EPiC: Earthquake Prediction in California

Caroline Suen David Lo Frank Li

December 10, 2010

1 Introduction

Earthquake prediction has long been the holy grail for seismologists around the world. The various factors affecting earthquakes are far from being understood, and there exist no known correlations between large-scale earthquake activity and periodic geological events. Nonetheless, we hope that machine learning will give us greater insight into the patterns underlying earthquake activity, even if we cannot predict the time, location, and strength of the earthquakes accurately.

2 Background

With the exception of a brief period in the 1970s, earthquake prediction was generally considered to be infeasible by seismologists. Then, in 1975, Chinese officials ordered the evacuation of Haicheng one day before a magnitude 7.3 earthquake struck [15]. This led to a flurry of optimism toward earthquake prediction [9, 11], which was subsequently checked by the failed prediction of the magnitude 7.8 Tangshan earthquake of 1976. Another failure occurred in Parkfield, California in the early 1980s. Up to then, magnitude 6.0 earthquakes had occurred at fairly regular 22-year intervals. This led researchers to predict that an earthquake would strike by 1993; no such earthquake arrived until 2004. To this day, the Haicheng earthquake remains the only successful earthquake prediction in history. Critically reviewing earlier reports of successful earthquake predictions, an international panel of geologists famously concluded in 1997 that "earthquakes cannot be predicted" [4].

More recently, researchers in China have suggested that neural networks ensembles and support vector machines could be used to predict the magnitudes of strong earthquakes [5, 16], but more research needs to be done to corroborate their findings. For now the best earthquake forecast is vague at best: the USGS predicts a 63% probability of one or more magnitude 6.7 earthquakes in the San Francisco Bay Area between 2007 and 2036 [1]. Rather than attempt to issue earthquake predictions, we hope to analyze past data for periodic patterns that may advance our understanding of earthquake dynamics.

3 Methodology

3.1 Data

For data quality purposes, we decided to focus on earthquakes in California, which is heavily monitored for earthquake activity. We scraped, parsed, and removed duplicates from the following data sources:

- National Geophysical Data Center [7, 8]
- Northern California Earthquake Data Center [6]
- Southern California Earthquake Data Center [10]

Each of these data sources is freely available and contains magnitude, epicenter location, and time of occurrence information for earthquakes dating back to 1930.

The data set is not well populated before 1970, and the number of earthquakes records increases rapidly thereafter. This is partly due to advances in measurement technology and roughly corresponds to the Northern California Seismic Network being brought online.

3.2 Poisson Model

3.2.1 Overview

Our initial model sought to estimate the number of earthquakes at a particular location in the next year. In developing the model we had several relevant features to consider, including magnitude, latitude and longitude of the epicenter, depth of the epicenter, and starting time of each earthquake. For our initial model, we made two simplifications:

- **time invariance**: that the frequency of earthquakes at a particular location did not change over time
- **magnitude insensitivity**: that all earthquakes above a specified magnitude threshold C were considered equally

By not differentiating time and magnitude, we could model the number of earthquakes per year at a specified location as a Poisson distribution. That is,

$$f(loc, t) = f(loc) \sim Poisson(\lambda(loc)),$$

where f is the observed frequency of earthquakes above magnitude C at location loc in year t, and λ is the true (time-invariant) frequency of earthquakes above magnitude C at location loc.

However, earthquakes are extremely unlikely to have occurred at exactly the location of interest, so we decided to weight all earthquakes based on their distance from loc. We chose Poisson-like scaling factors similar to the standard used in locally weighted linear regression.

3.2.2 Definitions

Given a training set of m earthquakes, let $t^{(i)} \in \mathbb{R}$, $loc^{(i)} \in \mathbb{R}^2$, $mag^{(i)} \geq C$ denote the time (in years AD), location (in degrees latitude and longitude), and magnitude (in Richter units) of the ith earthquake. For notational convenience, we will rewrite location $loc = (lat, lon)$ in terms of its latitude and longitude components where appropriate.

We define the weighting function between two locations as

$$w(loc_a, loc_b) = \exp\left(\frac{-d(loc_a, loc_b)^2}{2\sigma^2}\right),$$

where d is the geodesic (great circle) distance between the two input locations and σ is a bandwidth parameter controlling the rate of dropoff of the weights.

