Predicting ground shaking intensities using DYFI data and estimating event terms to identify induced earthquakes

Abhineet Gupta [1]

ABSTRACT

There has been a significant increase in earthquakes in Central and Eastern US (CEUS) in recent years. This increase in seismicity has been associated with human activities like wastewater injection, and is referred to as induced seismicity. One of the components for hazard and risk calculation from induced seismicity is the level of ground shaking expected from an earthquake at a site of interest. In CEUS, because of historically low seismicity, there is limited information to predict these shaking intensities. Even with the recent increase in seismicity, the sparsity of seismic networks limits the available information.

US Geological Survey (USGS) collects and maintains a *Did you feel it?*(DYFI) database where users report online when they feel an earthquake. DYFI data is much more widely available than other ground motion data and is used here to generate a ground motion intensity prediction model. Additionally, we assess the hypothesis that intensities generated from induced earthquakes tend to be different than those from natural earthquakes. A mixed-effects regression model is used as our primary prediction model since it allows estimation of random effects associated with earthquakes and regions. We show results from various prediction functions used in our model and conclude that the intensity predictions could not be differentiated for induced events.

1 INTRODUCTION

There has been a dramatic increase in seismicity in CEUS in recent years (Ellsworth 2013). There is a possibility that this increased seismicity in CEUS is caused by anthropogenic processes and is referred to as induced or triggered seismicity. The earthquakes are a nuisance for people and some larger magnitude earthquakes have also caused structural damage. Hence, it is important to quantify seismic hazard and risk from this increased seismicity.

One of the major components in determining seismic hazard and risk is the expected level of ground shaking at a site. Level of ground shaking from a given earthquake is typically estimated using previously collected ground motion data in a region. However, in CEUS due to historically low seismicity and sparse seismic network, there is not enough ground motion data available to constrain the prediction models. In this study, we use DYFI data which is more widely available to develop an intensity prediction model and to assess if intensities from induced events differ from natural events.

Since DYFI data is recorded on a continuous scale, regression models are best suited for prediction. Assessing the difference in predictions from different earthquakes can be achieved by using a random effects model. Additionally, another random effect for regions can also be calculated to assess how intensities vary across regions. Thus, the intensity level can be predicted from a combination of fixed effects which are a function of earthquake magnitude, depth and earthquake-to-site distance, and random effects which are computed for each earthquake and each region. Hence, a mixed effects model is utilized as the primary model to achieve the objectives of this study.

The mixed-effects regression model is also compared to a support vector regression (SVR) model. This is primarily done to assess the effectiveness of mixed-effects regression compared to other popular regression models. However, it should be noted that SVR does not have the functionality to assess random effects. Hence the second component of this study to evaluate whether intensities differ for induced events could not be achieved using the SVR model.

2 THE DATA - DID YOU FEEL IT?

USGS has been collecting DYFI data from users since year 2000. When an earthquake happens, users can visit the USGS website and answer a simple questionnaire regarding the extent to which they felt the earthquake. Based on users' responses, USGS then assigns an intensity level to each report called the Community Decimal Intensity (CDI). CDI is indicative of the Modified Mercalli Intensity (MMI) which is assigned based on the damage from ground shaking in a given region. More information about DYFI can be obtained from the paper by (Wald et al. 2012).

[1] Graduate Student, Department of Civil and Env. Engg., Stanford University, Stanford, USA

DYFI data was collected through personal communication with USGS for all earthquakes from year 2000 to 2014-10-14 with a minimum magnitude of 3 and having a minimum of 5 DYFI responses. This data was collected for earthquakes between latitudes 25° and 49°, and longitudes $-105°$ and $-67°$. This data covered all of Central and Eastern US. A total of 712,761 individual CDI data points were collected from 798 earthquakes with this criteria. The data contained CDI values and user locations. It was combined with corresponding earthquake information like magnitude and depth, and the earthquake-to-user distance was calculated.

2.1 Data processing

The data collected as mentioned above was cleaned up and processed before using it for model fitting. There was a flag in the data that marked some values as suspect. These suspect values were completely removed. Data points with a CDI of less than 2 were also removed since USGS does not calculate a finer CDI between 1 and 2. Finally, only those data points were retained that had an earthquake-to-user distance of at most 200 km since that is our range of interest. The data after this initial cleanup is shown in Fig 1.

