Is it possible to generate good Earthquake Risk Models using Genetic Algorithms?

Blind Author 1
University 1
Address 1
Address 2
e-mail@address.com

Blind Author 2
University 1
Address 1
Address 2
e-mail@address.com

Blind Author 3
University 1
Address 1
Address 2
e-mail@address.com

Blind Author 4
University 1
Address 1
Address 2
e-mail@address.com

ABSTRACT

Understanding the mechanisms and patterns of earthquake occurrence is of crucial importance for assessing and mitigating the seismic risk. In this work we analyze the viability of using Evolutionary Computation (EC) as a means of generating models for the occurrence of earthquakes. Our proposal is made in the context of the "Collaboratory for the Study of Earthquake Predictability" (CSEP), an international effort to standardize the study and testing of earthquake forecasting models.

We use a standard Genetic Algorithm (GA) with real valued genome, where each allele corresponds to a bin in the forecast model. The design of an appropriate fitness function is the main challenge for this task, and we test three different proposals, all based on the log-likelihood of the candidate model against the training data set.

The resulting forecasting models are compared with statistical models traditionally employed by the CSEP community, using data from the Japan Meteorological Agency (JMA) earthquake catalog. Our results indicate promise for the use of GA as basis for constructing statistical earthquake forecast models. Based on these results, we identify research directions that we consider deserve more attention from the EC community.

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods and Search Heuristic Methods; J.2 [**Computer Application**]: Physical Sciences and Engineering—*Earth and Atmospheric Sciences*

GECCO'14 July 12-16,2014, Vancouver, BC, Canada.

General Terms

Real World Applications

Keywords

Earthquakes, Risk Models, Genetic Algorithms, Forecasting

1. INTRODUCTION

Earthquakes pose a huge risk for human society, in their potential for large scale loss of life and destruction of infrastructure. In the last decade, large earthquakes such as Sumatra (2004), Kashmir (2005), Sichuan (2008) and Tohoku (2011) caused terrible amounts of casualties.

Therefore, it is crucially important to understand the patterns and mechanisms behind the occurrence of earthquakes. These may allow us to create better seismic risk forecast models, with general information about about when and how often large earthquakes are expected to happen. Such information can be put to good use for mitigating damage through urban planning, civil engineering codes, emergency preparedness, et cetera.

** However, the specifics of earthquake generation mechanisms are not yet completely understood. In this sense, the development of accurate models for the risk assessment of earthquake activity is a very active area of study.

** In the study of Earthquake predictability, we highlight the RELM (Relative Earthquake Log-likelihood Model) and the CSEP

**** Evolutionary algorithms have shown themselves to be very effective in hard problems, such as (...). However, the use of EC techniques in seismology studies, and in particular in the development of earthquake risk models, is extremely scarse.

** We believe that a real contribution can be make by the use of evolutionary techniques in this field. Therefore in this paper we **** make a tutorial on earthquakes,

*** We illustrate these ideas by designing a simple genetic algorithm that creates an earthquake prediction model as defined by the CSEP. We compare this algorithm with a "skilless" model (random model), and the RI model. Using Data from the JMA catalog between 2000 and 2012

