Earthquake Risk Models with Genetic Algorithm Hybridization

Blind Author 1University 1
Address 1
Address 1
email@address.com

Blind Author 1University 1
Address 1
Address 1
email@address.com

Blind Author 1University 1
Address 1
Address 1
email@address.com

ABSTRACT

This project main goal is to refine a method, called the GAModel, by obtaining better accuracy in seismic risk analysis. The GAModel is a model which aims to generate forecasts by using only Evolutionary Computation (EC).

This document sumamarizes the GAModel and proposes three extended versions for the GAModel. The first model, ReducedGAModel, is similar to the GAModel, tough it has a different genotype representation. In the GAModel the genotype is related with the fenotype, and in ReducedGAModel the genotype is a version non-related with the fenotype. The second model named as Emp-ReducedGAModel, as the reducedGAModel, uses this new genotype idea and incomporates some emperical laws such as the modified Omori-Utsu formula. The last one, the Emp-GAModel, recovers the genotype-fenotype relation from the GAmodel but also incomporotes the same emperical laws that composes the ReducedGAModel.

These four models were evaluated and compared based on the predictabily experiments framework proposed by the "Collaboratory for the Study of Earthquake Predictability" (CSEP), an international effort to standardize the study and testing of earthquake forecasting models. The experiments were designed to compare 1-year earthquake rate forecasts for four regions in Japan in using the data from the Japan Meteorological Agency (JMA) earthquake catalog.

Keywords

Evolutionary Computation, Genetic Algorithms, Forecasting, Earthquakes

1. INTRODUCTION

Earthquakes may cause soil rupture or movement, tsunamis and more. They may cause great losses and that can be explicit by some examples such as the earthquakes in Tohoku (2011) and Nepal(2015). To be able to minimize the consequences of these events, we look to create forecast earthquake occurences models. Hence the characteristics of the earthquakes may vary both in time and place, these methods should be to adapt their behavior to be able to forecast earthquakes which follows the reality.

This project aims to obtain a better method, based in improvements to the GAModel [1], a statistical method of analysis of earthquakes risk using the Genetic Algorithm technique (GA). Two ideias where taken into account for this. The first, is to change the candidate solution representation. By that, we objective to make the GAModel more specialized, focusing only on areas on which earthquakes happened already.

The other ideia is based on the assumption that earthquakes cluster in both space and time, and the ideia is to apply the Genetic Algorithm technique (GA) with a some empirical laws, such as the modified Omori law. First, the background intensity (the independent earthquakes or mainshock), which is a function of the space, is forecasted using the GA. Then, we use some empirical laws to obtain the dependent earthquakes (aftershock) for a specific time interval.

With this two new ideas, we developed three new methods. We named them as the ReducedGAModel, the Emp-GAModel and the Emp-ReducedGAModel. The first one represents the idea of focusing on areas with occurences. The second one, adds what we call as the domain knowledge, using empirical laws yet to be described, on the section 2. The last one, is a mixed between the two ideas, which means that it not has the new representation as it uses the empirical laws as well.

The models resulted of those methods were analyzed using likelihood tests, namely the L-test, the N-test and the R-test, as suggested by Regional Earthquake Likelihood Model (RELM) [5].

For developing the methods and to be able to compare them we used the earthquake catalog from the Japanese Meteorological Agency (JMA), using event data from 2005 to 2010.

This paper is organized as: in section 2, we give a details of each of the forecast proposed covering the Collaboratory for the Study of Earthquake Predictability (CSEP) framework and the empirical laws. In section 3, we give the description of the tests proposed in [6]. After that, in 4, we define the target areas used for the experiment and the data from the JMA; we clarify the design followed during the experiments and how ew compared the models derived from our methods. Finally, we show the results and conclude this work in 5 and 6.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

2. 1-YEAR MODELS

In the Collaboratory for the Study of Earthquake Predictability (CSEP) framework, a forecast uses one common format which is a gridded rate forecast [10]. For this format a geographical region, during a start date and an end date, is divided in sections, the bins. The forecast will estimate the number (and sometimes the magnitude) of earthquakes that happens in this target region, during the target time interval, considered to be of one year for this study [1].

Large and independent earthquakes, also known as mainshocks, are followed by a wave of others earthquakes, the aftershocks [7]. Hence there is no physical measurement to identify mainshocks and its aftershocks [7], we divided the models in two groups: the ones that only forecasts mainshocks and those that forecast both mainshocks and aftershocks.