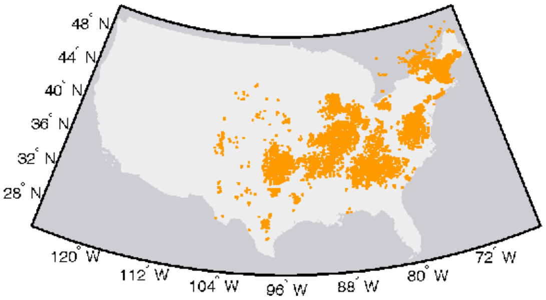

FIG. 1. DYFI data collected from Central and Eastern US

After cleaning the dataset, data points from only those earthquakes were retained which had a minimum of 5 recordings. On the next step, each user's location was assigned to a region, marked by a grid of 0.1° latitudes and longitudes. Finally, data points from only those regions were retained which had recordings from at least 5 unique earthquakes. After this processing, a dataset of 239,737 points was obtained which was subsequently used for model training.

2.2 Features

The primary features used for modeling were the earthquake magnitudes, earthquake depths and earthquake-to-user distances. Different models utilized different functions of these features. The details of feature implementation are described in the models section below. The other set of features used to model the random effects in mixed-effects regression models were the unique identifiers for each earthquake and each region.

3 TRAINING OF REGRESSION MODELS

Using the dataset described in the previous section, various models were trained and compared. The description and results from these models are shown below.

3.1 Mixed effects regression model

A mixed-effects regression model performs prediction by combining the contributions from fixed effects and random effects. In this case, the fixed effects are defined as a function of earthquake magnitude, depth and earthquake-to-user distance. The random effects are a property of each earthquake and each region, and can include the intercept terms and the slope terms associated with fixed effects features. This model can be represented by the equation below.

$$CDI = f(M, D, h) + \eta + \tau + \varepsilon$$

In the above equation, the function of earthquake magnitude M, earthquake-to-user distance D and earthquake depth h corresponds to the fixed effects. η corresponds to the random effect for earthquakes and is called the event term. Having a non-zero event term implies that the event yields an intensity which is consistently higher or lower than the mean prediction for all earthquakes. Similarly, τ defines the site term and is estimated for all regions where data was recorded. (Regions were described in the section on data processing.) ε refers to the residual for each data point.

During training of a mixed-effects regression model, coefficients for both fixed effects and random effects are determined. The random effects coefficients can be used to evaluate whether certain earthquakes or regions are more likely to cause higher or lower intensities. The results from training mixed-effects regression models are shown below.

3.1.1 Training results

Various models with different combinations of features and their functions were used during training. Results for some of these models are shown below.

Model description	η_{std}	τ_{std}	ε_{std}
Model 2 $CDI = 1 + M + D + \eta + \tau + \varepsilon$	0.296	0.262	0.924
Model 4 $CDI = 1 + M + D + \varepsilon$	NA	NA	0.974
Model 6 $CDI = 1 + M + D + h + \eta + \eta'.h + \tau + \varepsilon$	0.280	0.262	0.924
Model 7 (Atkinson et al. 2014) $CDI = 1 + M + \log(D_a) + D_a + B_a + M.\log(D_a) + \varepsilon$	NA	NA	0.962
Model 9 $CDI = 1 + M + \log(D_a) + B_a + M.\log(D_a) + \eta + \tau + \varepsilon$	0.294	0.206	0.918
Model 11 $CDI = 1 + M + \log(D_e) + B_e + M.\log(D_e) + \eta + \eta'.h + \tau + \varepsilon$	0.295	0.211	0.917

In the above models, models 1 to 6 are termed linear models since the fixed effects are linear functions; models 7 and above are termed attenuation models since their functional form is more representative of ground motion attenuation. In the model equations, parameters D_a and B_a are defined as $\sqrt{D^2 + 14^2}$ and $\max[0, \log(D_a/50)]$ respectively, as described by (Atkinson et al. 2014). Parameters D_e and B_e are defined as $\sqrt{D^2 + h^2}$ and $\max[0, \log(D_e/50)]$ respectively.

As observed from the table, model 6 for linear models and model 11 for attenuation models had the smallest residual standard deviations. A 10-fold cross-validation was performed for these two models to obtain another comparison metric. Cross-validation errors *rMSE* of 2.814 and 2.788 were obtained for model 6 and model 11, respectively.