** Our Results

– Our main goal is to understand X

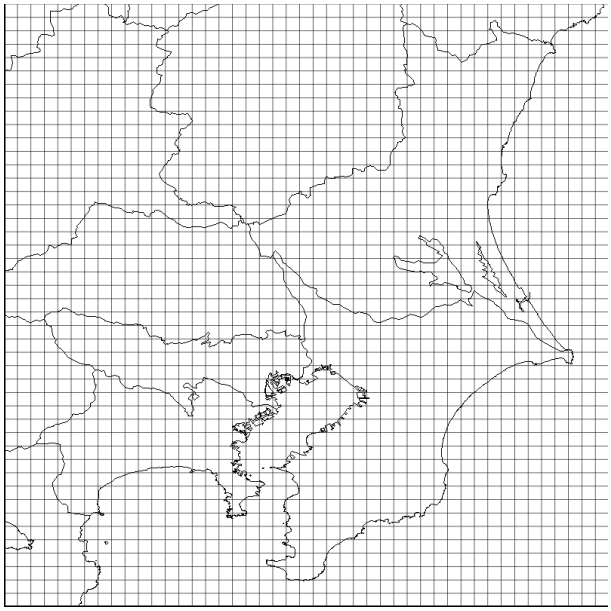

Figure 1: Target area for the Kanto region and its bins

** Outline of the paper In section 2, we describe the Earthquake predictability problem.

2. THE EARTHQUAKE PREDICTION PROBLEM

(WRITE) What is the Earthquake Prediction Problem

In this context, the Collaboratory Study for Earthquake Predictability (CSEP) is a global project to support rigorous research into the forecast models of earthquake rupture processes.

In order to objectively compare forecasts

2.1 Earthquake Forecast Model

The forecast target is defined as a geographical region, a starting date and an ending date. A forecast model will estimate the number (and sometimes the magnitude) of earthquakes happening within this target time/space.

A forecast is composed of *bins*. Each bins represents a geographical interval, measured in latitude and longitude, and sometimes a magnitude interval.

For example, in this paper we define the "kanto" region as as the area covered by latitude N34.8 to N36.3, and longitude E138.8 to E140.3. This area is divided into 2025 (45x45) bins of approximately $5km^2$ (Figure 1).

For each bin in the forecast we associate a (positive, integer) number of expected earthquakes for that bin. A good forecast is that which the number of estimated earthquakes in each bin corresponds to the actual number of earthquakes that occurs in the target time period.

2.2 The CSEP Tests

The CSEP framework uses six different tests to compare earthquake forecasts. These are divided into two groups: log-likelihood based tests, and alarm based tests.

Log likelihood tests are based on a direct analysis of the similarity between the forecast and the actual earthquake catalog [8]. A Poissonian probability distribution function is used to measure the likelihood between the forecast and the data. Three statistical tests are defined using this metric: the *L-test*, which compares a forecast with the data, the *N-test*, which compares the total number of earthquakes forecast with the total number of observed earthquakes, and the *R-test*, which compares multiple forecasts at the same time. These tests require that all the forecasts being compared have the same number and size of bins.

Alarm based tests, on the other hand, are based on a treshold analysis [10]. All bins with forecast values above the treshold are added to an "alarm set". The tests generate two values: a miss rate v, the proportion of earthquakes in the data that occurred in bins outside the alarm set; and the coverage rate τ, the proportion of the total bins in the model covered by the alarm set. Three alarm based tests are defined: the *Molcham Diagram* draws a path of v and τ based on varying values of the treshold, the *Area Skill Score (ASS)* summarizes a Molcham Diagram into a single number. A Receiver Operating Characteristic (ROC) can also be drawn using method of analysis. These testes do not take into account the total number of earthquakes, either from the forecast or the data.

2.3 The Relative Intensity Algorithm

The Relative Intensity (RI) algorithm is a commonly used benchmark for earthquake prediction models [4]. In this paper, we use it as goalpost to assess the suitability of evolutionary computation for this problem.

The working assumption behind the RI is that larger earthquakes are more likely to occur at locations of high seismicity in the past. Accordingly, the RI algorithm will estimate the number of earthquakes in a bin based on the number of earthquakes observed in the past for that bin, plus an attenuation factor that takes into account the seismicity of neighboring bins.

For more details on the implementation of the RI, please see the paper by Nanjo [4]. In the experiments in this paper, we use the following parameters: $b = 0.8$ and $s = 5$.

3. EVOLUTIONARY COMPUTATION FOR EARTHQUAKE RISK ANALYSIS

In the field of seismology and earthquake risk analysis, the few cases of Evolutionary Algorithm approaches have usually taken one of two forms. EC is often used to estimate parameter values for seismological models. These models can be used to describe and understand earthquakes.

