Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Earthquake Risk Induction Models with Genetic Algorithm

Yuri Cossich Lavinas

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Marcelo Ladeira

Coorientador
Prof. Dr. Claus de Castro Aranha

Brasília
2016

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Earthquake Risk Induction Models with Genetic Algorithm

Yuri Cossich Lavinas

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Marcelo Ladeira (Orientador)
CIC/UnB

Prof. Dr. Claus de Castro Aranha
University of Tsukuba

Prof. Dr. George Sand Leão Araújo de França
Observatório Sismológico/UnB

Prof. Dr. Rodrigo Bonifácio de Almeida
Coordenador do Bacharelado em Ciência da Computação

Brasília, 05 de Julho de 2016

Dedicatória

...

Agradecimentos

Agradeço à todas pessoas que de algum forma contribuíram para a realização deste trabalho.

Agradeço aos Professores Marcelo Ladeira e Claus Aranha por terem orientado esse trabalho. Por aceitarem este desafio, por sempre paticiparem das reuniões, mesmo nos mais diversos horários. Agradeço também por sempre se preocuparem comigo e com meu bem-estar. Por sempre apoiarem meus estudos e me desafiarem. Obrigado.

Agradeço ao Professor George Sand por participar da banca e se disponibilizar para conversas sempre que eu precisar.

Agradeço a minha mãe, Rosane, por estar sempre presente, me apoiar e nunca duvidar de mim. Agradeço por me ouvir especialmente quando era difícil de me entender.

Agradeço ao meu pai, Omir.

Agradeço ao meu irmão Eric por me ser meu amigo, estar sempre comigo e me entender sempre.

Agradeço aos especialmente aos meus amigos Pimenta, Pedrinho, Shima, Artur, Cata, Black, Maurício, Giordano, Haroldo, Paulo, Folle, Akemi, Mari, Cainã, Marcelo Rios, Thales, Paula, Luisa, Magami, Natsumi e todo mundo da Cjr e do Sigma.

Resumo

Este projeto visa desenvolver um modelo de previsão de riscos de terremotos com Algoritmos Geneticos (GA). Modelos de risco de terremotos descrevem o risco de ocorrência de atividades sísmicas em uma determinada área baseado em informações previamente obtidas de terremotos em regiões próximas da área de estudo. GA foi utilizada para aprender um modelo de risco usando informações previamente obtidas como base de treino. Baseado nos resultados obtidos, acreditamos ser possível obter melhores modelos se conhecimentos do domínio da aplicação, como conhecimentos oriundos da literatura ou modelos de distribuição de terremotos, poderem ser incorporados ao processo de aprendizado do Algoritmo Genético.

O objetivo principal é definir um método para estimar a probabilidade de ocorrências de terremotos no Japão usando dados históricos de terremotos para um grupo de determinadas regiões geográficas. Este trabalho se baseia no contexto do "Collaboratory for the Study of Earthquake Predictability" (CSEP), que visa padronizar os estudos e testes de modelos de previsão de riscos de terremotos.

Durante o desenvolvimento das atividades, passamos por três estágios. (1) Nós propusemos um método baseado em uma aplicação de GA e objetivamos gerar um método estatístico de análise de risco de terremotos. Estes foram analisados por seus valores de *log-likelihood*, como sugerido pelo *Regional Earthquake Likelihood Model* (RELM). (2) A seguir, modificiamos a representação do genoma, de uma representação baseada em área para uma representação baseada em ocorrências de terremotos, buscando obter uma convergência mais rápida dos valores de *log-likelihood* dos candidatos do GA e (3) usamos métodos da sismologia conhecidos para refinar os candidatos gerados pelo GA.

Em todas as estapas, os modelos de risco são comparados com dados reais, com modelos gerados pela aplicação do *Relative Intensity Algorithm* (RI) e com eles próprios. Os dados utilizados foram obtidos pela *Japan Metereological Angency* (JMA) e são relativos a atividades de terremotos no Japão entre os anos de 2000 e 2013.

Nós analisamos as contruibuições de cada modelo proposto usando metodologia descritas pelo CSEP e comparamos os modelos desenvolvidos. Os resultados apontam que modelos com terremotos mais estáveis possuem maiores valores de log-likelihood.

Palavras-chave: algoritmos genéticos, terremotos, log-likelihood

Abstract

This projects aims to develop an earthquake prevision risk model using Genetic Algorithms (GA). 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. We used GA to learn risk models using past earthquake occurrence as training data. Based on the results obtained, we believe that a much better model could be learned if domain knowledge, such as known theories and models on earthquake distribution, were incorporated into the Genetic Algorithm's training process.

The main goal is to define good methods to estimate the probability of earthquake occurrences in Japan using historical data of a group given geographical regions. This work is established in the context of the "Collaboratory for the Study of Earthquake Predictability" (CSEP), which seeks to standardise the studies and tests of earthquake risk prevision models.

To achieve the main goal, we passed three stages. (1) We proposed a method based in one application of GA and aims to develop statistical methods of analysis of earthquake risk. The risk models generated by this application were analysed by their log-likelihood values, as suggested by the Regional Earthquake Likelihood Model (RELM). (2) Then, we modify the genome representation from an area-based representation to an earthquake representation aiming to reach a faster convergence of the log-likelihood values of the GA's candidates and (3) we use known methods from seismology (such as the Omori-Utsu formula) to refine the candidates generated by the GA.

In all stages, the risk models are compared with real data, with the models generated by the application of the Relative Intensity Algorithm (RI) and with themselves. The data used was obtained from the Japan Meteorological Agency (JMA) and are related with earthquake activity in Japan between the years of 2000 and 2013.

We analyse the contributions from each risk model using the methodologies described in the CSEP and compare their quality. Our results indicate that models with more stable earthquakes obtain higher log-likelihood values.

Keywords: genetic algorithms, earthquake, log-likelihood

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nesse capítulo, apresentamos uma especificação geral do problema abordado, sua relevância e quais são os objetivos do estudo.

1.1 Terremotos

Terremotos causam muito danos ao meio ambiente e suas consequências podem representar, direta ou indiretamente, riscos aos seres humanos. Eles se manifestam por tremores e movimentações terrestres. Terremotos podem causar outros abalos sísmicos, deslizamentos de terras, atividades vulcânicas, etc.

Existem muitos exemplos que mostram a força devastadora dos terremotos. Em Abril de 2015, aconteceu um tremor no Nepal, com magnitude de momento (M_w) de 7.8 e considerado o maior terremoto desde 1934. Ele destruiu muitos prédios e infraestrutura, desencadeou muitos deslizamentos de terra e pedras em regiões montanhosas [32]. Muito outros terremotos dependentes deste [31], aconteceram após ele, incluindo dois grandes *aftershocks* com magnitude (M) de 6.7 e 7.3 que causaram ainda mais danos [32].

Outro exemplo aconteceu em Março de 2011, no Japão. Ele foi um terremoto de magnitude de momento de M_w 9.0 [29], e foi considerado como o terremoto me maior magnitude que já aconteceu no Japão. Ele casou *tsunamis* que chegaram mais de 39 metros, moveu a ilha principal do Japão em mais de 2 metros como também alterou o eixo da Terra. Este terremoto casou mais de 14 mil mortes, fez mais de 244.000 pessoas desabrigadas e ainda provocou o derretimento do complexo *Fukushima Daiichi Nuclear Power Plant*. Muitos outros terremotos seguiram este evento [17]. Também em 2011, um terremoto de magnitude de momento M_w 7.1 aconteceu em Van, Turquia e causou muitas

mortes e muitos danos. Este são alguns exemplos recentes do quão perigosos podem ser os terremotos. [8].

Este terremotos, e muitos outros, podem causar danos a sociedade e tem características em comum. Eles não somente são terremotos muito fortes, como também aconteceram em áreas habitadas, causando ainda mais danos. Para diminuir o quanto for possível futuros desastres, muito pode ser feito. Isto inclui desenvolver estruturas com técnicas mais resistentes a terremotos, criar sistemas de alarme de terremotos, criar sistemas de engenharia mais preciso, etc.

Para prevenir quantas casualidades quanto possível, é necessário entender os padrões e mecanismos por trás das ocorrências do terremotos. Precisamos aprender se existe alguma relação entre a localização dos eventos e o tempo destes, como essa relação ocorre, et cetera. Assim, é possível criar melhores modelos de previsão de ocorrência de terremotos, indicando quais regiões demonstram uma probabilidade maior de ocorrências de terremotos em um certo período de tempo.

Até agora, é difícil de entender claramente como as diferentes variáveis sísmicas (tempo da ocorrência, magnitude, local, profundidade, etc.) influenciam os abalos sísmicos e se existem um modelos matemáticos capazes de fornecer modelos detalhados e informações precisas sobre as relações e como estimá-las. Portanto, desenvolver um modelo de previsão de risco de terremotos pode se comprovar muito complexo.

1.2 Predição de Terremotos

Koza et al. [14], disseram que Computação Evolutiva (EC) pode estimar, por tentativa e erro e baseado em uma grande quantidade de dados, melhores soluções para problemas que seres humanos podem ter dificuldade em resolver. EC é a família do sub-campo da inteligência artificial que visa extrair padrões e resolver problemas usando uma grande quantidade de dados históricos. Podemos também dizer que sem qualquer conhecimento sobre o problema a ser controlado, a EC pode aprender e estimar soluções para o problema [16].

Algoritmos Genéticos (GA) é a técnica de EC que é utilizada neste estudo. Ela é constituída pela categoria de busca heurística, são algoritmos estocásticos e o método de busca deles é baseado em herança genética e sobrevivência do mais adaptado [15]. São

técnicas aplicáveis em casos onde é difícil entendimento e que o conhecimento sobre o caso ainda está suficientemente disponível.

Em vista dessas informações e nas dificuldades de entender como terremotos se comportam, procuramos explorar dados históricos de terremoto usando GA. É esperado encontrar novas ideias sobre o estudo de terremotos, seus padrões e mecanismos por trás de suas ocorrências. Para isso, precisamos primeiro determinar o problema de previsão, depois verificar se é adequado utilizar EC para o problema de gerar um modelo de previsão de sismos.