3.2 Support vector regression

Support vector regression (SVR) model was used to compare mixed-effects regression model with another model. SVR was implemented using the algorithms developed by (Chang and Lin 2011). ε-SVR with l_1 regularization was used with cost parameter C . A linear kernel (no kernel) were implemented since it provided an equation form which could be easily implemented later. The SVR algorithm can be described as shown below.

$$
\begin{aligned}
\min_{w,b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{m}\xi_i + C\sum_{i=1}^{m}\xi_i^* \\
s.t \quad & y^{(i)} - \boldsymbol{w}^T\phi(\boldsymbol{x}^{(i)}) - b \leq \varepsilon + \xi_i, && i = 1, 2, \ldots, m \\
& \boldsymbol{w}^T\phi(\boldsymbol{x}^{(i)}) + b - y^{(i)} \leq \varepsilon + \xi_i^*, && i = 1, 2, \ldots, m \\
& \xi_i, \xi_i^* \geq 0 && i = 1, 2, \ldots, m
\end{aligned}
$$

The features used for ε-SVR were the earthquake magnitude M, earthquake depth h, earthquake-to-user distance D, epicenter distance $D_e = \sqrt{D^2 + h^2}$, $\log(D_e)$ and $M.\log(D_e)$. The output labels were the corresponding CDI at each data point. All the features and output labels were scaled to be in the range $[0, 1]$ as recommended in the algorithm implementation.

The parameter ε was fixed at 0.05. Parameter C for linear kernel was determined using grid search. Running the algorithm on the complete dataset took a very long time. Hence during grid search a subset of 25,000 data points was randomly selected for parameter selection. For linear kernel, the mean square error was minimized at $C = 2^4$. This parameter value was used to train the complete dataset.

3.2.1 Training results

While training the complete dataset, the computation time was found to be very long. Hence, at this stage of the study, the training was done on a random subset of 50,000 data points. Using ε-SVR with the parameter value described above, a residual standard deviation of 0.9623 was calculated for linear kernel. This was about the same standard deviation as obtained from mixed-effects regression model without random effects. Hence, for this set of analyses, SVR did not perform better than a mixed-effects regression model.

4 RESULTS

Observations and additional results obtained after training the models described in the previous section are shown below. Since SVR did not perform better than the mixed effects regression model, no further results were generated for SVR.

For mixed-effects regression models, it was observed that the difference in residual standard deviations among various models was not very large. However, as a general trend, the standard deviations reduced with inclusion of random effects. It was also observed that linear models had a higher standard deviation than a comparable attenuation model.

Despite the small differences in residual standard deviations, the importance of using attenuation models is demonstrated in Fig. 2. The attenuation models, as expected, better model the physical process of intensity attenuation with distance. The intensity reduces faster at smaller distances and then the decrease gets flatter. Since in hazard and risk assessment, we are usually interested in intensity behavior at short distance ranges, the attenuation models are the better choice for mixed-effects regression.

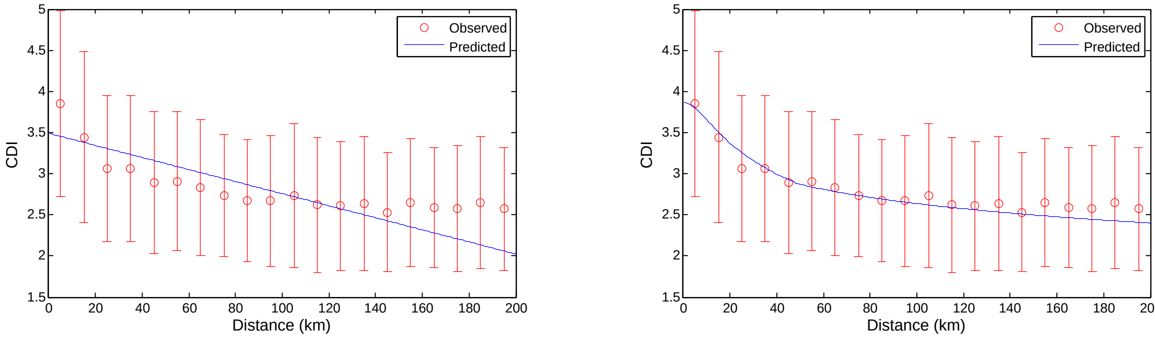

FIG. 2. Comparison of observed and predicted CDI variation with distance at M = 4 and h = 10 for model 6 (right) and model 11 (left)

To assess the second goal of this study about whether event terms differ for induced events, the event terms from model 11 were plotted on a map, as shown in Fig. 3. From a visual analysis, the event terms in regions where induced earthquakes have been observed did not appear to be different than for regions with natural earthquakes. Hence, at this point, we could not conclude that ground motion intensities generated from induced events differed from those generated from natural earthquakes.