It is easy to show that a numerically stable expression for $d(loc_a, loc_b)$ is

$$2r \arcsin\sqrt{\sin^2\left(\frac{\Delta lat}{2}\right) + \cos\left(lat_a\right)\cos\left(lat_b\right)\sin^2\left(\frac{\Delta lon}{2}\right)},$$

where $\Delta lat = lat_a - lat_b$, $\Delta lon = lon_a - lon_b$, $r =$ radius of the earth, and all latitudes and longitudes are in radians.

We now define

$$k(loc, y) = \sum_{i=1}^{m} \mathbf{1}\{t^{(i)} = y\} w(loc, loc^{(i)})$$

to be the weighted count of earthquakes at location loc in year y.

3.2.3 Mathematical Derivation

For each loc of interest, we wish to find the $\hat{\lambda}(loc)$ that maximizes the likelihood of our training set. Time invariance implies that each year's earthquakes are independent samples. For a training set spanning the years y_{start} and y_{end}, this yields

$$\hat{\lambda}(loc) = \arg\max_{\lambda} \prod_{y=y_{start}}^{y_{end}} \frac{\lambda^{k(loc,y)} e^{-\lambda}}{k(loc, y)!},$$

with an extension of the Poisson distribution to allow for nonintegral values of $k(loc, y)$.

Taking derivatives and maximizing yields

$$\hat{\lambda}(loc) = \frac{1}{y_{end} - y_{start} + 1} \sum_{i=1}^{m} \exp\left(\frac{-d(loc, loc^{(i)})^2}{2\sigma^2}\right)$$

3.3 Fourier Model

3.3.1 Overview

Our second model was motivated by periodic patterns observed in the data obtained using the Poisson model (see section 4 for more details). This model entails performing a Discrete Fourier Transform (DFT) [2] on the input data and then examining the frequency spectrum for dominant frequencies. If the input data is strongly periodic, we would expect to see several frequency peaks; conversely, if the input data is truly random, we would expect to see no discernible pattern. Therefore, to fit the input data to a Fourier model, we simply compute the DFT for the input data and pick proper coefficients for the frequencies that have the highest power. The Fourier model can be parametrized by the number of frequencies that are used.

3.3.2 Definitions

We used the Fourier model to analyze which region in California would have more earthquakes year to year. We define the north region as bounded by latitudes 42°N and 36°N and longitudes 128°W and 114°W. The south region is defined as bounded by latitudes 36°N and 30°N with the same longitude range.

Recall that $\hat{\lambda}(loc)$ estimates number of earthquakes at loc. Therefore, to count the number of earthquakes in a region, we simply sum up the $\hat{\lambda}$ values for all locations in the region, discretized to a grid of 0.1° by 0.1°.

3.3.3 Mathematical Derivation

We define the dominant frequencies chosen by the fitting process as $\omega_1, \ldots, \omega_n$ and the DFT coefficient of each value as $Y(j\omega_1), \ldots, Y(j\omega_n)$. Since we will be using a single-sided DFT, we define the amplitude component $A(\omega)$ for every frequency ω_i as $2|Y(j\omega_i)|$ for $i = 1...n$. Similarly, we define the phase $P(\omega)$ for every frequency ω_i as $\angle Y(j\omega_i)$. For simplicity, we assume that the input data has been normalized so that the mean is 0 (e.g. $Y(0) = 0$). Putting this together yields

$$y(t) = \sum_{i=1}^{n} A(\omega_i) \mathrm{Re}(\exp(j(\omega_i t + P(\omega_i)))$$

where Re denotes the real component.

4 Results and Analysis

4.1 Poisson Model

4.1.1 Results

We implemented our Poisson model fitting in Matlab to both train and predict earthquake frequency. For our initial run, we trained the Poisson model on all earthquakes after 1970, with magnitude greater than 2.0, and after some experimentation, with $\sigma = 10$km. We evaluated our model at each location on a grid of 0.01 degrees latitude and longitude, plotted the predicted earthquake frequency, and overlaid it onto Google Earth (see Table 1) [3]. Red areas are predicted to have relatively frequent earthquakes, whereas blue areas are predicted to have little earthquake activity. Our data resulted in more red areas in regions known to be earthquake prone, such as the San Francisco Bay Area and Los Angeles, providing a good first validation of our model, as the predictions appeared reasonable.