For example, Ramos [7] uses Genetic Algorithms to decide the location of sensing stations in a seismically active area in Mexico. Nicknam et. al [5] and Kennett and Sambridge [2] used evolutionary computation to determine the Fault Model parameters (such as epicenter location, strike, dip, etc) of a given earthquake. Evolutionary Computation has also been used to estimate the peak ground accelleration of seismically active areas [3, 1].

EC has also been used very few times as a method to generate seismic forecast model. One such approach is described by Zhang and Wang [11], where they fine tune a Neural Network with a GA, and then use this system to produce a forecast. Unfortunately that work did not provide enough

information to reproduce the proposed GA+ANN system or their results.

4. A FORECAST MODEL USING GENETIC ALGORITHMS

To investigate the ability of Evolutionary Computation to generate earthquake forecast models, we design and test a simple Genetic Algorithm. Let us call this system the *GAModel*.

An individual in GAModel will encode a forecast model as defined in the CSEP framework (2.1). The population will be trained on earthquake occurrence data for a fixed training period. The best generated individual will be compared with a model generated by the RI algorithm, and a skilless (random) model. This comparison is based on earthquake occurrence data for a test period immediately posterior to the training period.

By encoding the entire forecast model as one individual, we identified two main concerns while designing GAModel: First, as the forecast model contains a large number of bins, the genome of an individual will be respectively large. Evolutionary operators and parameters must be chosen carefully to guarantee convergence in a reasonable time, and avoid local optima.

Secondly, the design of the fitness function deserves a lot of attention, to avoid the risk of the system overfitting to the training data.

4.1 Genome Representation

In GAModel, each individual represents an entire forecast model. The genome is a real valued array X, where each element corresponds to one bin in the desired model (the number of bins n is defined by the problem). Each element $x_i \in X$ can take a value from $[0, 1)$. In the initial population, these values are sampled from a uniform distribution.

In the CSEP framework, a model is defined as a set of integer valued expectations, corresponding to the number of predicted earthquakes for each bin. To convert from the real valued chromossome to a integer forecast, we use a modification of the Poisson deviates extraction algorithm from [6] (Chapter 7.3.12).

Algorithm 1 Obtain a poisson deviate from a $[0, 1)$ value

```
Parameters 0 ≤ x < 1, μ ≥ 0
L ← exp(−μ), k ← 0, prob ← 1
repeat
  increment k
  prob ← prob * x
until prob > L
return k
```

In Algorithm 1, x is the value taken from the chromossome, and μ is the average number of earthquakes observed across the entire training data. Note that in the original algorithm, $k - 1$ is returned. Because the log likelihood calculation used for model comparison discards forecasts that estimate 0 events in bins where earthquakes are observed, we modify the original algorithm here to make sure all bins estimate at least one event.

4.2 Fitness Function

The main challenge when applying an Evolutionary Computation method to any new application domain is usually the definition of an appropriate fitness function. Accordingly, the largest part of our effort in this work was to define a good fitness function for GAModel.

4.2.1 Simple Log Likelihood Fitness Function

Our first candidate was the log-likelihood between the forecast generated by an individual and the observed earthquakes in the training data, as described by Schorlemmer et. al. [8]. In simple terms, the log likelihood is a measure of how close a forecast is to a given data set.

Let $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n | \lambda_i \in \mathbb{N}\}$ be a forecast with n bins. In this definition, λ_i is the number of earthquakes that is forecast to happen in bin i. To derive Λ_X from an individual $X = \{x_1, x_2, \ldots, x_n | 0 \leq x_i < 1\}$, we calculate each λ_i from x_i using Algorithm 1.

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n | \omega_i \in \mathbb{N}\}$ be the observed earthquakes in the training data. The log likelihood between an individual's forecast Λ_X and the observed data Ω can be calculated as:

$$L(\Lambda_X|\Omega) = \sum_{i=0}^{n} -\lambda_i + \omega_i * \ln(\lambda_i) - ln(\omega_i!) \qquad (1)$$

There are two special cases that arise when any $\lambda_i = 0$. If $\lambda_i = 0$ and $\omega_i = 0$, then the value of the sum for that element is 1. If $\omega_i > 0$, then $L(\Lambda_X|\Omega) = -\infty$ and the forecast must be discarded. For more details on this, see [8].

