Improving the Generation of Earthquake Risk Models Using Evolutionary Algorithms tempered by Domain Knowledge

Yuri Lavinas
University of Brasilia
Computer Science Department
yclavinas@gmail.com

Marcelo Ladeira
University of Brasilia
Computer Science Department
mladeira@unb.br

Claus Aranha
University of Tsukuba
Graduate School of SIE
caranha@cs.tsukuba.ac.jp

***Abstract*—Earthquake Risk Models describe the risk of occurrence of seismic events on a given area based on information such as past earthquakes in nearby regions and the seismic properties of the area under study. These models can be used to help to better understand earthquakes, their patterns and their mechanisms.**

In previous work, we showed that Genetic Algorithms (GA) could generate risk models with the same degree of precision as the Relative Intensity (RI) method, which is considered a benchmark for this problem. However, a few shortcomings were also defined in that approach: (1) The representation of the model in the Genetic Algorithm was too sparse, (2) Domain knowledge was not used to create the model, and (3) The relationship between foreshocks and aftershocks were not taken into account.

In this work, we try to address these three concerns. We propose a new representation of a seismic risk model to be used as the genome of the Genetic Algorithm. We introduces a hybrid model that incorporates seismic theories about earthquake distribution (such as the Omori-Utsu formula). And we introduce two methods to filter the earthquake catalog in order to remove earthquakes that are likely to be aftershocks before generating the risk model.

We examine each of these changes through simulations using the catalog of Japanese earthquakes between 2000 and 2010. Our results show that (XXX)(YYY). These results allow us to draw recommendations for future development in this field.

1. Introduction

Earthquakes can cause great damage to human society through soil rupture, movement, tsunami, etc. Some recent earthquakes that highlight this destructive potential are the great East Japan Earthquake of 2011 (depicted in figure1), and the April 2015 earthquake in Nepal. One important tool for the enactment of policies that minimize the consequences of these events are earthquake occurrence models (also called risk models). These models can be used to identify patterns in the seismic mechanisms that generate earthquakes, and are important to increase our understanding of these events.

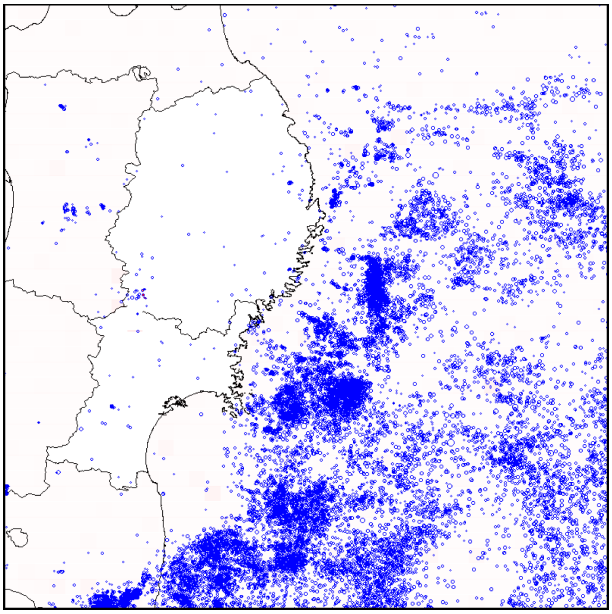

Figure 1. Seismic Activity in Eastern Japan in 2011. Each blue dot represents one earthquake

In our previous work [**?**], we proposed a way to generate earthquake risk models using a standard Genetic Algorithm (here called the GAModel). The GA model was shown to be competitive with the Relative Intensity (RI) model, while not using any a-priori information about the distribution of earthquake occurrences. We summarize the GAModel, and other relevant literature, in Section 2. However, we have identified two key issues with the GAModel. Addressing these issues will be the focus of this paper.

The first issue is that the genome representation used by GAModel has too many parameters (over 2000 for regular cases). Even though a majority of these parameters do not contribute for the accuracy of the final risk model, the size of the search space implies a slower optimization time. To address this issue, we propose a new genome representation for an earthquake risk model, which we call ReducedGAModel. In the ReducedGAModel, only areas with minimal probability of an earthquake are represented as parameters in

the evolutionary process. By reducing the search space, this representation is expected to also increase the convergence speed of the evolutionary optimization process.

The second issue is that GAModel does not take into account any sort of domain knowledge, such as the assumption that earthquakes cluster in both time and space. Heuristic search methods such as Genetic Algorithms usually benefit from the introduction of domain knowledge to the search. Therefore, we propose a hybrid version of the GAModel which incorporates seismic models of earthquake decay. This version, named Emp-GAModel (Empiric GA Model), generates a model with a much smaller number of earthquakes than the regular GAModel. For each earthquake in this model, a sequence of aftershocks is generated using an adaptation of the Epidemic Type Aftershock Sequence model (ETAS). We expect that this hybrid approach will produce more accurate models.