We also performed further quantification of the general accuracy of our model. We overlaid the predicted earthquake density on a map of known fault lines and a map of earthquakes recorded in the past week (see Table 1). As expected, earthquake-dense areas correspond with known fault locations, while areas that don't experience much earthquake activity do not have many nearby faults. Turning to real recorded earthquakes, we similarly observe that earthquakes tend to occur more frequently in areas marked as hot. However, we do see that the hot region near Los Angeles did not seem to have many quakes, while a not-so-hot region near San Diego is extremely earthquake dense. Because the recent data only includes one week's worth of earthquakes, we concluded that the lack of earthquakes near Los Angeles could be coincidence, or it could also point at flaws in our model. We investigate this further in the next section.

4.1.2 Validation

We performed validation on the predictive power of our Poisson model. To do so, we used cross validation where we trained the predictor on earthquakes after 1970 and with magnitude larger than 4.0, with 5 years of data withheld to use as a test set. We then compared the number of earthquakes predicted by the model versus the number of earthquakes that actually occurred, for each location. For this particular example, our test set is on the years 1980 to 1985. Table 2 shows earthquake density predicted by the model, earthquake density for the test set, and the error between the test density and the predicted density. If our model has high predictive power, then we expect the magnitude of the error to be small all around.

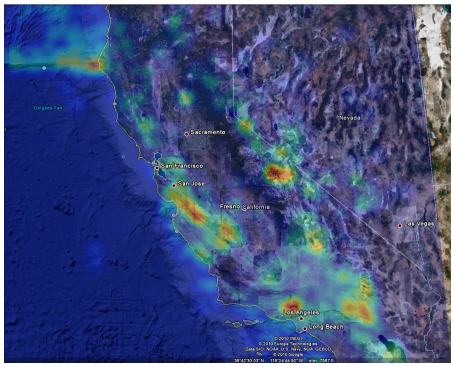
Certain regions have a significantly higher earthquake density in the test set (red regions), while other regions have a significantly lower earthquake density (dark blue regions). From this, we immediately see that the Poisson model has extremely high error rates for some seismically active regions. Indeed, we see that regions that are seismically active in 1980-1985 are not what the model predicted. Thankfully, regions that are not seismically active tend to have low error rates, as expected. However, when we compare two different test sets (1975-1980 vs 1980-1985), we see that the distribution is quite different.

As seen above, our analysis regrettably suggests that one of the central assumptions to our Poisson model, namely time invariance, is likely not valid. On further inspection, we concluded that time invariance on the frequency of earthquakes may not always be a reasonable assumption. For example, a large earthquake frequently causes numerous aftershocks of considerable magnitude afterwards, all localized in a small region in a short period of time. When a Poisson model is fitted to this type of data, it erroneously predicts an extremely high frequency of earthquakes for that region. However, even though our Poisson model may not be the optimal model for predicting the number of earthquakes for seismically active regions, it still provides a fairly accurate indication of what regions are seismically active. We expand on this idea in our Fourier model.

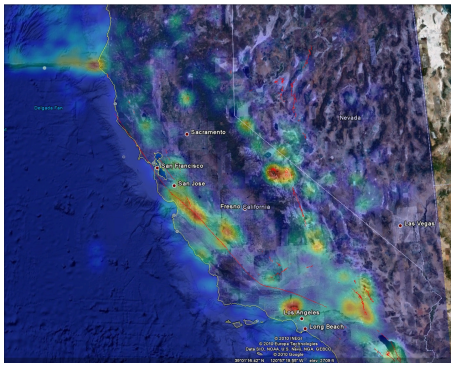
4.2 Fourier Model

4.2.1 Results

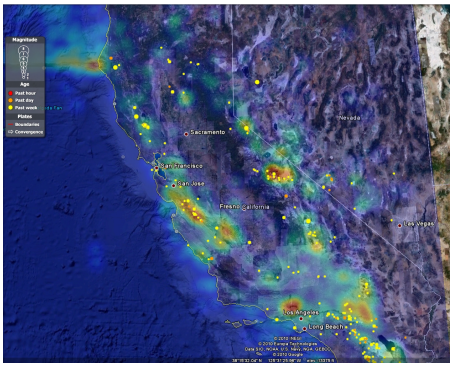
Because of the differing seismically active "hot spots" present between 1980-1985 and 1975-1980, we hypothesized that the location of the hot spots might be periodic over time. To test our hypothesis we split California into northern and southern regions, and performed Fourier analysis with differing numbers of coefficients in order to predict, for each year, whether the northern or southern region would have more earthquakes. We used a two-way split because Fourier analysis works best on continuous-valued functions, and mapping from two regions to two values that give an approximately continuous function can be done. Our dataset spanned from 1950 to 2010. For each year n years after 1950, we computed the Fourier coefficients based on the previous n years, then computed which region was expected to have a higher weight of earthquakes based on the approximation at year n.