In the Simple Log Likelihood fitness function, the value of $L(\Lambda_X|\Omega)$ is taken as the fitness value of the individual.

Early testing with the Simple Log Likelihood function showed that GAModel had a very strong tendency to overfit to the training data. This is natural, since there are differences between the seismocity of a larger period and a shorter one.

To avoid this overfitting, we have developed two alternative fitness functions.

4.2.2 Simulated Log Likelihood Fitness Function

In the Simulated Log Likelihood fitness function, we generate k "simulations" from the observed data Ω, and compare an individual's forecast with all the simulated data.

Let $\hat{\Omega} = \{\hat{\omega}_1, \hat{\omega}_2, \ldots, \hat{\omega}_3 | \hat{\omega}_i \in \mathbb{N}\}$ be a simulated data set generated from Ω. Each $\hat{\omega}_i$ is taken randomly from a Poissonian distribution with mean ω_i.

To calculate the Simulate Log Likelihood fitness, we calculate the Log Likelihood of the forecast Λ_X generated by an individual X against each of the k simulated data sets $(L(\Lambda_X|\hat{\Omega}_1), \ldots, L(\Lambda_X|\hat{\Omega}_k))$. We use the lowest of the k log likelihood values as the fitness of X.

The idea behind this fitness function is that by training against a number of similar, but slightly different training data sets at the same time, we might be able to avoid overfitting the forecast to the observed training data set.

4.2.3 Time-slice Log Likelihood Fitness Function

In the time-slice log likelihood fitness function we break up the training data set into smaller *slices*. These slices are based on the chronology of the earthquakes contained in the training catalog. The duration of each slice is the same as the duration of the test catalog.

Let's consider an example where the target period for the forecast is one year, from 1/1/2014 to 1/1/2015, and the

training data is taken from the 10 year period of 1/1/2004 to 1/1/2014. The time-slice log likelihood fitness function will divide the training data into ten 1-year periods, from 2004 to 2005, 2005 to 2006, and so on.

When an individual X is evaluated, we calculate the log likelihood of its forecast Λ_X against each of the time slices $(\Omega_{2004}, \Omega_{2005}, \ldots, \Omega_{2013})$. The lowest value is used as the fitness of X.

The idea behind this fitness function is that the catalog data available for training will normally span a period of time much longer than the desired forecast. By breaking the training data into smaller periods, we are trying to make the evolutionary algorithm learn any time-repeating pattern that might exist in the data.

4.3 Evolutionary Operators and Parameters

GAModel uses a regular generational genetic algorithm. For selection, we use Elitism and Tournament selection.

For the crossover operator we use *aNDX*, recently proposed by Someya [9]. If the value of a chromossome after the crossover falls outside the $(0, 1]$ boundary, it is truncated to these limits. For the mutation operator, we sample entirely new values from $(0, 1]$ for each mutated chromossome.

Population Size	500
Generation Number	100
Elite Size	1
Tournament Size	50
Crossover Chance	0.9
aNDX parents	4
aNDX zeta	0.5
aNDX sigma	1
Mutation Chance (individual)	0.8
Mutation Chance (chromossome)	$(\text{genome size})^{-1}$

Table 1: Parameters for GAModel

The parameters used for the evolutionary computation are described in Table 1. At this stage, we are not yet particularly worried with convergence speed of the system. Because of this, not a lot of effort was spent fine tuning these paremeter's values. Instead, these values were choosen by trial and error on a data region not used in the experiments of section 5.1, until an acceptable convergence time was found.

5. EXPERIMENTS

5.1 Experimental Data

The data used in these experiments comes from the Japan Meteorological Agency's (JMA) catalog. The catalog lists earthquakes recorded by the sensing station network in Japan, along with the time, magnitude, latitude, longitude and depth of the hypocenter.