The two proposed adaptations are described in detail in Section 3. To analyze their contributions, we perform a simulated comparison of the GAModel, the ReducedGAModel, the EMP-GAModel, and a combined ReducedEmp-GAModel, which combine both adaptations.

This simulated comparison follows Regional Earthquake Likelihood Model (RELM) framework described by the Collaboratory for the Study of Earthquake Predictability (CSEP) [**?**]. This framework dictates the format of the risk model and the utility function used to evaluate the quality of a risk model (defined as the log-likelihood of the past earthquake occurrence data given the model). The data used in the comparisons is the earthquake occurrence catalog from the Japanese Meteorological Agency (JMA). We focus on earthquakes occurring in the Japanese archipelago between 2000 and 2010. The RELM framework is described in section **??**

Furthermore, we examine the effect of "de-clustering" a catalog in the generation of risk models. In seismological jargon, "de-clustering" refers to the act of identifying groups of main shock-aftershock earthquakes, and removing all but the main shocks from the catalog, which is considered the representative earthquake for the group. Accordingly, a "de-clustered" earthquake catalog is considered to be easier to study, given that the de-clustering process removes redundant information. On the other hand, there are some researchers that do not agree with this idea.

In our experiments, we use three catalogs: one that did not undergo de-clustering (original catalog), and two de-clustered catalogs, generated by two different de-clustering methods: Window and Single Link Cluster. The de-clustering strategies are described in section **??**

Our experimental results, detailed in section **??**, indicate that XXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXXXXXxx XXXXXXXXXXXXXXX. We conclude that YYYYYYY YYYYYYYYY YYYYYYYYYYYYYYYYYY YYYYYYYYYYYYYYYYYy YYYYYYYYYYY.

2. Bibliography and GAModel review

In this section we will briefly discuss some reports of the application of Evolutionary Computation and related method for Earthquake Risk Analysis.

The usage of Evolutionary Computation in the field of earthquake risk models is somewhat sporadic. Zhang and Wang [**?**] used Genetic Algorithms to fine tune an Artificial Neural Network (ANN) and use this system to produce a forecast model. Zhou and Zu [**?**] also proposed a combination of ANN and EC, but their system only forecasts the magnitude parameter of earthquakes. Sadat, in the paper [**?**], follows the idea of Zhou and Zu, aiming to predict the magnitude of the earthquakes in North Iran, but in this case, they used ANN and GA.

Nicknam et al. [**?**] simulated some components from a seismogram station and predicted seismograms for other stations. They combined the Empirical Green's Function (EGF) with GA.

Kennett and Sambridge [**?**] used GA and associated teleseisms procedures to determine the Fault Model parameters of an earthquake.

Some seismological models were developed aiming to estimate parameter values by using Evolutionary Computation. For example, EC was used to estimate the Peak Ground Acceleration (PGA) of seismically active areas [**?**], [**?**], [**?**], [**?**]. The works done by Kerh et al. [**?**], [**?**] are basely a combination of ANN and GA to estimate or predict PGA in Taiwan.

Ramos and Vázques [**?**] used Genetic Algorithms to decide the location of sensing stations. In this work they achieved, in general, better results with the GA method when compared with the Seismic Alert System (SAS) method and a greedy algorithm method.

Saeidian et al. [**?**] also based on the same idea of locating sensing stations. They do a comparison in performance between the GA and Bees Algorithm (BA) to decide which of those techniques would perform better when choosing the location of sensing stations. He found out that the GA was faster than the BA.

Huda and Santosa [**?**] published a paper in which the goal is to find, via GA, the speed of the waves P and S in the mantle and in the earth crust. P waves are indicated as the first fault found in seismological data and S waves are the changes caused in the phase of a P wave [**?**].

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

$$\Lambda(t, x, y, M|\Upsilon_t) = \mu(x, y) \tag{1}$$

where you can denote the number of earthquakes forecast in all bins as $\Lambda(t, x, y)$ [**?**] given that Υ_t is the earthquake observation data up to time t.

The ReducedGAModel, which represents the idea of changing the candidates solution representation (see Section 1), is a method with a similar description of the GAModel. The difference is that, in the ReducedGAModel each possible solution represents only a fraction of the forecast where we expect to find specific risk areas.

2.0.1. GAModel. The GAModel is based on the framework suggested by the CSEP. The CSEP framework, a forecast model uses a gridded rate forecast [**?**], one common format in the literature. The forecast will estimate the risk of earthquake occurrence in this target region, during the target time interval. For this study we considered the target time interval of one year [**?**].

In the CSEP framework one forecast is defined as a region in a specific time interval and is divided in bins. Each bin represents a geographical interval. The whole target area of study is covered by a group of bins where each bin has an earthquake forecast value. This groups of bin represent the $\mu(x, y)$, the background intensity [**?**]. In the GAModel, each possible solution is represented as an entire forecast model.