2.1 Genome Representation

The GAModel each individual represents an entire forecast model [1]. For the ReducedGAModel and each individual represents a subarea of the forecast model. This subarea is related to earthquakes past events locations only, so the genome size is usually smaller than the one used in the GAModel and the Emp-GAModel.

The Emp-ReducedGAModel and the Emp-GAModel are only different from the ReducedGAmodel and from the GAmodel, respectively, by the use of equations after the forecast is provided. This means that the theirs genome representation are the same as the GAModel and the ReducedGAModel,correspondingly.

For all methods, 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 $x_i \in X$ takes a value from $[0, 1)$. In the initial population, these values are sampled from a uniform distribution. For more details of the genome represatation, please refer to [1].

2.2 Fitness Function

For the fitness function we used the log-likelihood value. The fittest individual among all the others, is preserved in the next generation, to make the solution of one generation as good as the its last generation. The bins, a gene (a element) of the genome representation, b_n, define the set β and n is the size of the set β:

$$\beta := b_1, b_2, ..., b_n, n = |\beta| \tag{1}$$

The probability values of the model j, expressed by the symbol Λ, is made of expectations λ_i^j by bin b_i. The vector is define as:

$$\Lambda^j = \left(\lambda_1^j, \lambda_2^j, \ldots, \lambda_i^j\right); \lambda_i^j := \lambda_i^j(b_i), b_i \in \beta \tag{2}$$

The vector of earthquake quantity expectations is defined as: earthquake by time. The Ω vector ÃI' composed by observations ω_i per bin b_i, as the Λ vector:

$$\Omega = \left(\omega_1, \omega_2, \ldots, \omega_i\right); \omega_i = \omega_i(b_i), b_i \in \beta \tag{3}$$

The calculation of the log-likelihood value for the ω_i observation with a given expectation λ is defined as:

$$L(\omega_i|\lambda_i^j) = -\lambda_i^j + \omega_i \log \lambda_i^j - \log \omega_i! \tag{4}$$

The joint probability is the product of the likelihood of each bin, so the logarithm $L(\Omega|\Lambda^j)$ is the sum of for $L(\omega_i|\lambda_i^j)$ every bin b_i:

$$\begin{aligned} L^j = L(\Omega|\Lambda^j) &= \sum_{i=1}^{n} L(\omega_i|\lambda_i^j) \\ &= \sum_{i=1}^{n} -\lambda_i^j + \omega_i \log \lambda_i^j - \log \omega_i! \end{aligned} \tag{5}$$

The fitness function is a coded version of the equation 5. It uses the probabilities of the bins of each individual of model for the λ values.

2.3 Evolutionary Operators and Parameters

2.4 Mainshock Models

The GAModel is purely based on the framework suggested by the CSEP. In it, one forecast is defined as a place for a specfic time and is divided in bins. Each bin represents a geographical interval. The whole target area of study is covered by these bins and represents the $\mu(x, y)$, the background intensity [11]. In the GAModel, each possible solution is represented as an entire forecast model.

In this context the GAModel is considered as one method to generate space-rate-time forecasts. It also could be described as:

$$\lambda(t, x, y, M|\Upsilon_t) = \mu(x, y) \tag{6}$$

where you can denote the number of earthquakes forecast in a bin as $\lambda(t, x, y)$ [10] given that Υ_t is the earthquake observation data up to time t.

The ReducedGAModel, which represents the first ideia (see 1), is a method described as the GAModel, but each possible solution represents a fraction of the forecast where we expect to find especific risk areas. The GAModel defines a expected number of earthquakes for every single bin in the target region. That could lead to exhaustive and, sometimes worthless, searches. That is caused by the number of bins in the forecast and also because some in some bins there are no earthquake occurances in the observation data. To minimize this, the ReducedGAModel only define expected number of earthquakes in bins that already had some occurance in the past, giving some sort of guideline to the GA.

To make it clear, we use the same example as the one used in [1].The "Kanto" region, one of the four areas used in this study, is divided into 2025 bins (a grid of 45x45 squares). Each bin has an area of approximately 25km2 (Figure ??). The GAModel then calculates an expected number of earthquakes for every bin on a determinated time interval, so the GA searches for good values in 2025 bins.

The ReducedGAModel will first obtain the position of past occurances and will calculate some expected number of earthquakes for only the bins related to those positions. For example, if there are 10 bins with occurances in Kanto in the last year, it will make the GA search good values for only those 10 bins, leaving the other 2015 bins with the value zero, representing zero occurances. It is important to highlight that in the worst case, it will make the same amount of searches as the GAModel.