Previsão de terremotos é uma área de estudo em aberto. Nenhuma pesquisa ainda foi capaz de sugerir que um terremoto de larga escala pode ser previsto com acurácia [1] e muito cientistas acreditam que previsão de terremotos pode ser tanto impossível como as informações para fazer a predição estejam foram do alcance da ciência [5].

No contexto desse estudo, não buscamos prever nenhum terremoto individualmente ou suas características. O objetivo dessa pesquisa está focado na observação de terremotos se agruparem no espaço tempo. As técnicas de computação para aprender e gerar modelos de risco de ocorrência de terremotos. Existe diversas utilidades por detrás do estudo dos mecanismos dos terremotos, com o objetivo de gerar modelos estatísticos de previsão de riscos [25].

Posteriormente, buscaremos estudar modos de melhorar os métodos gerados, usando tanto GA como outras técnicas computacionais a também qualquer conhecimento sismológico. Para isso, propomos diferentes representações que buscam refinar a qualidade do algoritmo e para incorporar métodos da sismologia para os métodos já presentes.

Chapter 2

Introduction

In this chapter we present a general specification of the problem, its relevance and what are the goals of this study.

2.1 Earthquakes

Earthquakes may cause lots of damages environment and consequently may represent, directly or inderectly, a risk to human lives. They manifest themselves by shaking and moving of the ground. They may also cause tsunamis, landslides, volcano activities, etc.

There are many examples that show how devastating one large earthquake can be. In 25th of April 2015, there was a strong earthquake in Nepal, with moment magnitude (M_w) of 7.8 and considered the largest since 1934. It destroyed lots of buildings and infrastructure, and triggered numerous landslides and rock/boulder falls in the mountain areas [32]. Many other aftershock occurrences, which are dependent earthquakes [31], happened after it, including two major aftershocks (M) 6.7 and 7.3 earthquakes that caused additional that were also very destructive [32].

Another example happened in March 2011, Japan. It was a 9.0 M_w earthquake [29], and it is considered the most powerful earthquake to ever hit Japan. It caused tsunami that reached more than 39 meters, moved the main island in Japan more than 2 meters east and also changed the Earth axis. It was reported that it caused more than 14 thousand deaths, made more than 244,000 people homeless and provoked a meltdown of the Fukushima Daiichi Nuclear Power Plant complex. Many large aftershocks followed the main event [17]. Also in 2011 a magnitude M_w 7.1 earthquake hit Van, Turkey and caused lots of deaths and great damages. These are only three very recent examples of

large earthquake damages of how dangerous earthquakes can be [8].

Those earthquakes, and many others that hazard the human society, have some common characteristics. They are not only powerful earthquakes but they happened nearby populated areas, which increase the damaged provoked. To minimize as much as possible future earthquake disaster, a lot can be done. That includes developing good urban planing, for example to build structures with techniques that can withstand the forces of earthquakes, to create earthquake warning systems, to create more precise civil enginnering codes, and such.

To be able to prevent as many casualty as possible, we need to understand the patterns and mechanisms behind the occurrence of earthquakes. We need to know if there is any relationship between the earhquake locations and its time of occurrance, how they are related to each other, et cetera. Based on this, it is possible to create better seismic risk forecast models, indicating which regions show a higher probability of earthquake occurrence at certain periods in time.

Until by now, it has been difficult to clearly understand the many different seismic variables (time of occurrence, magnitude, local, depth, etc.) influences the earthquakes and either exists a mathemathical model capable of supplying detailed and precise information about the relations and ways to estimate them. Therefore, to develop a prediction earthquake risk model can prove itself very complex.

2.2 Earthquake Prediction

In [14], Koza et al. say that Evolutionary Computation (EC) may find, by try trial and error and based on a great amount of data, better solutions for problems that human beings may not find it easy to solve. EC is a family of subfield of artificial intelligence that aim to extract patterns and to solve problems using a great amount of historical data. We may also say that without any domain knowledge about the problem to be controlled, the EC learns about may learn and find solutions for the problem. [16].

Genetic Algorithm (GA) is the chosen EC technique that is used in this study. It constitutes a category of heuristic search, they are stochastic algorithms and their search method are based on genetic inheritance and survival of the fittest [15]. They are interesting to be used specially in cases that are difficult to understand and the knowledge

available is not sufficiently available.

Based on these information and on the difficulty to understand how earthquakes behave, we want to explore historical earthquake data using GA. It is expected that it will help to find new ideas about earthquakes, their patterns and their mechanisms behind earthquake occurrences. For doing so, we need first to outline the forecast problem, then verify the suitability of Evolutionary Computation to the problem of generating earthquake forecast models.

Earthquake prediction is a polemic subject. No research has even come close to suggesting that individual large scale earthquakes can be predicted [1] and presently some scientists think that earthquake prediction may not only be fully impossible, but also that the resources needed for such a prediction may be out of reach [5].

In the context of this study, we do not aim to predict any individual earthquake and its major characteristics. Our goal relies on the fact that earthquakes do cluster in time and space. We want to use computer techniques to learn and to generate risk models. There is a lot of value behind the study of earthquake mechanisms, with the goal of generating statistical models of earthquake risk [25].

Next, we will study ways to improve the generated methods using both GA and/or other computer techniques and any seismological knowledge. We propose different representations that aim to refine the algorithm quality and to incorporate seismology methods to refine the models proposed.

Chapter 3

The Earthquake Forecasting Problem

This chapter focus on the theoretical concepts used as base for this study. The main topics are the Collaboratory for the Study of Earthquake Predictability (CSEP) framework and its tests.

3.1 Earthquake Likelihood Model Testing

We started studies of the earthquake forecasting problem by determining and selecting ways to build earthquake forecast models, to evaluate and to compare them, as suggested by the Collaboratory for the Study of Earthquake Predictability (CSEP). It is an international partnership to promote rigorous study of the feasibility of earthquake forecasting and predictability [1].

For that, we gathered some important information about earthquake predicting needed for this study. Most of it is based on the paper *Earthquake Likelihood Model Testing* [27]. From this paper we gathered information that guided us into how to build, evaluate and compare earthquake forecast models efficiently.

A very difficult and yet very common problem when studying earthquake models is how to compare different kinds of models, that are based on different tests protocols. The CSEP proposes a methodology for rigorous scientific testing of these many different models. This group proposed an framework called the CSEP framework. It provides a

method to compare earthquakes risk models in an objectively and consistently way [1].

All forecast models proposed in this study are based in the CSEP framework where a forecast model uses a gridded rate forecast [34], one common format in the literature. For evaluate and compare these models we used the likelihood based tests. They are the L-test, the N-test and the R-test, as suggested by Regional Earthquake Likelihood Model (RELM) [27].

The principle behind each consistency test is the same. One calculates a goodness-of-fit statistic for the forecast and the observed data. One then estimates the distribution of this statistic assuming that the forecast is the data-generating model (by simulating catalogues that are consistent with the forecast). One then compares the calculated statistic with the estimated distribution; if the calculated statistic falls in lower tail of the estimated distribution, this implies that the observation is inconsistent with the forecast, or that the forecast should be "rejected". For the CSEP consistency tests used here, the likelihood is the fundamental metric, but this approach would be similar for different statistical measurements [5].

3.1.1 Vector of expectations

The CSEP framework uses a gridded rate forecast, see 3.1. This gridded forecast may be structured by a vector of earthquake expectations, occurrences probabilities, that are directly related to a vector of real earthquake observations.

Based on this structure, it is possible to calculate the log-likelihood value of a model with the real data observed. It is also possible to use comparison tests based on the calculation of the log-likelihood.

3.1.2 The Log-Likelihood Function

To calculate the log-likelihood value we need both vectors cited above, in section 3.1.1. One of them is the vector of earthquake expectations and the other is the vector of real earthquake observations. On them, each element is considered a bin.

Each bin, b_n, define the set β and n is the size of the set β:

$$\beta := b_1, b_2, ..., b_n, n = |\beta|. \tag{3.1}$$

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:

$$\Lambda^j = \left(\lambda_1^j, \lambda_2^j, \ldots, \lambda_i^j\right); \lambda_i^j := \lambda_i^j(b_i), b_i \in \beta \tag{3.2}$$

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

$$\Omega = \left(\omega_1, \omega_2, \ldots, \omega_i\right); \omega_i = \omega_i(b_i), b_i \in \beta \tag{3.3}$$

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

$$L(\omega_i|\lambda_i^j) = -\lambda_i^j + \omega_i \log \lambda_i^j - \log \omega_i! \tag{3.4}$$

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:

$$\begin{aligned} L^j = L(\Omega|\Lambda^j) &= \sum_{i=1}^{n} L(\omega_i|\lambda_i^j) \\ &= \sum_{i=1}^{n} -\lambda_i^j + \omega_i \log \lambda_i^j - \log \omega_i! \end{aligned} \tag{3.5}$$

The fitness function is a coded version of the equation 3.5. It uses the probabilities of the bins of each individual of model for the λ values.

3.1.3 Uncertainties in Earthquake Parameters

It is important to say that the earthquake parameters, sources, as the location, magnitude and focal mechanism, cannot be estimated without uncertainties. Therefore, each parameter uncertainty has to be included in the testing [27]. Moreover, by estimating it, it is possible to judge the reliability and robustness of the forecast testing [5]. Also, each observation must be treated as independent ones. This is not the case of the aftershocks, once they are directly dependent on another prior earthquake.

3.2 Tests for evaluating Models

In the paper *Earthquake Likelihood Model Testing* [27], it is proposed some statistical tests that are used in this study, developed by the The Regional Earthquake Likelihood Models (RELM). They were used to compare and evaluate the every forecast models. These tests are based on the log-likelihood score that compares the probability of the events predicted by the model with the observed events.

To evaluate the data-consistency of the forecast models we used the N-Test, the Number Test, and the L-Test, the Likelihood Test. Therefore, assuming a given forecast model as the null hypothesis, the distribution of an observable test is simulated. If the observed test statistic falls into the upper or lower tail of this distribution, the forecast is rejected [28].