Genome Representation. In the GAModel each individual represents an entire forecast model. Each gene is a real value, corresponding to one bin in the desired model. The values are sampled from the interval [0, 1). These real values are converted to a integer forecast, we use the same modification of the Poisson deviates extraction algorithm 2.0.1 used in [**?**].In it x is the real value that will be converted and μ is the mean of the earthquakes observations in the real data.

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

[H]

$L \leftarrow \exp(-\mu), k \leftarrow 1, prob \leftarrow 1 * x$
repeat
  increment k
  $prob \leftarrow prob * x$
until $prob > L$
return k
while $prob > L$ **do**
  $k \leftarrow k + 1$
  $prob \leftarrow prob * x$
end while
return k

The individual 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 and they are randomly generated. For more details of the genome representation, please refer to [**?**].

We show an example to clarify how the GAModel works. The Kanto region, one of the four areas used in both studies, is divided into 2025 bins (a grid of 45x45 squares). The GAModel calculates an expected number of earthquakes for every bin on a determined time interval, so the GA searches for good values in 2025 bins.

Fitness Function. The GAModel uses the log-likelihood value, as fitness function. The fittest individual among all the others, is preserved in the next generation, to guarantee that the best solution of the present generation is at least as good as the one from former generation. The bins b_n, define the set β and n is the size of the set β:

$$\beta := b_1, b_2, ..., b_n, n = |\beta|. \tag{2}$$

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:

$$\Lambda^j = \left(\lambda_1^j, \lambda_2^j, \ldots, \lambda_i^j\right); \lambda_i^j := \lambda_i^j(b_i), b_i \in \beta \tag{3}$$

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

$$\Omega = \left(\omega_1, \omega_2, \ldots, \omega_i\right); \omega_i = \omega_i(b_i), b_i \in \beta \tag{4}$$

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

$$L(\omega_i|\lambda_i^j) = -\lambda_i^j + \omega_i \log \lambda_i^j - \log \omega_i! \tag{5}$$

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:

$$\begin{aligned} L^j = L(\Omega|\Lambda^j) &= \sum_{i=1}^{n} L(\omega_i|\lambda_i^j) \\ &= \sum_{i=1}^{n} -\lambda_i^j + \omega_i \log \lambda_i^j - \log \omega_i! \end{aligned} \tag{6}$$

Evolutionary Operators. The GAModel use a combination of operators made available by the Distributed Evolutionary Algorithms in Python (DEAP) [**?**]. We used the One Point Crossover for the crossover operator, the Polynomial Bounded Mutation for the mutation operator and for selection, we used Tournament selection and Elitism. The parameters are described in the Table 1.

Table 1. PARAMETERS USED IN GAMODEL AND EMP-GAMODEL

Population Size	500
Generation Number	100
Elite Size	1
Tournament Size	3
Crossover Chance	0.9
Mutation Chance (individual)	0.1
Polynomial Bounded parameters	eta = 1, low = 0, up = 1

The parameters of the Polynomial Bounded mutation function are:

1) eta = 1. Crowding degree of the mutation. A high eta will produce a mutant resembling its parent, while a small eta will produce a solution much more different;
2) low = 0. The lower bound of the search space;
3) up = 1. The upper bound of the search space.

The chance of applying both mutation and crossover operators takes into account only their chance of occurrence. This means that it may be the case that one of them or both are not applied.

3. The Forecast Models Using Genetic Algorithm

All forecast models proposed in this paper are based in the CSEP framework.

Each individual has its own representation of the framework based on different perceptions of what are the best aspects of the framework.

Fo all methods, the population is trained on earthquake event data for a training period, which is anterior to the target test period. After completing the evaluation limit, the best individual is chosen to be the final forecast.

3.1. 1-year Forecast Models

Large and independent earthquakes, also known as mainshocks, are followed by a wave of others earthquakes, the aftershocks [**?**]. Hence there is no physical measurement to identify mainshocks and its aftershocks [**?**], we divided the forecast models in two groups: the ones that only forecasts mainshocks, using only GA techniques, and those that forecast both mainshocks and aftershocks using both GA techniques and empirical laws, such as the modified Omori law. These laws are use to derive the aftershocks from a synthetic data of mainshocks.

Both classes forecast earthquakes with magnitude greater than 3.0 for every scenario proposed.

3.2. Mainshock Models

There are two mainshock models. The ones generated by the GAModel and the ones generated by the ReducedGAModel.

3.2.1. The ReducedGAModel. 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 because in some bins there are no earthquake occurrences in the observation data. That means that the GAModel has a lot of parameters and may of its bins have null values (values equal to 0). To avoid such unnecessary task we proposed the ReducedGAModel.

With this method, we aim to minimise the search space and the quantity of parameters the GA has to deal with. For that we changed the individual representation. The individuals in the ReducedGAModel only define expected number of earthquakes in bins that already had some occurrence in the past, giving a direction to where the GA should search. That helps the ReducedGAModel in the search for better solutions and it makes the convergence faster once the space search is smaller.

