107 BDI 3s in the data, the SVM model correctly classified 97 (91%), mistook 3 for BDI 1s and 7 for BDI 2s. Additionally, of the 195 BDI 1s in the data, the SVM model correctly classified

Actual BDI \ Predicted BDI	1	2	3	4
1	172	23	0	0
2	64	137	2	0
3	3	7	97	0
4	0	0	1	6

FIG. 7: Confusion Matrix

Algorithm	Training		Testing	
	F-Score	Error	F-Score	Error
Random Forest	0.953	1.57%	0.877	17.20%
Support Vector Machines	0.834	19.55%	0.879	17.03%
Neural Networks	0.790	22.80%	0.815	21.50%

FIG. 8: Comparison of Algorithm Accuracy

172 (88%), and mistook 23 for BDI 2s. The poorest classification was of the BDI 2s, wherein 66 of the 203 were mis-classified. Thus, for this dataset, the model was mislabeling the lower levels of damage. This is non-critical considering the lower levels of damage generally do not contribute to major portion of damage as the structure remain more or less elastic.

Using the final feature list, the F score for the SVM models predictions from the American Canyon earthquake (Napa) was 0.879, and given the amount of randomness and outliers in damage predictions, this F score indicates fairly good results. Training and testing was conducted for RF, SVM, and NN algorithms. Results are shown in Figure 8.

Also, there was an attempt made to visually compare DYFI CDIs (scaled from 1-4) with predicted BDIs. Figure 4 shows BDI and CDI damage from the American Canyon 2014 earthquake for RF and SVM models. Figure 9 shows damage from the Northridge 1994 earthquake, and Figure 10 shows damage from a 2014 earthquake offshore of Northern California.

The RF, SVM and NN plots in Figures 10a), b) and c). match the general shape and damage levels of the DYFI data in 4c) very well. The SVM plot predicted smoother boundaries with fewer outliers, especially in the lower damage states. As mentioned before, the machine learning model fills in the knowledge gaps where DYFI data does not exist.

An attempt was also made to study the variation of number of hidden layers in NN, and its robustness to estimate damage. Reasonable results were observed even at four number of hidden layers. However, 1000 hidden layers showed higher BDI around water bodies in Napa area. This could be attributed to over fitting. If not, this could be interesting as it shows higher damage around soft soil deposits in Napa, which is typically the case in most earthquakes.

The DYFI data in Figure 9c) is not very extensive, and thus it

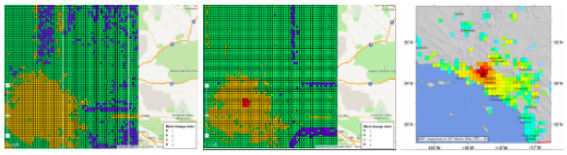

FIG. 9: a) RF BDI, b) SVM BDI, and c) Scaled CDI damage state plots of the Northridge 1994 earthquake

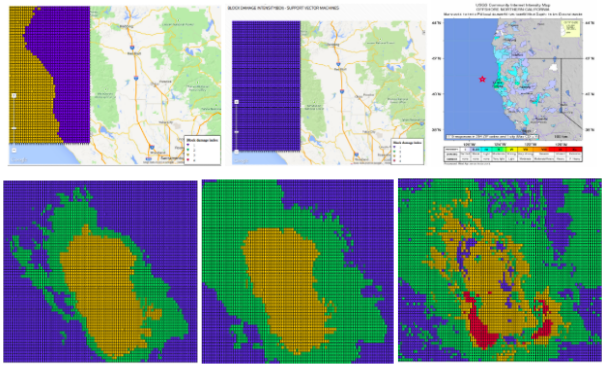

FIG. 10: a) RF BDI, b) SVM BDI, and c) Scaled CDI damage state plots of the offshore Northern California 2014 earthquake d). NN BDI (4 hidden layers). e) NN BDI (64 hidden layers). f). NN BDI (1000 hidden layers)

is somewhat difficult to visually assess the RF and SVM performance. In general, however, it appears that the trends between predicted and recorded damage are similar. The SVM better captures the higher damage states near the epicenter.

Figure 10 a),b),c) explores what happens when a magnitude 6.8 earthquake happens offshore, and thus large damage states are not expected on land. The predicted BDIs on land are all very low, matching the CDI data. Interestingly, because the soil in the ocean is very soft, the RF model predicts a higher damage state (BDI = 3). On the other hand, the SVM algorithm appears to be more sensitive to ShakeMap values (recorded on land) and less sensitive to soil type, thus predicting a low damage state for all points in the ocean.

Considering all the above results, the SVM is very stable across a broad range of earthquakes, and furnishes the best results. RF and NN, even though giving an acceptable accuracy, sometimes improbably mislabel several locations.

IV. WEB APPLICATION

A web application was created to implement the machine learning model described herein and make educated predictions with regard to the probable. The website is right now in beta version and requires one to actually request for a time slot when the administrator activates the MATLAB background programs on the server. See `http://web.stanford.edu/~ahmadw/cgi-bin/index.php` There are two main modes of the application, the homeowner mode and the community disaster response center mode.

A. For the Homeowner

The average homeowner knows little about earthquake engineering, but they are interested in their risk exposure. User inputs are few, and typically within homeowner knowledge. The website requires the user to input their homes location, replacement value of the home (which includes structural and non-structural components, but not property value), and the replacement value of contents. The website makes four BDI predictions using Sa intensities from the hazard curve corresponding to return periods of 2475, 475, 50, and 20 years. The algorithm makes 10 predictions per hazard level, takes the mean BDI, and rounds to the nearest whole number.