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

3.2.1 L-test - Data-consistency 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 following the formula, where $\widehat{L}_k$ is the value of the log-likelihood of the model j, in the *bin* k and $\widetilde{L_q^j}$ is the value of the log-likelihood of the simulation j in the *bin* q:

$$\gamma_q^j = \frac{\left|\left\{\widehat{L}_k^j | \widehat{L}_k^j \leq \widetilde{L}_q^j, \widehat{L}_k^j \in \widehat{L}^j, \widetilde{L}_q^j \in \widetilde{L}^j\right\}\right|}{|\widehat{L}^j|} \tag{3.6}$$

The analysis of the results can be splitted 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 simulations and is consistent with the data.
3. Case 3: γ^j is high. This means that the log-likelihood of the data 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.

3.2.2 Number test or N-Test

The N(umber)-Test also analyses the consistency of the model, but it compares the number of observations with the number of events of the simulations. This test is necessary to supply the under predicting problem, which may pass unnoticed by the L-Test.

This measure 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 distribution, then the model is consistent with the observation, nor estimating too much events nor too few of them.

3.2.3 The Likelihood Test or R-Test

The R(atio)-Test compares two forecast models against themselves. The log-likelihood is calculated for both models and then the difference between them is calculated, named the observed likelihood ratio. This value indicates which one of the model better fits the observations.

The likelihood ratio is calculated for each simulated catalogue. 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 [28].

3.2.4 Evaluation

The evaluation process is made as follow: first, the data-consistency is tested by the L-Test and R-test. If the model passes these tests, meaning that it was not rejected by them, they are compared with other forecast models, which were also not reject, with the R-Test. The model that best fits the R-Test is then chose as the best model [27].

Chapter 4

State of Art

In this Chapter we will briefly explain Genetic Algorithms and then discuss some reports of the application of Evolutionary Computation and related method for Earthquake Risk Analysis.

4.1 What are Genetic Algorithms

The main goal of a Genetic Algorithm (GA) is to find approximated solutions in problems of search and optimisation. Based on Koza [13], GA are mechanism of search based on natural selection and genetic. They explore historical data to find optimum search points with some performance increment, as said by Goldberg [6].

4.1.1 How does GA work

A GA uses those mechanisms to generate solutions to optimisation and search problems. The first step is to create an initial population of possible solutions. Frequently, the initial population is randomly generated once it is common to ignore the main aspects that influence the algorithm performance.

Each possible solution of a population is called an individual. Every individual is a possible solution of a problem. Those individuals have its fitness value estimated by a fitness function. A fitness function should determine how suitable a individual is to a given problem. The most suitable individuals are graded with better values and the not so suitable ones have a lower value.

After measuring the population fitness value, some individuals are then selected by a process that takes into account each individual fitness value to influence the next population. The individuals with better values have a higher chance to be selected. The individuals selected take part in the variation process. This process may alter some of the individual characteristics using the crossover and mutation operators.

The crossover operator is a operator that is used to vary the characteristics of a group of individuals. For that a number of parents, a group of individuals from the current population, are selected. In most of the cases, the parents are chosen to compose a pair that will exchange information that will take to compose the child, a new individual that will belong to the next generation.

Another important operator is called the mutation operator. It is a operator with the purpose of avoiding the loss of diversity of information. It works by changing the characteristics of an individual, looking to add new information to the next population.

It is common to have a evolutionary operator that allows the fittest individual from the current generation to take part in the next generation. This operator is called Elitism and it is used to assure that the next generation best solution is at least as good as in the current generation.

4.2 Evolutionary Computation and Earthquake Risk Prevision

The usage of Evolutionary Computation in the field of earthquake risk models is somewhat sporadic.

Zhang and Wang [35], Zhou and Zu [36] and Sadat et al. [24], used Artificial Neural Network (ANN) related with earthquake prediction. In all these works, they combine the ANN with some other technique, to achieve better results. They used a group of earthquake parameters, as the accumulated release energy, magnitude in a specific area and others. Some parameters in this group are not available in our earthquake database.

Those papers use the available parameters of the earthquakes that happened in the area of study to create a risk predictor of earthquake or to propose a magnitude range for future earthquakes. They object to consider each variables influence the most the results

so that their methods can achieve higher performance. We may compare and/or evaluate our method by comparing it to the works cited before.

Nicknam et al. [19] simulated some components from a seismogram station and predicted seismograms for other stations. They combined the Empirical Green's Function (EGF) with GA. The EGF method is used to synthesise acceleration time histories and the GA approach is developed to optimise the seismological model. They found that this method obtained good agreement with the observed data, but are not sure that results are free from uncertainties. Nicknam et al worked with more than 30 seismological model parameters. Although, that amount of parameters is not available to us, we can use the information from this paper to exam two options. The first, one may investigate if more earthquake parameters will improve our method and the other one is to analyse how they dealt with so many variables. Then we may consider to do the same and observe the results.

Kennett and Sambridge [9] used GA and associated teleseisms procedures to determine the Fault Model parameters of an earthquake. By doing so, they demonstrated that non-linear inversion can be achieved for teleseismic problems without any calculation of waves travel times. They used only P-wave data and expect that if more data could be introduced, the method would accomplish better results.

Some seismological models were developed aiming to estimate parameter values by using Evolutionary Computation. For example, Evolutionary Computation was used to estimate the Peak Ground Acceleration (PGA) of seismically active areas [12, 2, 10, 11].

The works done by Kerh [10, 11] are basely a combination of ANN and GA to estimate or predict PGA in Taiwan. These work are based on the benefits of mixing both techniques. They state that the usage of a purely ANN method to estimate PGA may fall into a local minimum and that can be avoid by combining ANN with GA, hence GA is a good method to find global optimums.

Their goal was to decide which areas may be considered potentially hazardous areas. They focused on urban areas, these works are important to revalidate building regulations, urban development and such. The earthquake variables that were used in these work are: local magnitude, focal depth, and epicentre distance. Both magnitude and depth are already used in our work, which is not the case of the epicentre distance variable. They also state that PGA is inversely proportional to epicentre distance, so to add data

about this variable may be useful to our work, once it could provide useful information to predict risk models both direct or indirectly.

Ramos and Vázques [23] 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. In some cases, the SAS has a better response time than the GA. They consider it to be once caused because the SAS only alerts when earthquakes with magnitude bigger than 5.0 degrees in the Richter scale occurs, while the GA deals with all the earthquakes.

Ramos's and Vázques's work is a important work because it helps the population to avoid bigger disasters caused by earthquakes by increasing the time response of the Seismic-Sense Stations. It has some similar feature as the one present in this document: it uses GA to prevent earthquake disasters and tries to locate targets in a given area (though the targets of this work is sensing stations and ours works target is the earthquakes location) and it proposes a methodology to do a GA parameter setting to find which combination of values for the GA parameters achieve higher results. It is interesting to state that once a solution places a station in an area that is not possible to have sensors, this possible solution suffers some penalties

Saeidian et al. [26] also based on the same idea of locating sensing stations. His work differs from the work of Ramos because it makes 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 [7] published a paper in which the goal is to find, via Genetic Algorithm, 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 [7]. This research aims to obtain a structure of the Japanese underground and geographically focuses in the same region as our work, though it uses data from two kinds of waves which are not available to us.

Chapter 5

Models

As stated in the Section 3.1, all forecast models proposed for this study are based in the Collaboratory for the Study of Earthquake Predictability (CSEP) framework.

We propose four forecast model methods. The main difference between them is that they have different genome representation. The genome for each forecast model focus on different aspects of the framework, therefore their represantion vary.

The first method is the GAModel [1], a statistical method of analysis of earthquakes risk using the Genetic Algorithm (GA). It is a straight application of the CSEP framework. The next method is a specialization of GAModel. It focuses only on areas on which earthquakes happened already in a near past. This will lead to a faster convergence, once the amount of parameters is smaller and consequently, the search space gets smaller. We called it ReducedGAModel. These methods used only computational algorithms and techniques.

Another method is the Emp-GAModel. This method incorporates some geophysical knowledge. It is a hybridisation of the models generated by the GAModel with some empirical laws that will be discussed further, in Section 5.3. We also applied these empirical laws in the ReducedGAModel, and name it Emp-ReducedGAModel.

For all methods, the population is evolved taking into account earthquake event data for a training period, which is anterior to the target test period. After completing the evaluation stop criteria, the best individual is chosen to be the representative forecast model for that method.

5.1 1-year Forecast Models

Based on the gridded rate forecast explained in the Section 3.2.4, we developed earthquake forecasts methods that will estimate the risk of earthquakes occurrence to a target region, during a time interval. Some of the methods also may estimate the magnitude of these shocks. For this study we considered the target time interval of one year [1].

There is no physical relationship to identify mainshocks and its aftershocks [28], we divided the forecast models in two classes: 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.

Mainshocks are large and independent earthquakes. They are followed by a wave of others earthquakes, the aftershocks [28].

5.2 Mainshock Methods

The mainshock methods are considered as methods to generate space-rate-time forecasts. They could be described as:

$$\Lambda(t, x, y, M|\Upsilon_t) = \mu(x, y) \tag{5.1}$$

where the number of earthquakes forecast in all bins can denoted as $\Lambda(t, x, y)$ [34] given that Υ_t is the earthquake observation data up to time t.

5.2.1 GAModel

The GAModel is completely based on the framework suggested by the CSEP. In it, 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 these bins where each bin has an earthquake forecast value. This groups of bin represent the $\mu(x, y)$, the background intensity [37]. In the GAModel, each possible solution is represented as an entire forecast model.

The GAModel forecasts only earthquakes with magnitude greater than 3.0, for every scenario proposed. The space interval for the magnitude is 0.1, named as cells. That results in magnitude cells of [3.0, 3.1), [3.1, 3.2), until [9.9, 10).

Genome Representation

In the GAModel each individual represents an entire forecast model. Each gene of the individual 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 5.2.1 used in [1]. In the algorithm 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

$L \leftarrow \exp(-\mu), k \leftarrow 1, prob \leftarrow 1 * x$
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 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 and they are randomly generated. For more details of the genome representation, please refer to [1].

To clarify how the GAModel works, we use the same example as the one used in [1]. The Kanto region, one of the four areas used in both studies, is divided into 2025 bins (a grid of 45x45 squares). Each bin has an area of approximately $25km^2$. 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.

Fitness Function

To compare the individual data with the observated data, we use the log-likelihood calculation as fitness function. This equation allow us to compare events in the observated data with the values of occurrences obtained by a model. The models that have more similarity with to the observated data have bigger log-likelihood values. 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 fitness function is a coded version of the Equation 3.5. It uses the probabilities of the bins of each individual of the model for the λ values.