The ReducedGAModel has a similar description of the GAModel.The difference is that, in the ReducedGAModel, each possible solution represents only a fraction of the forecast where we expect to find specific risk areas. To do so, this method obtain the position of past occurrences. Then it calculates some expected number of earthquakes only for the bins related to those positions. These positions may vary during the evolution of the method, including positions that never had earthquake events before. That is important to add some variation to the method.

Genome Representation. The genome representation in the ReducedGAModel is a simplified version of the genome of the GAModel. For the ReducedGAModel, the genome is a list of ordered pairs. The first element of the pair are the coordinates of a bin in the model. The second element of the pair is a number that indicates an earthquake occurrence estimate for this bin.

To calculate the number of elements of the individual we use the real data from the prior 5 years and create a list of every bin that had events in it.

In the ReducedGAModel, each individual is a list of a sub-region of the forecast model. This list initially refers to bins where earthquake events happened in the past. During a execution of the ReducedGAModel, the list may refer to positions that never had occurrences before. Each element of the list, a gene, also contains one real value between [0,1). In the initial population, these values are sampled from a uniform distribution and they are randomly generated. When needed, every real value is converted to a integer forecast by the same Algorithm, as in the GAModel.

To generate the forecast model we need to do an intermediate step. We map every location from the list with a bin in the forecast model.

The genome size is usually smaller than the one used in the GAModel and the Emp-GAModel, once the amount of sub-regions where earthquakes with magnitude above 3.0 happened for any given area is smaller then the total number of genes of the individual.

To illustrate, we give the following example. Lets consider that there are 10 bins with occurrences in Kanto in the last 5 years. It will make the GA start searching for good values for only those 10 bins, leaving the other 2015 bins empty, representing zero occurrence. It is important to highlight that in the worst case, it will make the same amount of searches as the GAModel. The final forecast model will maintain the amount of bins with occurrence, but the number of events for every bin and their location may change.

Fitness Function. The fitness function is the same as in the GAModel, 2.0.1. Here is also important to generate the forecast model by applying the map function on the individual.

Evolutionary Operators. All operators in the ReducedGAModel are the same as the operators of the GAModel, except the mutation function. We use a simple mutation operator which samples entirely two new values, both sampled from uniform distributions. The first, is a new real value from [0,1) and the second one, a new integer

value from [0,x), where x is the maximum amount of bins a model can have in the target region. For the parameters see Table 1.

As in the GAModel, the chance of applying both mutation operator function and crossover operator are independent and they may or may not be used.

3.3. Mainshock+Aftershock Models

The mainshock and aftershock methods are a two-step methods. The first step is as defined for the mainshocks methods, therefore, we first use GA techniques to obtain a synthetic mainshock data. The second step is to use seismological empirical equations to obtain the aftershocks from the mainshocks.

Hence earthquakes cluster in space and inspired by the space-time epidemic-type aftershock sequence (ETAS), we proposed two methods, called Emp-GAModel and Emp-ReducedGAModel. They represent the idea of associating the GA with seismological empirical equations. They are described as:

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

That can be expanded to:

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

methods use $\mu(x, y)$ as defined for mainshock methods 3.2. It is calculated as an expected number of earthquakes for every bin in the target region, given that Υ_t is the earthquake observation data up to time t.

3.3.1. Empirical Equations. The Omori law, $g(t)$, which is considered one empirical formula of great success [**?**] [**?**] [**?**], is a power law that relates the earthquake occurrence and its magnitude with the decay of aftershocks activity with time. For this approach we used the probability density function (PDF) form of the modified Omori law [**?**]:

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

The variable p is a index of this equation and the variable c is a constant, given in days. In the paper [**?**], Utsu summarise most of the studies in Japan and described the range for these variables. For p the range is between 0.9 and 1.4 and for c 0.003 and 0.3 days. These values were based on the Davidon-Fletcher-Powell optimisation procedure and used in ETAS [**?**]. Also there is the variable t that is the time limit to when a mainshock may influence the cause a aftershock.

Based on paper [**?**], we set the values of 1.3 for p and 0.003 for c for our experiments. We set the time interval t between a mainshock and its aftershocks at one month. In the paper, it says that if the t value is too short, the number of aftershocks is too small, but if it is too big, we may also consider background activity and suggest the use of a 30 days period.