Density of earthquakes

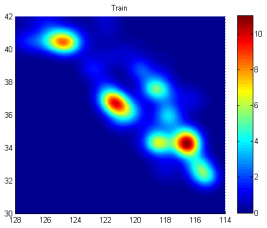


Density with fault lines

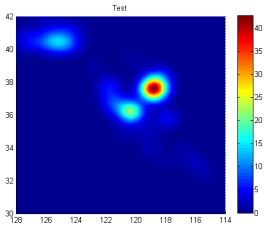


Density with recent quakes

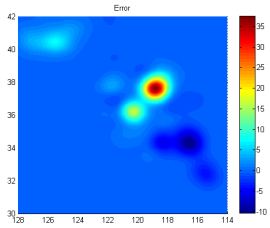
Table 1: Densities



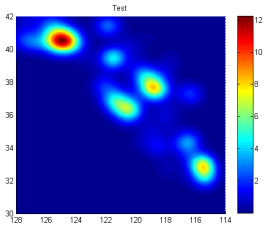
Density of earthquakes on training set (1970-2010, excluding 1980-1985)



Density of earthquakes on testing set (1980-1985)



Error between testing set and training set



Density of earthquakes on testing set (1975-1980)

Table 2: Model and test densities

4.2.2 Validation

To perform testing, we compared the predicted region with the actual region from the data. We were unable to use leave-one-out-cross-validation, as performing Fourier analysis requires the data to not contain any gaps from year to year. The graphs below show the accuracies of the predictions for every coefficient choice between one and eight and a graph of the predictions using the optimal number of coefficients, three, overlaid with the test data results (see Table 3). Test error is defined as the number of mispredicted regions divided by the total number of predictions. We slightly redefine training error, due to the sequential training process of training on year n, $n+1$, etc. Instead, we define training error as the total number of training errors over all years divided by the total number of training points over all years.

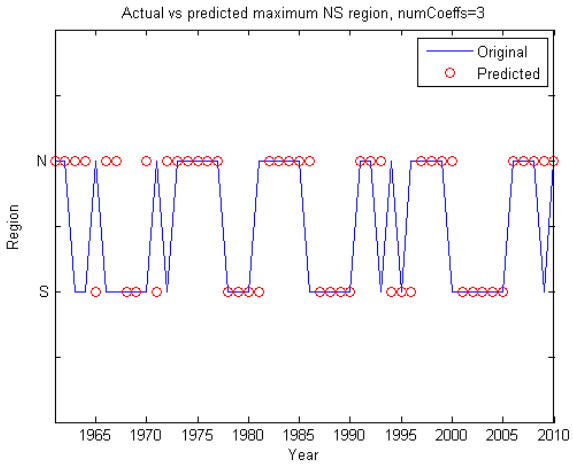
As seen by the graph, three coefficients is the optimal number of coefficients to use, as it yields the highest test accuracy. Beyond three coefficients we see that overfitting becomes an issue, as training accuracy increases to nearly 100% while testing accuracy drops significantly. At three coefficients our model has approximately 75% test accuracy. When we filter our data to consider only earthquakes of magnitude 4.0 or greater, using three coefficients yields a test accuracy of 66%. These facts suggest that the top three Fourier frequencies achieve better-than-random predictive performance, as the distribution of frequent earthquake regions is split evenly between north and south. These top three frequences (8 years, 12.8 years, and 16 years) may be of geological significance.

5 Conclusions and Next Steps

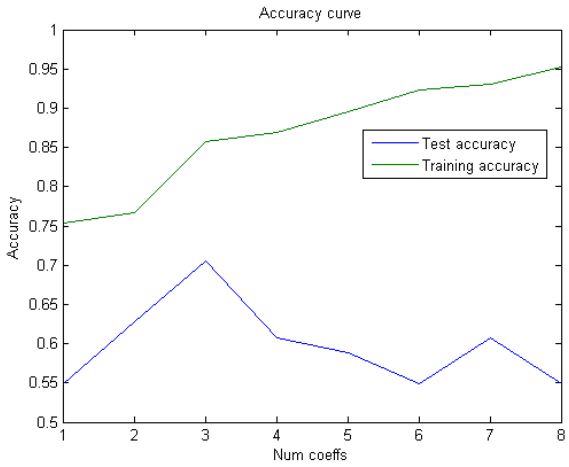
In our project we mathematically derived a weighted Poisson model that estimates the frequency of earthquakes by year for a given location. Our Fourier model extended the Poisson model to predict whether north or south California will have more earthquakes each year (in 2011, it's projected to be the northern region).