Evolutionary Operators

The GAModel use a combination of operators made available by the Distributed Evolutionary Algorithms in Python (DEAP) [4]. 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 5.1.

Tabela 5.1: 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 operator function and crossover operator function 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.

5.2.2 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 also 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. As said in the last paragraph, 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.

The ReducedGAModel, as the GAModel, forecasts only earthquakes with magnitude greater than 3.0, for every scenario proposed. The space interval for the magnitude is 0.1, named as cells. That results in magnitude cells of [3.0, 3.1), [3.1, 3.2), until [9.9, 10).

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 estimative for this bin.

To calculate the size of the individual we use the real data from the worse 5 years and create a list of every bin that had events in it, even if only once.

In the ReducedGAModel, each individual is a list of a subregion of the forecast model. This list initially refers to bins where earthquake events happened in the past. During the develop 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 EmpGAModel, once the amount of subregions where earthquakes with magnitude above 3.0 happened for any given area is smaller then the total number of genes of the individual.

To exemplify, we use a similar example as the one in the Section 5.2.1. 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, 5.2.1. 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 5.2.

Tabela 5.2: Parameters used in ReducedGAModel

Population Size	500
Generation Number	100
Elite Size	1
Tournament Size	3
Crossover Chance	0.9
Mutation Chance (individual)	0.1

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.

5.3 Mainshock with Aftershock Methods

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:

$$\Lambda(t, x, y, M|\Upsilon_t) = \mu(x, y)J(M) \tag{5.2}$$

That can be expanded to:

$$\Lambda(t, x, y|\Upsilon_t) = \mu(x, y) + \sum_{t_i \in t} K(M_i)g(t - t_i)P(x, y) \tag{5.3}$$

methods use $\mu(x, y)$ as defined for mainshock methods 5.2. 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.

Empirical Equations

The Omori law, $g(t)$, which is considered one empirical formula of great success [37] [30] [22], 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 [37]:

$$g(t) = \frac{(p-1)}{c(1+\frac{t}{c})^{(}-p)} \tag{5.4}$$

The variable p is a index of this equation and the variable c is a constant, given in days. In the paper [30], 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 [30]. Also there is the variable t that is the time limit to when a mainshock may influence the cause a aftershock.

Based on paper [33], 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 $M_c = 3.0$:

$$K(M_i) = A\ exp([\alpha(M_i - M_c)]) \tag{5.5}$$

In the paper [21], 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 [33], was used:

$$A = e^{(1.02M-4)} \tag{5.6}$$

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:

$$\sum_{t_i \in t} K(M_i)g(t-t_i) \tag{5.7}$$

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:

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

The last equation, the $J(M)$ from equation 5.2 is a simulation of the event magnitude by Gutenberg-Richter's Law, using the function etasim, from the SAPP R package [20].

5.3.1 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.

The genome representationm the fitness function and evolutionary operatores are the same as in the GAModel, see Sections 5.2.1, 5.2.1 and 5.2.1 for more information.

5.3.2 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.

The genome representation is the same as in the ReducedGAModel,Section 5.2.2. The Emp-ReducedGAModel use the same combination of operators that the ReducedGAModel. For more explanation, please see 5.2.1.

The fitness function is the same as for all methods,Section 5.2.1. Here is also important to generate the forecast model by applying the map function on the individual as in the last Section, 5.2.2.

Chapter 6

Data Analysis

In this Chapter 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.

6.1 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 2.2. This earthquake triggered too many after called aftershocks in all Japan. It is considered that big earthquakes may cause others earthquakes [37]. In Figure 6.1 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 tranning base to earthquakes until 2010.

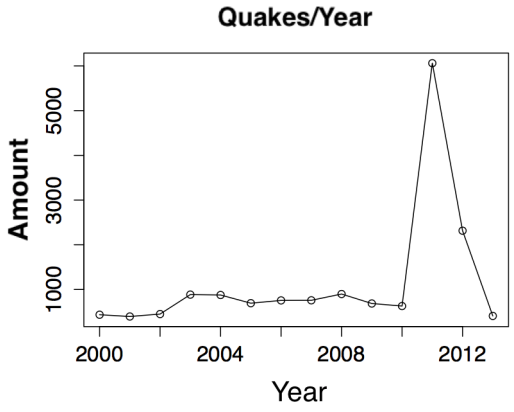

Figura 6.1: Amount of earthquake by year.

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.

6.2 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 6.2 shows how we defined them.

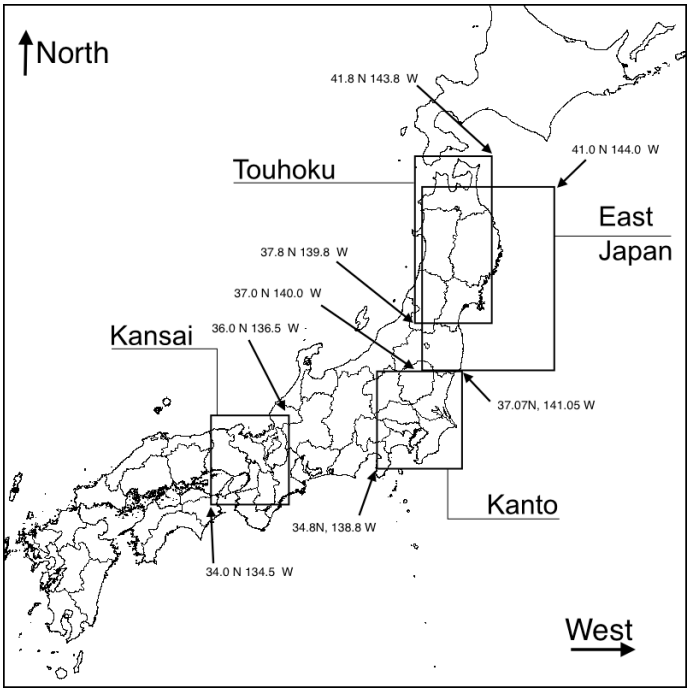

Figura 6.2: Japan and the areas used in this studied.

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.

6.3 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 6.3, 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.

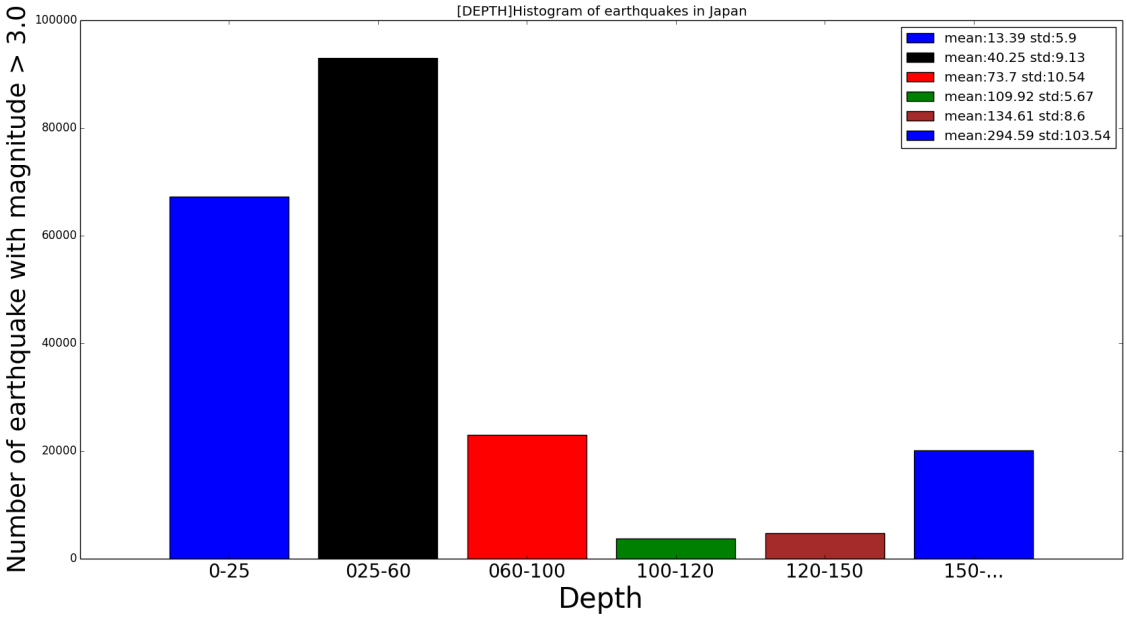

Figura 6.3: Depth Histogram of earthquakes.

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 [33].

6.3.1 Mainshocks and Aftershocks - Clustering

In the Section 4.2, 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 6.4.

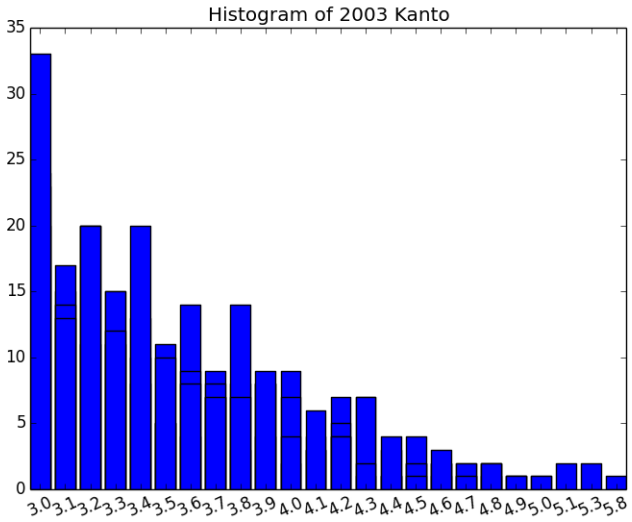

Figura 6.4: Histogram of earthquakes stronger than 3.0 showing the Gutenberg-Richter relation in Kanto

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 [31].

Chapter 7

Experimental Design

In this Chapter we report the experimental design of the models described earlier.

The first experiment was a pilot experiment. The goal was to estimate the number of repetitions needed to compare the GAModel with the RI Algorithm. For that, we used the Power of the Student t-test. We created scenarios composed by the GAModel applied in the region of Kanto for the years of 2000 to 2010. For each scenario, the GAModel was applied 10 times, leading to 10 observations/scenario. Than, we used these observations log-likelihood values for calculating the Power of the Student t-test.