For $K(M_i)$, the total amount of triggered events, we count aftershocks within a given area, A, using the following formula, where M_c is the magnitude threshold, with $M_c = 3.0$:

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

In the paper [**?**], it states that α should be equal to the inverse of the magnitude of an event, or $magnitude^{-1}$. To obtain A, the following equation from [**?**], was used:

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

With the $K(Mi)$ and $g(t)$, the PDF Omori, equations it is possible to calculate the total number of earthquakes. For that we must sum the product of the equations, varying t:

$$\sum_{t_i \in t} K(M_i)g(t - t_i) \tag{12}$$

This result will lead to a number of aftershocks related to a single mainshock Then, we can use the $P(x, y)$ equation to distribute the aftershock to the bins near the mainshocks position. $P(x, y)$ calculates the position of the aftershocks with base on the origin of the mainshock. It is a simple space distributing function, that allocates the aftershocks in one of the following positions: upper, lower, left or right. It runs for a number of steps, getting further from the origin at each step or as when there are no more events to be allocated.$P(x, y)$ can be split into 4 equations, one for each position:

$$\begin{aligned}
model[x + y] &= (aftershocks - [model[x] - 2 * x])/4; \\
model[x - y] &= (aftershocks - [model[x] - 2 * x])/4; \\
model[x - y * row] &= (aftershocks - [model[x] - 2 * x])/4; \\
model[x + y * row] &= (aftershocks - [model[x] - 2 * x])/4
\end{aligned}$$

and lastly, the $J(M)$ is obtained by using the function etasim, from the SAPP *R* package [**?**] that simulates magnitude by Gutenberg-Richter's Law.

3.3.2. Emp-GAModel. The Emp-GAModel is a specialisation of the GAModel. This is achieved by the use of empirical equations after the forecast is provided. This means that the its genome representation are the same as in the GAModel.

Genome Representation. The genome representation is the same as in the GAModel, Section 2.0.1.

Fitness Function. The fitness function is the same as in the GAModel,Section 2.0.1, and the ReducedGAModel.

Evolutionary Operators. The Emp-GAModel use the same combination of operators that the GAModel. For more explanation, please see the Section 2.0.1.

Emp-ReducedGAModel. The Emp-ReducedGAModel is a specialisation of the ReducedGAModel. This is achieved by the use of empirical equations after the forecast is provided. This means that the its genome representation are the same of ReducedGAModel.

Genome Representation. The genome representation is the same as in the ReducedGAModel,Section 3.2.1.

Fitness Function. The fitness function is the same as for all methods,Section 2.0.1. Here is also important to generate the forecast model by applying the map function on the individual as in the last Section, 3.2.1.

Evolutionary Operators. The Emp-ReducedGAModel use the same combination of operators that the ReducedGAModel. For more explanation, please see 2.0.1.

4. Experimental Data

Here we describe the earthquake catalogue, how we used it and the regions in Japan selected for the experiments.

We also preprocessed the catalogue. We wanted to analyse how earthquakes characteristic changed with the magnitude and the depth. Also we explain briefly how we classified the mainshocks and aftershocks.

4.0.3. Earthquake data. The goal of this research is to find existing patterns in the occurrence of earthquakes. For that it is essential to access trustful data and to explore its details. From the *Japan Meteorological Agency* web page we obtained earthquake data about earthquakes in Japan. In this data there are information about earthquakes that happened in or nearby Japan, with the variables: time of the occurrence, magnitude, latitude and longitude and epicentre depth, for the years of 2000 to 2013.

During the preprocessing phase, we discovered a higher number of occurrences of earthquakes during the year of 2011, when a 9.0 M_w earthquake happened, see Section **??**. This earthquake triggered too many after called aftershocks in all Japan. It is considered that big earthquakes may cause others earthquakes [**?**]. In Figure 2 it is possible to visualise a great number of earthquakes for the year of 2011. Because of this abnormal behaviour and because we decided to focus on more stable occurrences, we limited the training base to earthquakes until 2010.

Based on the statement done before and considering that we want earthquakes that follow more stable patterns, we selected the ones that happened in land areas or very shallow sea areas, with maximum depth of 100km.

4.0.4. Regions. For the experiments, the data was changed into slices for every year. Each slice is as follows: if the base contains data about a time interval of 10 years, it will be split in 10 slices.

We also selected some sub-areas in Japan to better extract and understand earthquakes characteristics and patterns. Those areas are Kanto, Kansai, Touhoku and East Japan. The Figure 3 shows how we defined them.

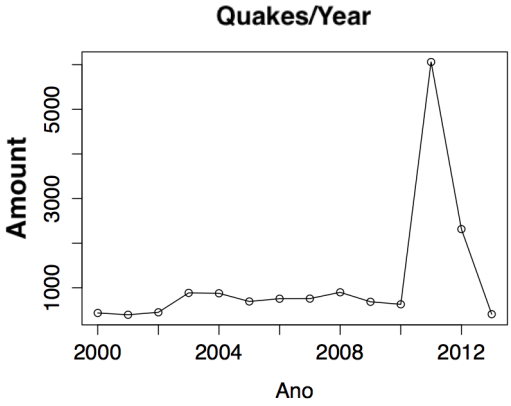

Figure 2. Amount of earthquake by year.

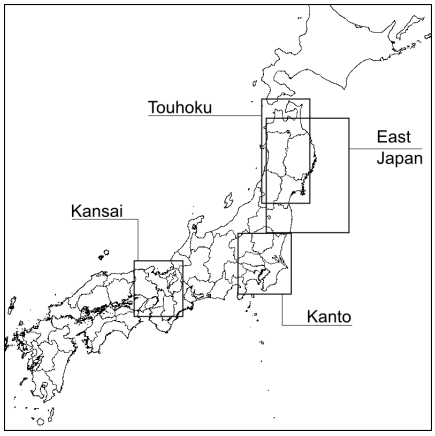

Figure 3. Japan and the areas used in this studied.