For the loss calculations, it takes a weighted average of all 10 iterations. Moreover, the user also gets an idea of the potential losses which he or she could face annually as well as recovery time for all the four hazard levels. This information could be

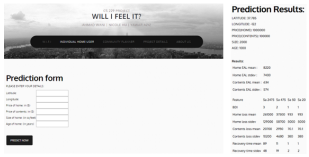

FIG. 11: Homeowner Web Module a) Input and b) Output

useful in household financial planning in order to protect assets against seismic risk. A screen capture of the homeowner web module input and output are shown in Figure 11a) and b), respectively.

B. For the Community Disaster Response Center

The community disaster response center module is used only at the onset of an event. The USGS publishes the ShakeMap within seconds after each event. The raw file can be found the in the downloads section of the ShakeMap of each earthquake. It can directly be uploaded without any pre-processing, and the website would automatically consider the spectral acceleration at 0.3s (assuming it to be a low-rise residential wooden structure).

After entering basic earthquake information like epicenter latitude/ longitude and magnitude, the application generates three maps, each of which gives a predicted damage state distribution of neighboring areas (+-100km from the epicenter) in the default mode. Figure 11 shows the community disaster response center input mode. The three output maps are generated using the SVM, RF and NN algorithms (similar to those shown through Figures 4, 9 and 10).

V. CONCLUSION AND FUTURE WORK

Despite the highly uncertain nature of earthquake engineering problems, augmenting the PBEE framework with machine learning achieved good accuracy in damage prediction, and has a promising future. The SVM helped to get a plausible representation of damage. In fact, this means machine learning can replace waiting for DYFI data when predicting community-wide damage. Further, as mentioned previously, this approach can fill in the geographic gaps in community-wide damage assessment giving near-immediate, and fairly accurate results. Despite these promising initial results, much future work can be done to further the efforts in this study. First of all, the model could be expanded beyond the state of California. Since the web application is still in the beta stage, further refinements could be made in it.

Comprehensive housing data could potentially improve damage state estimates. Additionally, expanding the scope to account for several types of structures, accounting for their current seismic health, type of construction material, and lateral resisting system would allow for better damage analysis for the community including businesses, mid-rises, etc., and thereby a more accurate estimate of loss. However, these initiatives would also require more precise DYFI data (correct up to 4 decimals) so features could be from the DYFI respondents location, and not based on the nearest neighbor. The benefit to saving lives and reducing property losses (especially in the disaster response mode) may outweigh the potential privacy concerns. This application could also help refine insurance premiums to better align with each homeowners risk exposure.

Empirical equations (extracted from paramertic learning techniques) relating damage state to the input features would be useful. As mentioned previously, Monte Carlo method was used to obtain data for higher CDI's since there was not much training data available. A potential improvement could be using the shaking intensity values of large events at other parts of the world (like Tohoku, Japan, 2010) which are not necessarily in a similar scenario, and using transfer learning techniques to extrapolate them to California. This would potentially enable the tool to predict damage states for severe catastrophes as well.

VI. ACKNOWLEDGEMENTS

The authors would like to thank Professors Gregory Deierlein and Eduardo Miranda for their insightful guidance throughout the project. In addition, we would also like to acknowledge Timothy Frank (AFIT PhD Fellow) for his insights. Finally, the authors would like to acknowledge Vince Quitoriano, Nico Luco, and David Wald from the USGS for supporting the projects vision and providing DYFI data.

VII. REFERENCES

[1] Alarifi, A., Alarifi, N., & Al-Humidan, S. (2012). Earthquake magnitude prediction using artificial neural network in northern Red Sea area. Journal of King Saud University, 24(4), 301-313.

[2] Ben-Hur, A. & Weston, J. (2010). A Users Guide to Support Vector Machines. Data Mining Techniques for the Life Sciences. Methods in Molecular Biology, 609, 223-239.

[3] Dengler, L., & Dewey, J. (1998). An intensity survey of households affected by the Northridge, California, earthquake of 17 January 1994. Bulletin of the Seismological Society of America, 88(2), 441462.

[4] Deierlein, G. (2004). Overview of a comprehensive framework for earthquake performance assessment. Performance-Based Seismic Design: Concepts and Implementation. P. Fajfar and H. Krawinkler, Eds. PEER Special Publication, PEER 2004/2005, September, 12pp.

[5] FEMA P-58-1. Seismic Performance Assessment of Buildings (2012). Federal Emergency Management Agency, Washington, D.C.

[6] HAZUS MR5. (2013). Federal Emergency Management Agency, Washington, D.C.

[7] International Conference of Building Officials, 1997. Uniform Building Code, Whittier, CA.

[8] Irfanoglu, I., & Freeman, S. (2005). Using the Earthquake Engineering Intensity Scale to Improve Urban Area Earthquake Emergency Response. Proceedings, 8th U.S. National Conference on Earthquake Engineering, San Francisco, CA.

[10] Krawinkler, H., & Miranda, E. (2004). Performance-Based Earthquake Engineering.. In Earthquake Engineering: From Engineering Seismology to Performance-Based Engineering, 41pp.

[11] Ramirez, C., & Miranda, E. (2012). Significance of residual drifts in building earthquake loss estimation. Earthquake Engineering & Structural Dynamics, 41(11), 14771493.

[12] US Census Bureau.
www.census.gov/housing/census/publications/

[13] USGS/Google/Zillow APIs.

[14] Wu, S., & Beck, J. L. (2012). Synergistic combination of systems for structural health monitoring and earthquake early warning for structural health prognosis and diagnosis.(pp. 83481Z-83481Z). International Society for Optics and Photonics.