Based on the results from the Power of the Student t-test, we were able to compare the RI Algorithm with the GAModel and with the randomModel. We generated 1 model per scenario for both the control method, the randomModel, and for the RI Algorithm. Then, we applied the Student t-test with the confidence level of 95%. to compare with the means from the models of the GAModel were higher than the models from the RI Algorithm and the randomModel.

The next experiment was made to compare the all the models proposed in the Chapter 5.3.2 with each other and to discover which method would achieve higher log-likelihood values. We created new scenarios, applying the methods for all regions and for the years of 2005-2010. We also used 3 kinds of catalogues: the JMA and the de-clustered catalogues form the Window method and the SLC method. Then, 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.

We also made a magnitude experiment. This experiment was done to explore the influence of the magnitude in all models generated. We split them into slices composed of earthquakes that have magnitude in a given magnitude interval. The we calculated the log-likelihood of these slices and applied the ANOVA test to compared these sliced-models.

7.1 The GAModel

We used the log-likelihood value as the fitness function. For the experiments, we divided the data into annual slices, as described in the Chapter 6, for the region of Kanto.

Once no effort was made to analyse the GA initial population and number of generations, we chose them by trial and error, until an acceptable convergence time were achieved.

7.2 The GAModel Initial Experiment

We create some scenarios to compare the GAModel with the RI - Relative Intensity Algorithm. These scenarios are defined as space/time regions. Each scenario contains the earthquakes for the Kanto region for a given year.

As being a stochastic method we used the Power of the statistical test and estimated the number of repetitions, n, needed for detecting a significant variation. For the Power of the Student t-test we used a pilot experiment to estimate n.

This pilot experiment was done by applying the GAModel in the region of Kanto for the years of 2000-2010. We created 10 observations for each scenario. We used them for calculating the Power of the Student t-test. The *delta* and *power* values were chosen based on the results of the pilot experiment. The standard deviation is the same as the observations. All results indicate that 10 repetitions are enough to compare the GAModel via Student's t-test.

An example of the Power of the Student t-test is given in the next table 7.1:

We also needed the log-likelihood of the RI method. As being a deterministic method, one observation for each scenario is enough. This value is used as the target for the Student's t-test. Also, we wanted to verify if the GAModel has quality higher than a

Year	2006
Delta	50 (for all scenario)
Standard Deviation	25.16513
Significance Level	0.05
Power of test	0.95
Alternative	two.sided
n:	5.58517

Tabela 7.1: Power t-test.

model with the values in the bins sampled from a Poisson distribution with mean one. We named this models as the randomModel and it has no data awareness.

7.2.1 The Relative Intensity Algorithm

The RI *Relative Intensity* (RI) algorithm is frequently used as reference for comparing methods [18]. The main idea behind the RI is that larger earthquakes are more likely to occur at locations of high seismology in the past.

The log-likelihood data for the RI for each scenario were given by Aranha et al. [1].

7.2.2 Evolutionary Operators

The GAModel use a combination of operators made available by the Distributed Evolutionary Algorithms in Python (DEAP) [4]. 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. These parameters are presented in the Table 7.2.

7.2.3 The GAModel and The RI Algorithm

We compared the method proposed, ReducedGAModel, with the non-variant method, GAModel, and with the Relative Intensity Algorithm, using its log-likelihood values. The values are compared via Student's t-test, so we could understand the if there is any statistic significant difference between the methods. The data used was gotten from the JMA catalogue.

Tabela 7.2: Parameters used in 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	beta = 1, low = 0, up = 1

7.2.4 Hypotheses

There are three tests hypothesis for this experiment that we want to analyse. For all, the confidence level is set to 95%.

The first is if the mean values of the log-likelihood for the ReducedGAModel are equal to the RI values.

$$\begin{cases} H_0 : \mu = \text{RI log-likelihood value} \\ H_1 : \mu! = \text{RI log-likelihood value} \end{cases}$$

The second is if the mean values of the log-likelihood for the ReducedGAModel are equal to the GAModel values.

$$\begin{cases} H_0 : \mu = \text{GAModel log-likelihood value} \\ H_1 : \mu! = \text{GAModel log-likelihood value} \end{cases}$$

And the last hypothesis is if the mean values of the log-likelihood for the GAModel are equal to the RI values.

$$\begin{cases} H_0 : \mu = \text{RI log-likelihood value} \\ H_1 : \mu! = \text{RI log-likelihood value} \end{cases}$$

7.3 The Models

Based on our promising results and because we aim to improve them, we developed the ReducedGAModel. It a simplified version of the GAModel. We want to compare the behaviour of this new method against the GAModel method.

We also wanted to explore how adding domain knowledge would improve the average quality of the GAModel or the ReducedGAModel They are versions of GAMo-

del/ReducedGAModel combined with empirical laws.

7.4 The Mainshock Models and Mainshock with Aftershock Method Experiments

Here we describe the catalogues and the evolutionary operators used for the experiments. Then we specify the models comparison.

7.4.1 The catalogues

The data used was from JMA catalogue, with the minimum magnitude of 3.0 and the two de-clustered catalogues, obtained from the methods explained in the Section 6.3.1. The models that use these catalogues have in the word Window appended at their names, for the methods that used the Window declustering, or *SLC*, for the methods that used the Single Link Cluster.

7.4.2 Evolutionary Operators

For the ReducedGAModel the only different operator is 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 length of the genome. For the parameters see Table 7.3.

Tabela 7.3: Parameters used in ReducedGAModel

Population Size	500
Generation Number	100
Elite Size	1
Tournament Size	3
Crossover Chance	0.9
Mutation Chance (individual)	0.1

The Emp-GAModel uses the same operators as the GAModel. The parameters are presented in the Table 7.2.

The Emp-ReducedGAModel uses the same operators as the ReducedGAModel. The parameters are presented in the Table 7.3.

7.4.3 Models Comparison

For this new experiment, we used even more scenarios (space/time regions) than the others. Each scenario contains the earthquakes for the regions of Kanto, Kansai, Touhoku and East Japan for a given year (2005-2010). We wanted to explore if there exists any influence in the performance of the models that are caused by the depth of an earthquake. So, the scenarios are also composed by introducing a three groups depths thresholds. They are: of earthquakes with depth smaller than 25km, or between 0km and 60km or even between 0km and 100km.

These methods are stochastic methods and hence are variations of the GAModel, we decided to maintain the number of repetitions without redoing the Power of the Student t-test.

7.4.4 Statistical Analysis of the Results

The goal is to discover if there is any variation between the methods and which are the most influential variables. For achieve that, we will use the ANOVA test, because it indicates it the means of several groups are equal or not for a given confidence interval. The confidence interval was set to 95%.

There are some tests hypothesis for this experiment that we want to analyse. They all can be generalised as follows:

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

Then, if there is no statistical significant difference between the means, we apply a post hoc methodology analysis. The chosen post hoc was the HSD Tukey. We apply it on the results obtained from the ANOVA test to specify which groups differ.

Tukey's methodology analysis shows the means of a case with the means of every other case. Doing so, it identifies differences between means :

$$\left\{\mu_a - \mu_b, \text{where } \mu_a \text{ is the mean of the first group and } \mu_b \text{ is the mean of the second group.}\right.$$

In the case where statistical significant difference exists, we explore this by pairing the measures observations of two groups [3].

That is:

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

7.5 Magnitude Experiment

In this experiment, we focused on studying how the magnitude of an earthquake affects the model quality, because the patterns of the earthquakes are depended of its magnitude. We wanted to explore the relation between the magnitude of the earthquakes and how would the models behave on those situations.

For that, we created magnitude intervals, where each interval is named as a slice. A slice is an closed interval of 1.0 degree starting from 3.0 degrees of magnitude, see Section 7.4.1, and ending in 10.0 degrees. For example, $[3.0 - 4.0]$ or $[7.0 - 8.0]$ are two different slices. For each model, we selected only the earthquakes that belong to a slice. Then, we calculate the log-likelihood value.

7.5.1 Catalogues and Models

This experiment used the same catalogues used in the previous experiment 7.4.

The models also are the models from the last experiment. We also created the new models, considering the slices and add them to the comparison. That lead to a comparison with the models from the last experiment and the models sliced.

7.5.2 Statistical Analysis

The goal is to discover if the magnitude influences any variation in the methods and how it does.

For this experiment, we followed the same design from the Section 7.4.4.

Chapter 8

Results

This Chapter is dedicated to show the results from the experiments explained in the last Chapter.

8.1 Results from the GAModel Experiment

The results of the experiments are in the Table 8.1 and in the box-plots Figure 8.1 and Figure 8.2. In the Table 8.1, the column labelled Random shows the result of the RandomModel, and the idea is the same for the ones labelled GAModel and RI. The "p-value" shows the significance value of the *Student's t-test* for the null hypothesis " The mean of the log-likelihood values is greater that the values for RI".

As expected, the RandomModel has lower values than the GAModel. When compared with the RI, the results show that the GAModel is competitive with the RI and that it is promising to use GA to generate earthquake forecasts.

Scenario	Log Likelihood				
Year	Random	RI	GAModel		p-value
2000	-2413.89	-2124.44	**-2094.05**	(8.80)	0.01
2001	-2418.14	-2103.19	**-2101.65**	(69.49)	0.57
2002	-2385.04	**-2094.43**	-2100.01	(72.62)	0.07
2003	-2401.00	-2104.65	**-2100.76**	(156)	0.35
2004	-2421.92	-2101.92	**-2098.30**	(55.28)	0.16
2005	-2643.38	-2248.40	**-2114.00**	(779)	0.01
2006	-2616.50	-2226.93	**-2115.6**	(633)	0.01
2007	-2451.68	**-2109.13**	-2122.03	(615)	0.13
2008	-2433.23	**-2112.92**	-4435.34	(657)	0.14
2009	-2884.74	-2438.10	**-2113.1**	(814)	0.01
2010	-2418.18	-2114.60	**-2112.07**	(843)	0.79

Tabela 8.1: GAModel Experiments Results. The highest results are shown in bold lines.