They are described as follows:

Kanto. Kanto is the region around Tokyo. It is an area with high seismologic activity during the years we studied. Its coordinates are 34.8 North, 138.8 West, with 2025 bins. Each bin covers an area of approximately 25km^2.

Kansai. Kansai is the region that includes Kyoto, Osaka and many others historical cities. In this area, rather than Kanto area, there is a small seismic activity. Its coordinates are 34 North, 134.5 West, with 1600 bins. Each bin covers an area of approximately 25km^2.

Touhoku. Touhoku is the region in the North of the main Japanese island. It has some clusters of seismic activities during the years we studied. Its coordinates are 37.8 North, 139.8 West, with 800 bins. Each bin covers an area of approximately 100km^2.

East Japan. Is the region that is related with the east coast of Japan. It is the most different area, because it has earthquakes that happened both in land or in the sea. It was in this region that the 9.0 M_w earthquake happened. Its coordinates are 37 North, 140 West, with 1600 bins. Each bin covers an area of approximately 100km^2.

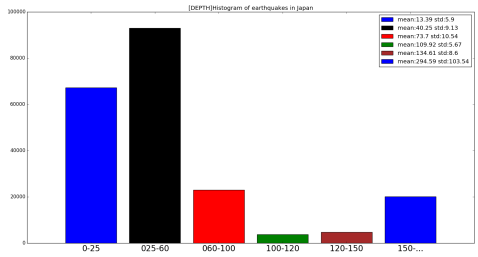

Figure 4. Depth Histogram of earthquakes.

4.0.5. Depth Histogram of Earthquakes. The patterns of earthquakes are dependent of the epicentre. We wanted to explore the relation between the depth of the earthquakes and how would our models behave on those situations.

In Figure 4, it is possible to understand that most of the earthquakes happened with depths smaller or equal to 100 km. The earthquakes deeper than 100 km are fewer and more distant, as it is in the same Figure.

The reason we decided to groups as: earthquakes with depth until 25 km, until 60 km or until 100 km. This is because shallow earthquakes are considered to be more independent earthquakes [**?**].

4.0.6. Mainshocks and Aftershocks - Clustering. In the Section **??**, we explained that we have two kinds of models, the ones that only consider aftershocks and those that consider both mainshocks and aftershocks. Therefore, it is needed to isolate, to classify the earthquakes into one of these two groups.

The question is how should it be done. The simplest way, is to select earthquakes with magnitude above 3.0 in the Richter Scale and then to consider those as the mainshocks. The distribution of earthquakes after this selection is exemplified in the Figure 5.

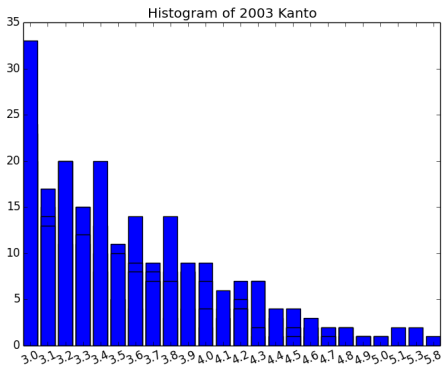

Figure 5. Histogram of earthquakes stronger than 3.0 in the Richter Scale in Kanto

The problem with this simple idea is: if a big mainshock happens and it triggers some aftershocks with magnitude higher than 3.0 in the Richter Scale it would be considered as a mainshock. To avoid this problem we used two methods proposed in the literature: Window Methods and the Single-Link Cluster. For more information about these methods, see reference [**?**].

The models that used the Window declustering catalogue have in the word Window appended at their names or SLC, for the methods that used the Single Link Cluster.

5. Experimental Design

The first experiment was made to compare the all the models proposed with each other and to discover which method would achieve higher log-likelihood values. We created some scenarios (space/time regions), and we applied the methods for for the regions of Kanto, Kansai, Touhoku and East Japan for a given year (2005-2010) with earthquakes with depth lesser than 100km. We also used 3 kinds of catalogues with the minimum magnitude of 3.0: the JMA and the declustered catalogues form the Window method and the SLC method.

We compared the means of the models log-likelihood values using the ANOVA test. If a group of variables considered for the ANOVA test showed no statistically significant difference, we applied the Paired Student t-test, in the case all groups showed statistically significant difference, the Tukey HSD methodology analysis was used.

The second experiment was made a to compare how the magnitude of the earthquakes influence the models generated. We used the same scenarios from the first experiment. We split the models obtained from these scenarios into slices composed of earthquakes that have magnitude in a given magnitude interval. We calculated the log-likelihood of these slices-models and applied the ANOVA test and the Tukey HSD to compared them.