2.5 Mainshock+Aftershock Models

Hence earthquakes cluster in space and inspired by the space-time epidemic-type aftershock sequence (ETAS), the Emp-GAModel, represents the second ideia (see 1) and is described as:

$$\lambda(t, x, y, M|\Upsilon_t) = \mu(x, y)J(M) \tag{7}$$

adicionar RI nanjo no lugar da F

$$\lambda(t, x, y|\Upsilon_t) = \mu(x, y) + \sum_{t_i \in t} K(M_i)g(t - t_i)F \tag{8}$$

The Emp-GAModel uses $\mu(x, y)$ as defined for the GAModel, so it calculates an expected number of earthquakes for every bin in the target region. The Omori law, $g(t)$, which is considered one empirical formula of great success [11] [8] [4], is a power law that relates the earthquake occurance and its magnitude with the decay of aftershocks activity with time. For this approach we used the probabilty density function (pdf) form of the modifed Omori law [11]:

$$g(t) = \frac{(p-1)}{c(1+\frac{t}{c})^(-p)} \tag{9}$$

In the paper [8], Utsu says that most p and c values, for various earthquake data sets fall in the range between 0.9 and 1.4, and between 0.003 and 0.3 days, respectively. These values were based on the Davidon-Fletcher-Powell optimization procedure and used in ETAS [8]. For the experiments done for this paper, we choose the values of 1.1 for p and 0.003 for c, arbitrarily.

For $K(M_i)$, the total amount of triggered events, we count aftershocks within a calculated aftershock area, A, using the formula, where M_c is the magnitude threshold:

$$K(M_i) = A\ exp([\alpha(M - M_c)]) \tag{10}$$

From [3], states alpha as the inverse of the magnitude of an event, or $magnitide^{-1}$. To obtain A, the following equation from [9], was used:

$$A = e^{(1.02M-4)} \tag{11}$$

and lastly, the $J(M)$ is a simulation of the event magnitude by Gutenberg-Richter's Law, using 1 as the value of β [2]:

$$J(M) = \beta e^{-\beta(M-M_c)}, M \geq M_c \tag{12}$$

At last, the Emp-ReducedGAModel is a mix between the two ideias, so it is represented as the Emp-GAModel but its candidates take the same form as the ones in the ReducedGAModel.

3. TESTS FOR EVALUATING MODELS

In the paper *Earthquake Likelihood Model Testing* [5] is proposed some statistical tests that are used in this study to compare and evaluate the models, developed by the The Regional Earthquake Likelihood Models (RELM). These tests are based on the log-likelihood score that compares the probability of the model with the observed events.

To evalute the data-consistency of the models we used the N-Test, the Number Test, and the L-Test, or Likelihood Test. These tests fall are significance tests. Therfore, assuming a given forecast model as the null hypothesys, the distribuition of an observable test is simulated. If the observed test statistic falls into the upper or lower tail of this distribuition, the forecast is rejected [7].

To be able to compare the model that passed the N-Test and the L-test, the R-Test, hypotheses Comparison Test is used. It calculates the relative performance of a model, by comparing the Log-likelihood values between two models.

3.1 Likelihood Test or L-Test

The L(ikelihood) Test considers that the likelihood value of the model is consistent with the value obtain with the simulations. The value is calculated by the formula, where $\widehat{L}_k$ is the value of the Log-likelihood of the model j, in the *bin* i and $\widetilde{L}$ is the value of the Log-likelihood of the simulation j in the *bin* q:

$$\gamma_q^j = \frac{\left|\left\{\widehat{L}_k^j | \widehat{L}_k^j \leq \widetilde{L}_q^j, \widehat{L}_k^j \in \widehat{L}^j, \widetilde{L}_q^j \in \widetilde{L}^j\right\}\right|}{|\widehat{L}^j|} \tag{13}$$

The analisys of the results can be splited into 3 categories, as follows:

1. Case 1: γ^j is a low value, or in other words, the Log-likelihood of the model is lower then most of the Log-likelihood of the simulations. In this case, the model is rejected.
2. Case 2: γ^j falls near the half of the values obtained from the simluations and is consistent with the data.
3. Case 3: γ^j is high. This means that the Log-likelihood of the data da is higher that the Log-likelihood of the model and no conclusion can be made what so ever.

It is important to highlight that no model should be reject in case 3, if based only on the L-Test. In this case the consistency can or cannot be real, therefore these model should be tested by the N-Test so that further conclusions can be done.