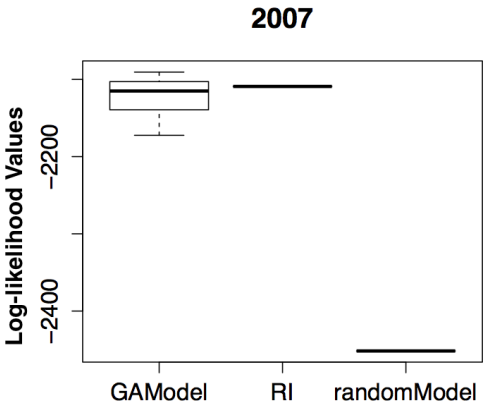

Figura 8.1: Box-plot of the values obtained by the models for the year 2007.

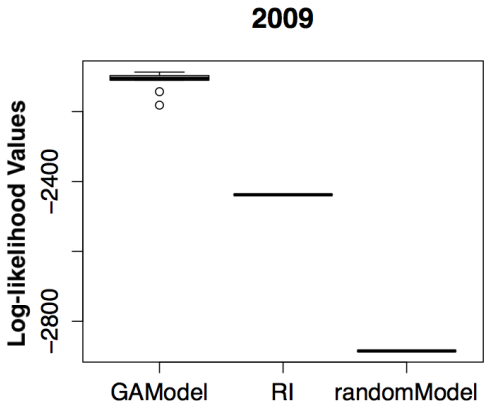

Figura 8.2: Box-plot of the values obtained by the models for the year 2009.

8.1.1 Models Example and The Real Data

The Figure 8.3 shows a model from the GAModel method for the year 2010. It indicates a low earthquake intensity as green while the more intensity areas, are shown as orange (for even higher cases, white is used). The Figure 8.3, from [1], shows a model from the RI Algorithm for the year 2010. It indicates a low earthquake intensity as white while the more intensity areas, are shown in red. For comparison reasons, we show the Figure 8.4, that shows the earthquake occurrences for the same year.

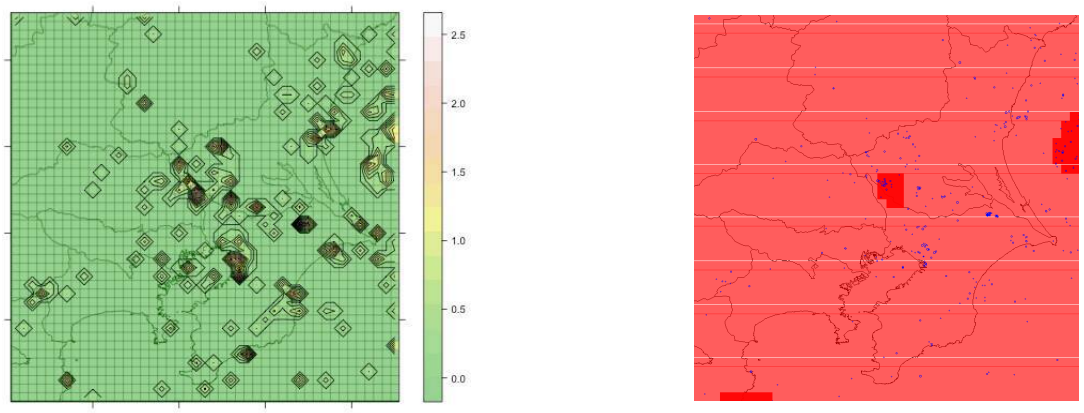

Figura 8.3: The Picture on the left, is the GAModel model for the year of 2010 in Kanto and the one on the right, is the RI model for the year of 2010 in Kanto.

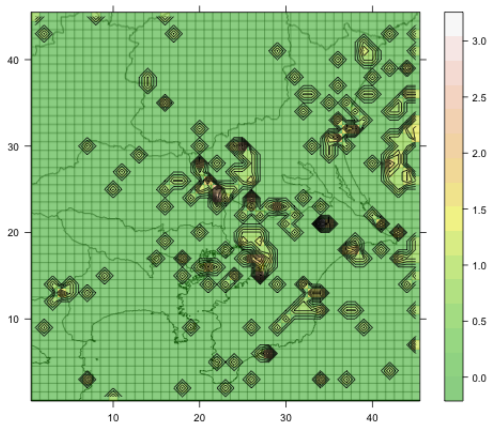

Figura 8.4: Earthquake occurrences in the year of 2010 in Kanto.

8.2 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 depths, the years and regions on the log-likelihood value. In this study there are the models: ReducedGAModel, Emp-ReducedGAModelSLC, Emp-GAModel, Emp-ReducedGAModel, GAModelWindow, ReducedGAModelWindow, GAModelSLC, ReducedGAModelSLC, Emp-GAModelClustered, Emp-ReducedGAModelWindow, GAModel and Emp-GAModelSLC .

Based on the results of this first test, it is evident that all variables are significantly different. The results of the experiments are in the Table 8.2. For all, the confidence

interval is set to 5%.

Variable	Degrees of Freedom	Sum Sq	Mean Sq	F Value	Pr(>F)
Model	11	1.984e+08	18035604	113.87	<2e-16
Depth	2	2.955e+07	14772789	93.27	<2e-16
Year	5	1.065e+09	213025401	1344.95	<2e-16
Region	3	2.188e+09	729443498	4605.38	<2e-16

Tabela 8.2: ANOVA Test Results Values - Mainshock Models Mainshock and Aftershock Models.

Because we found statistically significant result, we applied a Post hoc comparisons using the Tukey HSD analysis methodology. It compared each condition with all others. For example, it compares the values from the GAModel with the GAModelWindow, see Figure 8.5. It indicated that the models that used the catalogues from the Window Method or the Single Link Cluster, when compared with all other models, achieve greater log-likelihood values. Furthermore, we noticed that the depths conditions show a greater influence when the depth in smaller or equal to 25 km, see Figure 8.6.

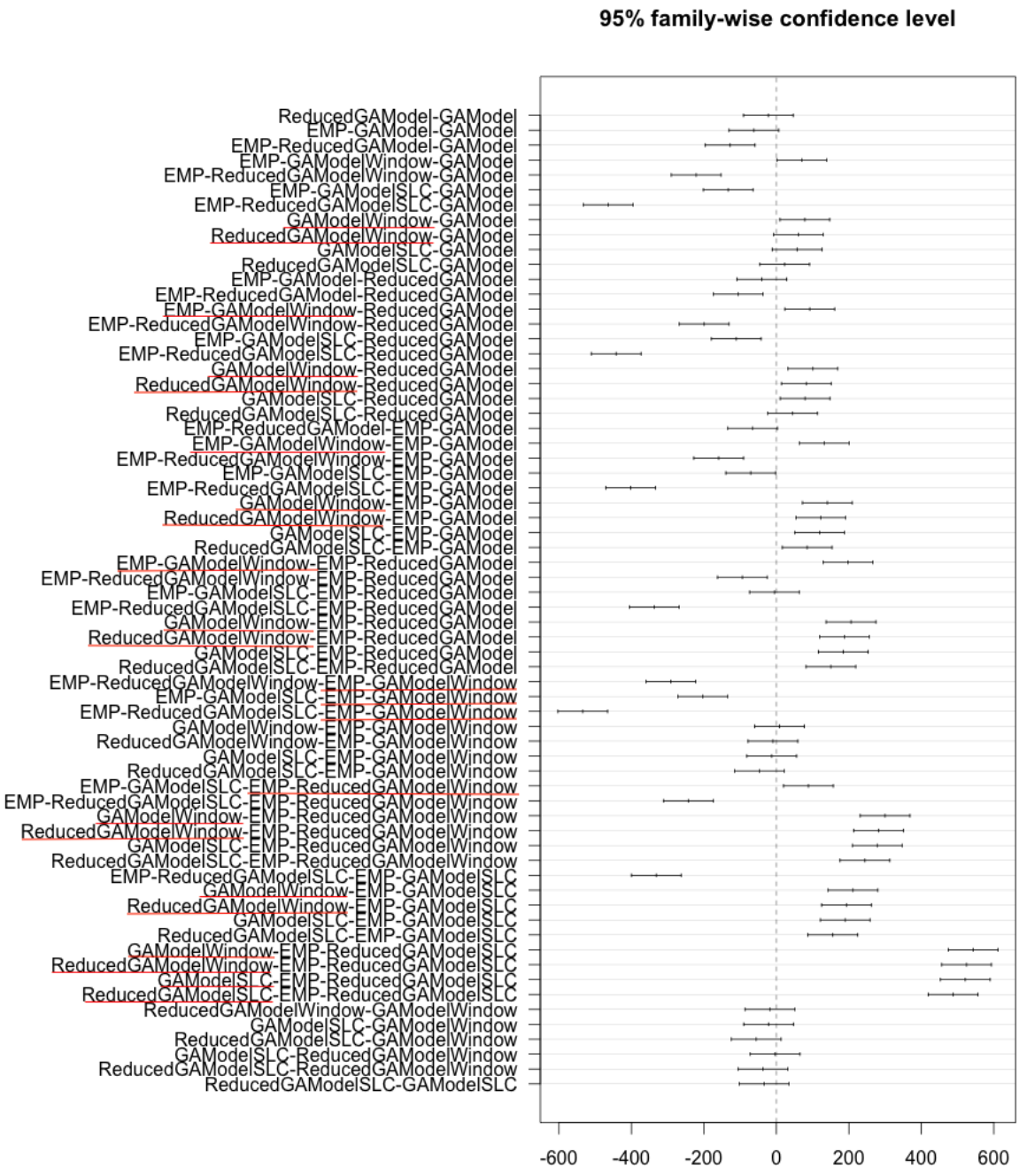

Figura 8.5: Intervals of Confidence 95% of differences between the Mainshock Models Mainshock and the Aftershock Models, taken two by two.