5.1. Details of the Statistical Analysis

The goal is to discover if there is any variation between the methods and which are the most influential variables. To achieve that, we used the ANOVA test, because it indicates that the means of several groups are equal or not for a given confidence interval. The confidence interval was set to 95%, meaning that if the "p-value" is smaller than 0.05 it signifies that there exists a statistical significant evidence that the variables variance are different from each other.

The hypothesis for this experiment can be generalised as follows:

$$\begin{cases} H_0 : \text{The population means are equal.} \\ H_1 : \text{The population means are different.} \end{cases}$$

The Tukey HSD is applied on the results obtained from the ANOVA test to specify which groups differ, in the case any group has a "p-value" lesser than 0.05. Tukey's methodology analysis shows the means of a case with the

means of every other case. Doing so, it identifies differences between means :

$$\begin{cases} \mu_a - \mu_b, \text{where } \mu_a \text{ is the mean of the first group} \\ \mu_b \text{ is the mean of the second group.} \end{cases}$$

In the case where statistical significant difference exists, we explore this by pairing the measures observations of two groups. That is:

$$\begin{cases} H_0 : \mu = 0, \text{the difference between observations is 0.} \\ H_1 : \mu! = 0, \text{difference between observations is not 0.} \end{cases}$$

5.2. Results from The Mainshock Models Mainshock with Aftershock Models Experiment

An one-way between subjects ANOVA was conducted to compare the effects of the models, the years and regions on the log-likelihood value. The models compared are: ReducedGAModel, Emp-ReducedGAModelSLC, Emp-GAModel, Emp-ReducedGAModel, GAModel-Window, ReducedGAModelWindow, GAModelSLC, ReducedGAModelSLC, Emp-GAModelWindow, Emp-ReducedGAModelWindow, GAModel and Emp-GAModelSLC.

Based on the results of the ANOVA test, it is evident that all variables are significantly different. The results of the experiments are in the Table 2. All variables had a "p-value" lower than 0.05, indicating that they can be considered different.

Because we found statistically significant result, we applied a Post-hoc comparisons using the Tukey HSD analysis methodology on the ANOVA result. It compared each condition with all others. For example, it compares the values from the GAModel with the GAModelWindow.

It indicated that the models GAModelSLC, Emp-ReducedGAModelWindow, GAModelWindow, ReducedGAModelWindow, ReducedGAModelSLC achieve statistically better or equal results in terms of log-likelihood when compared with the other models. When they are compared with themselves, they are statistically equal.

We applied the ANOVA test now only with this 5 models. The models group have a "p-value" of 0.171, indicating that they have similar log-likelihood values. The results are in the Table 3.

Therefore, to confirm that the models are statistically equal, we conducted the Tukey HSD on this ANOVA result. This time, we found statistically significant difference only for the year and region groups. To show that the models results are not statistically different from each other, we applied a pairing analysis.

From the the pairing analysis, we decided to use the *ReducedGAModelSLC* as the representative method of this study. That is because, in all cases when its values were compared, it showed a better performance in the means of the log-likelihood values and in only one case the "p-value" was higher than 0.05. For the results, see the Table 4.

5.2.1. The Models Examples And The Real Data. The Figure **??** shows a model from the ReducedGAModelSLC method for the year 2005 in East Japan. The next Figure, **??** shows a model from the ReducedGAModel **??** method for the year 2005 in East Japan.

All Figures, **?? ?? ?? ??**, indicate a low earthquake intensity as white while the more intensity areas, are shown in red. They are, in order, the data visualisation for the model from: the GAModel, the ReducedGAModel, the Emp-GAModel and the Emp-ReducedGAModel for East Japan in 2005. The Figure **??** represents the earthquake occurrences in the same region and year.

5.3. Magnitude Experiment

The goal is to discover if there is any variation in the methods when considering the magnitude of the earthquakes in the models. We wanted to explore the relation between the magnitude of the earthquakes and how would the models behave on those situations. To achieve that, we used the ANOVA test The confidence interval was set to 95%.

5.3.1. Magnitude Study. We compared those split-models against themselves. Based on the results of this test, it is evident that all variables are still significantly different. The results of the ANOVA test are in the Table 5. For all, as before, we choose the confidence level to be 95%.

We found statistically significant result and, as before, we applied the Tukey HSD test. It indicated that the interval $[3.0 - 4.0]$ always performed, in terms of log-likelihood values, worse than all other intervals. this phenomenon also happens in the interval $[4.0 - 5.0]$, though in this case, the difference is not as big as the last one. The other intervals show no significant difference.

From the results found, we decided to chose only earthquakes with magnitude higher than 4.0 as our threshold value.

6. Conclusions

Variable	Degrees of Freedom	Sum Sq	Mean Sq	F Value	Pr(>F)
Model	15	149303768	9953585	63.72	<2e-16
Year	5	414016420	82803284	530.06	<2e-16
Region	3	869821655	289940552	1856.02	<2e-16

Table 2. ANOVA TEST RESULTS VALUES - MAINSHOCK MODELS MAINSHOCK AND AFTERSHOCK MODELS.