5 CONCLUSIONS

In this study, various regression models were trained for predicting ground motion intensities from an earthquake. Intensity prediction is one of the major components in hazard and risk estimation. The models were trained using DYFI data from CEUS. Another important component of this study was to assess whether intensities from induced earthquakes differed from those from natural earthquakes. This was assessed through evaluation of random effects associated with different earthquakes. Since the regression model required both fixed and random effects, a mixed-effects model was used as the primary model for prediction. Results from mixed-effects regression were also compared with SVR, which did not include random effects. For the analyses performed in this study, SVR did not perform better than the mixed-effects regression model.

It was observed that inclusion of random effects reduced the residual error in intensity prediction. It was also observed that even though difference in residual errors were relatively small, a regression model should be selected based on the physical attenuation model for intensity to enable better prediction at shorter distances.

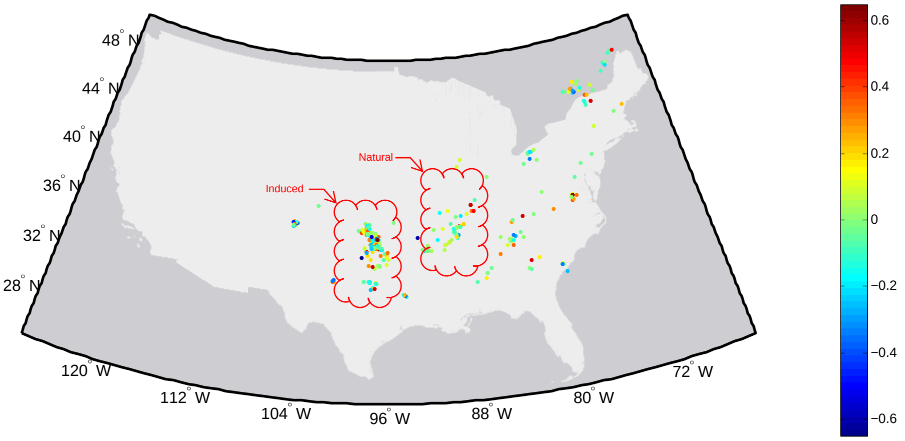

FIG. 3. Event terms from model 11 compared for induced and natural earthquakes

We also concluded that a significant difference in event terms could not be observed between induced and natural earthquakes; hence, event terms could not be effectively used to distinguish induced from natural events.

6 FUTURE WORK

The work done as part of this study was essential in understanding the effective utilization of large datasets and the implications of using different models for prediction. However, the scope of work that can be done with this dataset is very broad and there are numerous avenues for further exploration.

One of the first next steps for future assessment should be filling in the finer details in the models. SVR should be extended to the complete dataset to verify the residual standard deviations obtained from the subset of 50,000 data points. Results should be assessed further, for example, by validating the predictions from the models in this study with those by (Atkinson et al. 2014). The variation of predictions should be compared with observed data at different magnitude and distance levels. It should also be assessed if predictions have a bias due to different number of responses for each earthquake or region. There is also a possibility that different prediction models are better suited for short and long distances.

After verifying the applicability of the models, the event terms and site terms could be further analyzed. Spatio-temporal analysis could be used to determine auto-correlations. The event and site terms could also be linked to certain physical properties, for example, the faults causing the earthquakes or the soil conditions at a given region.

Finally, the prediction model can be used as an input for hazard and risk calculation. For the case of induced seismicity, risk estimation at a certain site can be used as a decision support tool to inform decisions on operations that might cause earthquakes. Thus, the prediction models developed in this study could be ultimately used in seismic risk mitigation.

REFERENCES

Atkinson, G. M., Worden, C. B., and Wald, D. J. (2014). “Intensity prediction equations for north america.” *Bulletin of the Seismological Society of America*.

Chang, C.-C. and Lin, C.-J. (2011). “LIBSVM: a library for support vector machines.” *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.

Ellsworth, W. L. (2013). “Injection-induced earthquakes.” *Science*, 341(6142), 1225942.

Wald, D. J., Quitoriano, V., Worden, C. B., Hopper, M., and Dewey, J. W. (2012). “USGS did you feel it? internet-based macroseismic intensity maps.” *Annals of Geophysics*, 54(6).