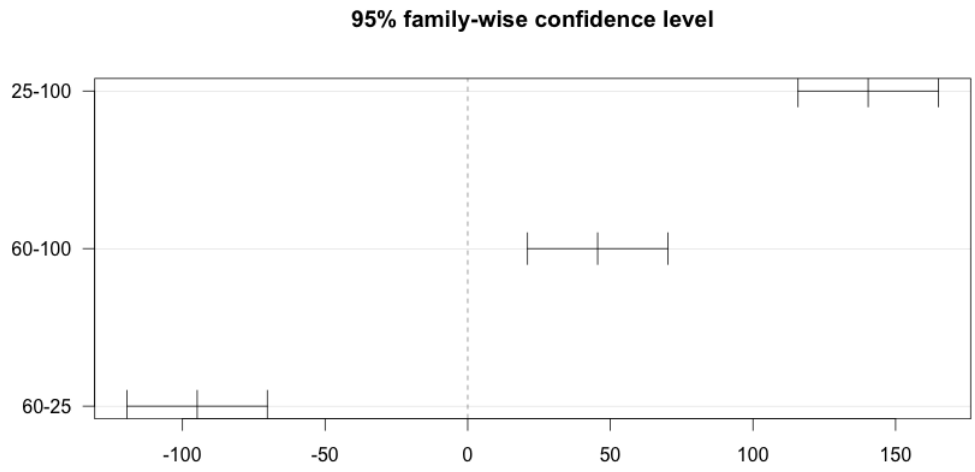

Figura 8.6: Intervals of Confidence 95% of differences between the depths, taken two by two.

Then, we decided to compare only the those models against themselves. That is, models used for this new comparison are those that used the catalogues from the Window Method or the Single Link Cluster and with earthquakes with depth smaller or equal to 25 km.

Variable	Degrees of Freedom	Sum Sq	Mean Sq	F Value	Pr(>F)
Model	7	21991488	3141641	21.07	<2e-16
Year	5	240753808	48150762	322.94	<2e-16
Region	3	374602684	124867561	837.48	<2e-16

Tabela 8.3: ANOVA Test Results Values - Models with depth smaller or equal to 25 km..

Based on the results of this test, it is evident that all variables are still significantly different. The results of the experiments are in the Table 8.3. For all, as before, the confidence interval is set to 5%.

Again, we found statistically significant result, therefore, we applied the Tukey HSD test. The results are shown in Figure 8.7. It indicated that the models from the GAModel method, when compared with all other models, achieve greater log-likelihood values.

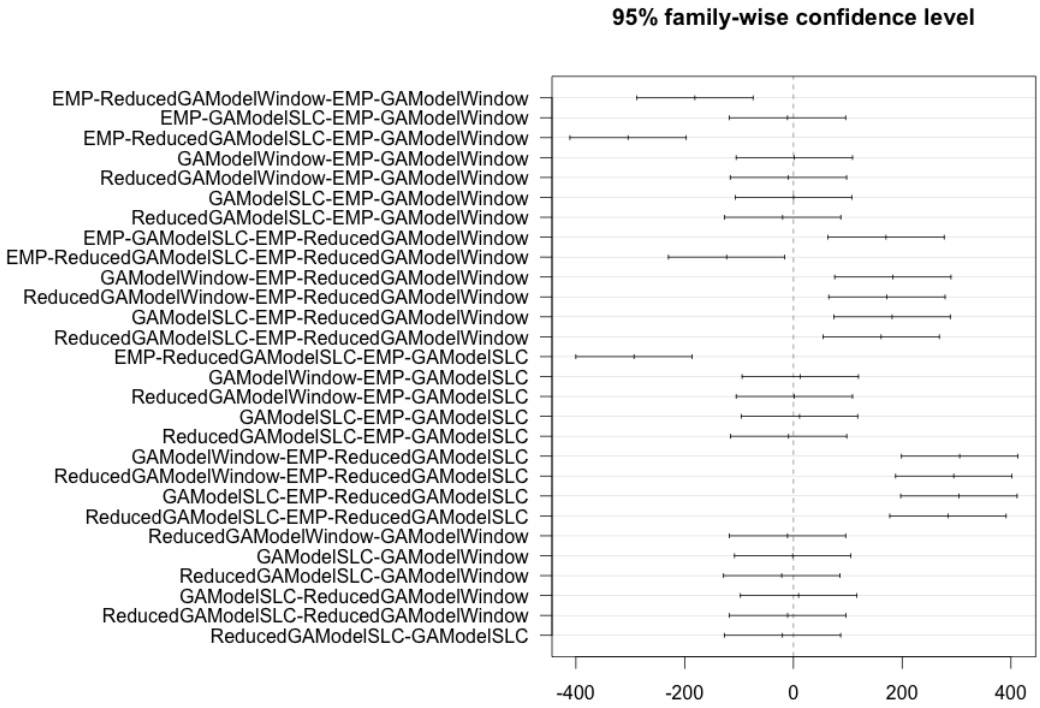

Figura 8.7: Intervals of Confidence 95% of differences between the models with depth smaller or equal to 25 km, taken two by two.

Based on the last results, we performed the ANOVA again, comparing only models from the GAModel method. The results of the experiments are in the Table 8.4.

Variable	Degrees of Freedom	Sum Sq	Mean Sq	F Value	Pr(>F)
Model	3	24462	8154	0.058	0.981
Year	5	121287074	24257415	173.600	<2e-16
Region	3	167719464	55906488	400.099	<2e-16

Tabela 8.4: Simple ANOVA Test Results- GAModels.

This time, we found statistically significant result only for the year and region condition. 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 *GAModelWindow* as the representative method of this study. That is because, in most cases when its values were compared, it showed a little better performance in the means of the log-likelihood values. For the results, see the Table 8.5.

In this Table, the column labelled $\mu_a - \mu_b$ shows the result of paired difference between the models referred in the Models Compared column. The "p-value" shows the significance value of the paired *Student's t-test* for the null hypothesis " The paired difference of the means of the models is equal" In Touhoku, the models Emp-GAModelSLC and the GAModelSLC had the exactly the same mean, therefore, the difference is 0 and the p-value could not be calculated.

Region	Models Compared	$\mu_a - \mu_b$	p-value
Kansai	**EMP-GAModelWindow** - Emp-GAModelSLC	4.405143	<0.04
	EMP-GAModelWindow - **GAModelWindow**	4.405143	<0.01
	EMP-GAModelWindow - **GAModelSLC**	-2.32017	<0.01
	EMP-GAModelSLC - **GAModelWindow**	-7.497507	<0.01
	EMP-GAModelSLC - **GAModelSLC**	-6.725214	<0.01
	GAModelWindow - GAModelSLC	0.772193	<0.03
Touhoku	EMP-GAModelWindow - **Emp-GAModelSLC**	-0.736392	<0.03
	EMP-GAModelWindow - **GAModelWindow**	-0.5847768	<0.03
	EMP-GAModelWindow - **GAModelSLC**	-0.736392	<0.03
	EMP-GAModelSLC - GAModelWindow	0.1516152	<0.04
	EMP-GAModelSLC - GAModelSLC	0	< NA
	GAModelWindow - **GAModelSLC**	-0.1516152	<0.04
East Japan	**EMP-GAModelWindow** - Emp-GAModelSLC	39.29195	<0.01
	EMP-GAModelWindow - **GAModelWindow**	-2.22763	<0.04
	EMP-GAModelWindow - GAModelSLC	2.558211	>0.06
	EMP-GAModelSLC - **GAModelWindow**	-41.51958	<0.01
	EMP-GAModelSLC - **GAModelSLC**	-36.73374	<0.01
	GAModelWindow - GAModelSLC	4.785841	<0.01
Kanto	**EMP-GAModelWindow** - Emp-GAModelSLC	0.6959113	>0.08
	EMP-GAModelWindow - **GAModelWindow**	-0.7598627	<0.01
	EMP-GAModelWindow - **GAModelSLC**	0.6668383	<0.03
	EMP-GAModelSLC - **GAModelWindow**	-1.45774	<0.01
	EMP-GAModelSLC - **GAModelSLC**	-1.36275	<0.01
	GAModelWindow - GAModelSLC	0.09302443	>0.2

Tabela 8.5: Paired Experiment Result.

8.2.1 The Models Examples And The Real Data

The Figure 8.8 shows a model from the GAModel method for the year 2005 in East Japan. The next Figure, 8.9 shows a model from the ReducedGAModel method for the year 2005 in East Japan.

All Figures, 8.8 8.8, 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 8.10 represents the earthquake occurrences in the same region and year.

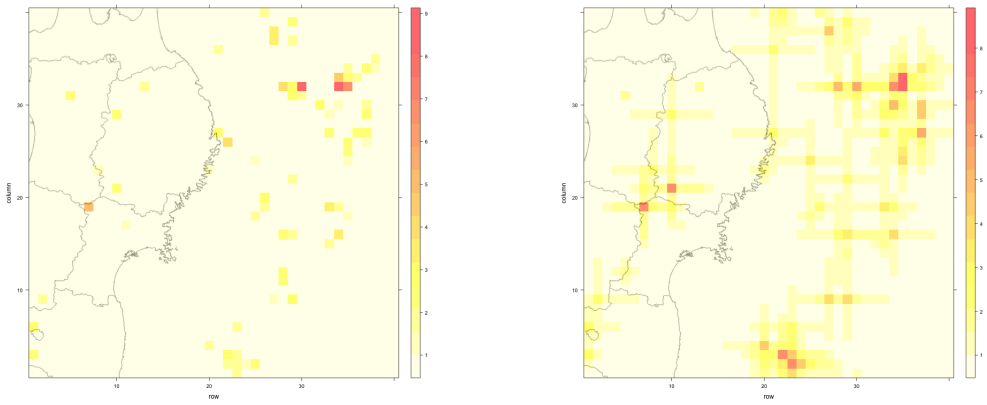

Figura 8.8: The Figure on the left is the GAModel model for the year of 2005, East Japan, and the one on the right Emp-ReducedGAModel model for the year of 2005, East Japan.

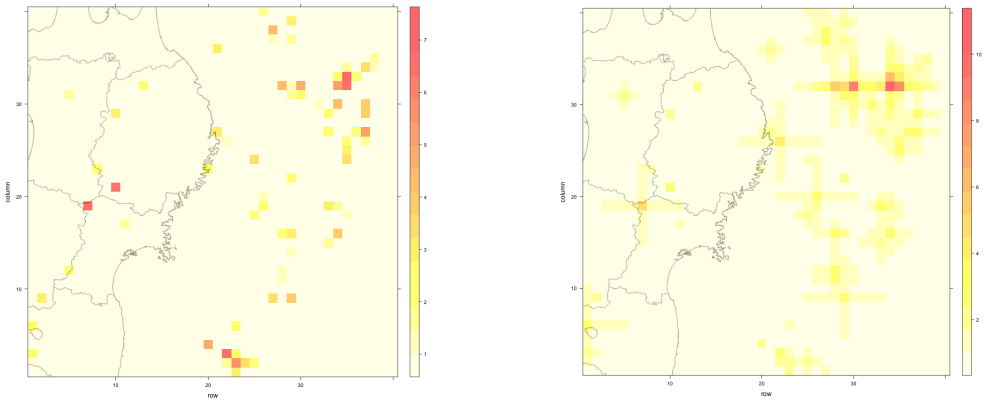

Figura 8.9: The Figure on the left is the ReducedGAModel model for the year of 2005, East Japan, and the one on the right Emp-GAModel model for the year of 2005, East Japan.