3.2 Number test or N-Test

The N(umber)-Test also analises the consistency of the model, but here, it compares the number os observations with the number of events of the simulations. This test is necessary to supply the underpredicting problem, which may pass unnotecied by the L-Test.

This mesures is estimated by the fraction of the total number of observations by the total number of observations of the model.

As the L-test, if the number of events falls near the half of the values of the distruition, then the model is consistent with the observation, nor estimating too much events nor few of them.

3.3 Hypotheses Comparison Test or R-Test

The Hypotheses Comparison, or the R(atio)-Test, compares two models against themselves. The log-likelihood is calculted for both models and then the difference between, the observed likelihood ratio, this value indicates which one of the model better fits the observations.

The likelihood ratio is calculated for each simulated catalog. If the fraction of simulated likelihood ratios less than

the observed likelihood ratio is very small, the model is reject. To make this test impartial, not given an advantage to any model, this procedure is applied symmetrically [7].

3.4 Evaluation

The evaluation process is made as follow: First, the data-consitency is tested by the L-Test and the R-test. If the model passes these tests, meaning that it was not rejected by them, it is compared with other models, which also were not reject, with the R-Test. The model that best fits the R-Test is then chose as the best model [5].

4. EXPERIMENTS

data, design, comparacao, regions, year(s)

To analyze the performance of the forecasts generated by the GAModel, the ReducedGAModel, the Emp-GAModel and finally the Emp-ReducedGAModel, we used the evalution method proposed by [5] and described in section 3.4.

Objecting a better understand of the patterns that most influence the earthquakes events and also to be able to determine the qualities of those forecasts, the data of the JMA catalog was divided into four groups. Each group constituent only of earthquakes that happened in a specific time interval for a given area of Japan. The experimental data will be described in details subsequently.

4.1 Experimental Data

The data used in these experiments comes from the Japan Meteorological Agency's (JMA) catalog. It is a list of earthquakes events which took place in Japan from 2000 to 2013. Each event is characterized by some typical earhtquake information such as magnitude, latitude, longitude, and depth.

For the experiments we consider events with magnitude above 5.0 which happened in four specific areas of Japan. Those areas (Kanto, Tohoku, Kansai and East Japan) represent different earthquake attributes and could lead to more information about the power of the forecasts and/or its pitfalls. Kanto, Touhoku and Kansai contain mainly inland earthquakes, which are considered to follow more stable patterns. East Japan includes also many off-shore earthquakes. For more information about the four regions as well as a map which locates them in Japan, please refer to [1].

4.1.1 Parameter Tuning

pysmac, smac pySMAC, a Python wrapper for the hyperparameter optimization tool SMAC 5 https://github.com/automl/pysmac

5. RESULTS

6. CONCLUSIONS

7. ACKNOWLEDGMENTS

Bogdan, Zechar, Zhuang
???

8. REFERENCES

[1] C. Aranha, Y. C. Lavinas, M. Ladeira, and B. Enescu. Is it possible to generate good earthquake risk models using genetic algorithms? In *Proceedings of the International Conference on Evolutionary Computation Theory and Applications*, pages 49–58, 2014.

[2] A. Helmstetter and D. Sornette. Predictability in the epidemic-type aftershock sequence model of interacting triggered seismicity. *Journal of Geophysical Research: Solid Earth (1978–2012)*, 108(B10), 2003.

[3] Y. Ogata and J. Zhuang. Space–time etas models and an improved extension. *Tectonophysics*, 413(1):13–23, 2006.

[4] F. Omori. On the after-shocks of earthquakes. 1895.

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

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

[7] D. Schorlemmer, J. D. Zechar, M. J. Werner, E. H. Field, D. D. Jackson, T. H. Jordan, and R. W. Group. First results of the regional earthquake likelihood models experiment. *Pure and Applied Geophysics*, 167(8-9):859–876, 2010.

[8] T. Utsu and Y. Ogata. The centenary of the omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth*, 43(1):1–33, 1995.

[9] Y. Yamanaka and K. Shimazaki. Scaling relationship between the number of aftershocks and the size of the main shock. *Journal of Physics of the Earth*, 38(4):305–324, 1990.

[10] J. D. Zechar. Evaluating earthquake predictions and earthquake forecasts: A guide for students and new researchers. *Community Online Resource for Statistical Seismicity Analysis*, pages 1–26, 2010.

[11] J. Zhuang, Y. Ogata, and D. Vere-Jones. Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research: Solid Earth (1978–2012)*, 109(B5), 2004.