The part of the catalog used in this work spans a period from 2000 to 2013, and has over 220.000 earthquakes recorded in the Japanese archipelago. In order to better understand the predictive power of GAModel, we divide the catalog into 5 areas, which are used in the experiments below. The relative position and size of these areas can be seen in Figure **??**.

1. **All Japan:**
2. **Sendai:**
3. **Kantou:**
4. **Kansai:**
5. **Touhoku:**

5.2 Experimental Design

To test the ability of a Genetic Algorithm to generate an effective earthquake risk model, we perform a forecast experiment. In this experiment we compare three forecasts: a randomly generated forecast, one generated by the RI algorithm, and a group of forecasts generated by the proposed method.

We define a "scenario" as a region and a forecast period. In this experiment, we will test 5 regions (all japan, pacific, kanto, tohoku, kansai), and 7 one-year periods (from 2005 to 2012), for a total of 40 scenarios.

Each scenario has a corresponding training window, which consists of the 5 years prior to the scenario's forecast time. This training period is used by the RI algorithm and by the proposed method to generate their respectives forecasts. For the sake of statistical testing, we generate 20 forecasts using the GAModel. Unless noted otherwise, all results reported are an average of these runs.

We compare the forecasts in three different ways: Using the log-likelihood between the forecast and the testing data, Using the ASS value, and visually by comparing the forecast maps of some significative scenarios.

5.3 Results and Comparison of forecast models

6. CONCLUSION

– We choose the parameters hapzardly - a more careful parameter selection might improve the result (or at least the running time)

– There is a large number of problems related to the development of forecast models: Parameter optimization, model construction, earthquake clustering, etc.

– On the evolutionary side, we think GA or ANNs might be a good idea.

– We encourage people to work on this with us.

6.1 Future Work

Acknowledgements

TODO: Check with Bogdan if we need to provide Acknowledgements for the JMA data.

7. REFERENCES

[1] A. F. Cabalar and A. Cevik. Genetic programming-based attenuation relationship: An application of recent earthquakes in turkey. *Computers and Geosciences*, 35:1884–1896, October 2009.

[2] B. L. N. Kennet and M. S. Sambridge. Earthquake location âĂŤ genetic algorithms for teleseisms. *Physics of the Earth and Planetary Interiors*, 75(1–3):103–110, December 1992.

[3] E. Kermani, Y. Jafarian, and M. H. Baziar. New predictive models for the v_{max}/a_{max} ratio of strong ground motions using genetic programming. *International Journal of Civil Engineering*, 7(4):236–247, December 2009.

[4] K. Z. Nanjo. Earthquake forecasts for the csep japan experiment based on the ri algorithm. *Earth Planets Space*, 63:261–274, 2011.

[5] A. Nicknam, R. Abbasnia, Y. Eslamian, M. Bozorgnasab, and E. A. Mosabbeb. Source parameters estimation of 2003 bam earthquake mw 6.5 using empirical green's function method, based on an evolutionary approach. *J. Earth Syst. Sci.*, 119(3):383–396, June 2010.

[6] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes, The Art of Scientific Computing*. Cambridge University Press, third edition, 2007.

[7] J. I. E. Rmaos and R. A. Vázques. Locating seismic-sense stations through genetic algorithms. In *Proceedings of the GECCO'11*, pages 941–948, Dublin, Ireland, July 2011. ACM.

[8] D. Schorlemmer, M. Gerstenberger, S. Wiemer, D. Jackson, and D. A. Rhoades. Earthquake likelihood model testing. *Seismological Research Letters*, 78(1):17–29, 2007.

[9] H. Someya. Striking a mean- and parent-centric balance in real-valued crossover operators. *Transactions on Evolutionary Computation*, 17(6):737–754, December 2013.

[10] J. D. Zechar and T. H. Jordan. The area skill score statistic for evaluating earthquake predictability experiments. *Pure and Applied Geophysics*, 167(8–9):893–906, August 2010.

[11] Q. Zhang and C. Wang. Using genetic algorithms to optimize artificial neural network: a case study on earthquake prediction. In *Second International Conference on Genetic and Evolutionary Computing*, pages 128–131. IEEE, 2012.