While the Poisson model has very little predictive power between years, it provides a good indicator of seismically active regions. Based on this, our Fourier model detects three major frequencies that should be investigated for geological significance. Our model's accurate and immediate predictions suggest that high-level earthquake prediction may be feasible.

It should also be possible to extend the Fourier model to be able to predict more than a binary result. For ex-



Comparison of actual and predicted



Learning curve for Fourier model

Table 3: Fourier model results

ample, California can be split into N non-overlapping bands. To perform a prediction, we would need to first do a pairwise Fourier prediction (using our existing model) between regions 1 and 2. The winner of this comparison would then be compared with region $3, 4, \ldots, N$. This approach would give a more specific prediction and could also lead to more interesting results.

As mentioned earlier, earthquake prediction has long been scoffed at as an infeasible hope. We hope that our insights into earthquake dynamics show that while today earthquake prediction might be a lofty and far-fetched goal, perhaps one day, there will be data, resources, and perhaps even machine learning algorithms available for us to understand and predict patterns of earthquakes activity.

References

[1] 2007 Working Group on California Earthquake Probabilities (2008), The Uniform California Earthquake Rupture Forecast, Version 2. Available: http://pubs.usgs.gov/of/2007/1437/of2007-1437_text.pdf

[2] Discrete Fourier Transform. Wolfram MathWorld. Available: http://mathworld.wolfram.com/DiscreteFourierTransform.html

[3] Google, KML Documentation. Available: https://code.google.com/apis/kml/documentation/

[4] Geller, Robert J. and Jackson, David D. and Kagan, Yan Y. and Mulargia, Francesco (1997), Earthquakes Cannot Be Predicted. *Science*, 5306:1616. doi: 10.1126/science.275.5306.1616

[5] Liu, Yue and Wang, Yuan and Li, Yuan and Zhang, Bofeng and Wu, Gengfeng (2004), Earthquake Prediction by RBF Neural Network Ensemble. *Advances in Neural Networks - ISNN 2004*, 3174:13-17.

[6] NCEDC, NCSN earthquake catalog. Available: http://www.ncedc.org/ncedc/catalog-search.html

[7] NOAA, The Seismicity Catalog CD-ROM Collection, 1996. Available: http://www.ngdc.noaa.gov/hazard/fliers/se-0208.shtml

[8] NOAA, Significant Earthquake Database. Available: http://ngdc.noaa.gov/nndc/struts/form?t=101650&s=1&d=1

[9] Press, Frank (1975), Earthquake Prediction. *Scientific American*, 232.5:14.

[10] SCEC, Southern California earthquake catalog. Available: http://data.scec.org/ftp/catalogs/SCEC_DC/

[11] Scholz, Christopher H., Sykes, Lynn R., Aggarwal, Yash P. (1973), Earthquake Prediction: A Physical Bases. *Science*, 4102:308-310.

[12] Stark, P. B. (1997), Earthquake prediction: the null hypothesis. *Geophysical Journal International*, 131: 495499. doi: 10.1111/j.1365-246X.1997.tb06593.x

[13] USGS, Google Earth/KML Files. Available: http://earthquake.usgs.gov/learn/kml.php

[14] USGS, Quaternary Faults in Google Earth. Available: http://earthquake.usgs.gov/hazards/qfaults/google.php

[15] Wang, Kelin and Chen, Qi-Fu and Sun, Shihong and Wang, Andong (2006), Predicting the 1975 Haicheng earthquake. *Bulletin of the Seismological Society of America*, 96.3:757-795. doi:10.1785/0120050191

[16] Wang, Wei and Liu, Yue and Li, Guo-zheng and Wu, Geng-feng and Ma, Qin-zhong and Zhao, Li-fei and Lin, Ming-zhou (2006), Support vector machine method for forecasting future strong earthquakes in Chinese mainland. *Acta Seismologica Sinica*, 19: 30-38.