Variable	Degrees of Freedom	Sum Sq	Mean Sq	F Value	Pr(>F)
Model	4	884882	221220	1.604	0.171
Year	5	150297410	30059482	217.955	<2e-16
Region	3	234225270	78075090	566.107	<2e-16

Table 3. ANOVA TEST RESULTS VALUES - EMP-REDUCEDGAMODELWINDOW, GAMODELWINDOW, REDUCEDGAMODELWINDOW, GAMODELSLC, REDUCEDGAMODELSLC.

Region	Models Compared	Mean of $\mu_a - \mu_b$	p-value
Kansai	**EMP-GAModelWindow** - GAModelWindow	38.67553	3.304e-05
	EMP-GAModelWindow - ReducedGAModelWindow	4.272185	0.2607
	EMP-GAModelWindow - GAModelSLC	112.0424	1.122e-05
	EMP-GAModelWindow - **ReducedGAModelSLC**	-1.787262	0.5673
	GAModelWindow - **ReducedGAModelWindow**	-34.40335	0.000963
	GAModelWindow - GAModelSLC	73.36687	9.065e-06
	GAModelWindow - **ReducedGAModelSLC**	-40.46279	6.32e-05
	ReducedGAModelWindow - GAModelSLC	107.7702	2.632e-05
	ReducedGAModelWindow - **ReducedGAModelSLC**	-6.059447	0.2982
	GAModelSLC - **ReducedGAModelSLC**	-113.8297	1.2e-05
Touhoku	EMP-**GAModelWindow** - GAModelWindow	3.34556	0.546
	EMP-GAModelWindow - ReducedGAModelWindow	81.60965	5.225e-07
	EMP-GAModelWindow - GAModelSLC	63.02216	0.01971
	EMP-GAModelWindow - **ReducedGAModelSLC**	-62.70586	0.007075
	GAModelWindow - ReducedGAModelWindow	78.26409	2.938e-05
	GAModelWindow - GAModelSLC	59.6766	0.04829
	GAModelWindow - **ReducedGAModelSLC**	-66.05142	0.001231
	ReducedGAModelWindow - **GAModelSLC**	-18.58749	0.3443
	ReducedGAModelWindow - **ReducedGAModelSLC**	-144.3155	0.000214
	GAModelSLC - **ReducedGAModelSLC**	-125.728	0.01216
East Japan	**EMP-GAModelWindow** - GAModelWindow	1.872764	0.9539
	EMP-GAModelWindow - ReducedGAModelWindow	194.4944	1.834e-06
	EMP-GAModelWindow - GAModelSLC	189.1155	0.0003456
	EMP-GAModelWindow - **ReducedGAModelSLC**	-274.9858	4.961e-05
	GAModelWindow - ReducedGAModelWindow	192.6217	0.003738
	GAModelWindow - GAModelSLC	187.2428	9.495e-06
	GAModelWindow - **ReducedGAModelSLC**	-276.8586	4.636e-05
	ReducedGAModelWindow - **GAModelSLC**	-5.378912	0.8576
	ReducedGAModelWindow - **ReducedGAModelSLC**	-469.4803	1.446e-05
	GAModelSLC - **ReducedGAModelSLC**	-464.1014	2.38e-06
Kanto	EMP-**GAModelWindow** - GAModelWindow	57.95612	0.00138
	EMP-GAModelWindow - ReducedGAModelWindow	79.60781	3.441e-05
	EMP-GAModelWindow - GAModelSLC	274.3114	5.717e-06
	EMP-GAModelWindow - **ReducedGAModelSLC**	-96.61803	6.22e-07
	GAModelWindow - ReducedGAModelWindow	21.65169	0.1105
	GAModelWindow - GAModelSLC	216.3553	2.302e-07
	GAModelWindow - **ReducedGAModelSLC**	-154.5741	1.741e-05
	ReducedGAModelWindow - GAModelSLC	194.7036	3.678e-05
	ReducedGAModelWindow -**ReducedGAModelSLC**	-176.2258	4.337e-06
	GAModelSLC - **ReducedGAModelSLC**	-370.9294	1.942e-06

Table 4. PAIRED EXPERIMENT RESULT. THE BOLD MODELS ARE THE ONES WITH BETTER RESULTS. THE REDUCEDGAMODELSLC ALWAYS ACHIEVED BETTER RESULTS, THEREFORE IT IS ALSO UNDERLINED.

Variable	Degrees of Freedom	Sum Sq	Mean Sq	F Value	Pr(>F)
Model	5	2.360e+09	4.720e+08	2828	<2e-16
Year	3	4.624e+09	1.541e+09	9234	<2e-16
Magnitude	7	3.726e+09	5.322e+08	3189	<2e-16

Table 5. ANOVA TEST RESULTS VALUES - MAGNITUDE STUDY.