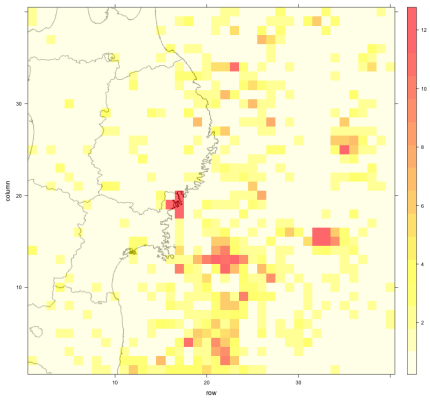

Figura 8.10: Earthquake occurrences in the year of 2005 in East Japan.

Figura 8.11: The Figure on the left is the Emp-ReducedGAModel model for the year of 2005, East Japan, and the one on the right GAModel model for the year of 2005, East Japan, East Japan.

8.3 Magnitude Study

From the results already obtain and showed in the Section 8.2, when selected the models with earthquakes with depth smaller or equal to 25 km and then we split the models in magnitude intervals, as defined in 7.5.

After that, 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 experiments are in the Table 8.6. For all, as before, the confidence interval is set to 5%.

Variable	Degrees of Freedom	Sum Sq	Mean Sq	F Value	Pr(>F)
Model	5	2.368e+09	4.737e+08	3058	<2e-16
Year	3	4.139e+09	1.380e+09	8906	<2e-16
Magnitude	7	3.212e+09	4.588e+08	2962	<2e-16

Tabela 8.6: ANOVA Test Results Values - Magnitude Study.

We found statistically significant result and, as before, we applied the Tukey HSD test. The results are shown in Figure 8.12 and the *NULL* field was used as the model with all magnitude intervals (the complete model).

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.

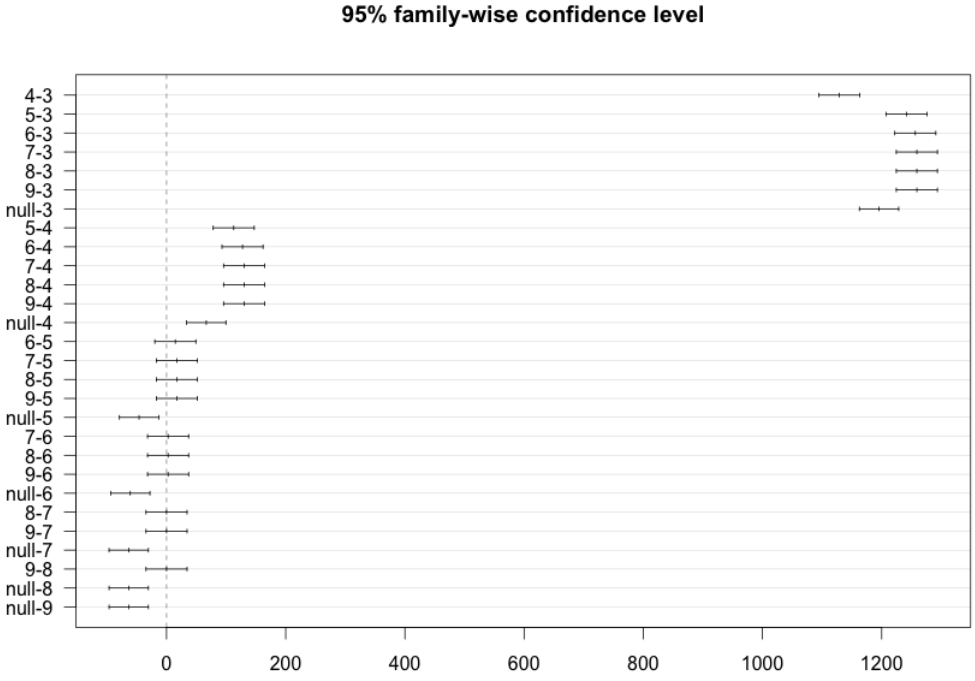

Figura 8.12: ANOVA results - Models from Magnitude Study.

Chapter 9

Conclusion

In this work we proposed some methods to generate earthquake risk model. We described how they were built and their characteristics. All models were based on the GAModel, and differ from it in the genome representation, the use of seismological equations or the catalogue used.

Initially, we studied the CSEP framework. Based on it we proposed a GA method to generate a earthquake risk model. We implemented the method and did some experiments comparing it with the model from the RI Algorithm. The scenario used for this experiments was composed by the years of 2000 to 2010 for the region of Kanto. The results showed that this method is competitive with the RI Algorithm.

Supported on this results and because we wanted to improve the performance of the method in terms of log-likelihood values, we proposed two methodologies. The first, is to change the genome representation. The other is to use seismological equations that would improve the accuracy of the methods by adding domain awareness to them.

The main goal of changing the genome representation is to minimise the search space. We expected that by doing so, the GA would have to consider less possibilities and would find a good solution with less computational effort and/or could lead to models with higher log-likelihoods models.

The new representation achieved similar results as the one from the GAModel. Hence we expected to achieve greater log-likelihood values than the GAModel, we consider that more efforts should be direct to refine this representation. We think that this representation would benefit if we consider areas that contained more than one bin, instead of how

it is now. Because it would have some flexibility in positioning the earthquakes.

Secondly, we studied the ETAS model and the seismological equations that it uses. That lead to the study of the Omori-Utsu formula and some others related formula. From this study, we encountered many difficulties, mostly related to lack of earthquake background knowledge. Although, during this study we realised that it would be interesting to analyse how the depths influence our methods and to consider a classification of the earthquakes into mainshocks and aftershocks.

The usage of the seismological equations showed no improvement to the models. We think, though, that there is still space to improve the usage of these equations. If we better understand how the seismological equations behave and when is the best time to use them, probably this would lead to some improvement to the models. Also, we may want to try some other group of equations, once we focused on the Omori-Utsu formula and its relative equations.

We compared the models generated for each scenario. From the statistical analysis, we could analyse how the earthquakes variables influence the models and which combination result in models with higher log-likelihood values.

From the analysis, we discovered that more stable earthquake are easier to predict. These earthquakes have depth smaller or equal to 25km, magnitude higher than 4.0 in the Richter Scale. We also discovered that to classify earthquakes into mainshocks and aftershocks improve our methods predicting ability.

After considering all analysis, the method that achieve better results in the statistical analysis, was the *GAModelWindow*. We propose a comparison of this method models with the RI Algorithm and the GAModel to proof test this method.

References

[1] Claus Aranha, Yuri Cossich Lavinas, Marcelo Ladeira, e Bogdan 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. 3, 6, 7, 8, 16, 17, 18, 33, 40

[2] Ali Firat Cabalar e Abdulkadir Cevik. Genetic programming-based attenuation relationship: An application of recent earthquakes in turkey. *Computers and Geosciences*, 35:1884–1896, October 2009. 14

[3] Felipe Campelo. Lecture notes on design and analysis of experiments. `https://github.com/fcampelo/Design-and-Analysis-of-Experiments`, 2015. Version 2.11, Chapter 7; Creative Commons BY-NC-SA 4.0. 37

[4] François-Michel De Rainville, Félix-Antoine Fortin, Marc-André Gardner, Marc Parizeau, e Christian Gagné. Deap: A python framework for evolutionary algorithms. In *Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computation Conference Companion*, GECCO Companion '12, pages 85–92, New York, NY, USA, 2012. ACM. 19, 33

[5] David Eberhard. *Multiscale seismicity analysis and forecasting: examples from the western Pacific and Iceland*. Tese (Doutorado), 2014. 3, 6, 8, 9

[6] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989. 12

[7] A. M. Huda e Bagus Santosa. Subsurface structure in japan based on p and s waves travel time analysis using genetic algorithm in japan seismological network. *International Journal of Science and Engineering*, 6(1), 2014. 15

[8] T Serkan Irmak, Bülent Doğan, e Ahmet Karakaş. Source mechanism of the 23 october, 2011, van (turkey) earthquake (m w= 7.1) and aftershocks with its tectonic implications. *Earth, Planets and Space*, 64(11):991–1003, 2012. 2, 5

[9] B. L. N. Kennet e M. S. Sambridge. Earthquake location — genetic algorithms for teleseisms. *Physics of the Earth and Planetary Interiors*, 75(1–3):103–110, December 1992. 14

[10] Tienfuan Kerh, David Gunaratnam, e Yaling Chan. Neural computing with genetic algorithm in evaluating potentially hazardous metropolitan areas result from earthquake. *Neural Comput. Appl.*, 19(4):521–529, June 2010. 14

[11] Tienfuan Kerh, Yu-Hsiang Su, e Ayman Mosallam. Incorporating global search capability of a genetic algorithm into neural computing to model seismic records and soil test data. *Neural Computing and Applications*, pages 1–12, 2015. 14

[12] E. Kermani, Y. Jafarian, e 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. 14

[13] John R Koza, Martin A Keane, e Matthew J Streeter. Genetic programming's human-competitive results. *IEEE Intelligent Systems*, pages 25–31, 2003. 12

[14] John R. Koza, Martin A. Keane, e Matthew J. Streeter. What's ai done for me lately? genetic programming's human-competitive results. *IEEE Intelligent Systems*, 18(3):25–31, 2003. 2, 5

[15] Zbigniew Michalewicz. Heuristic methods for evolutionary computation techniques. *Journal of Heuristics*, 1(2):177–206, 1996. 2, 5

[16] D. Michie, D. J. Spiegelhalter, e C.C. Taylor. Machine learning, neural and statistical classification, 1994. 2, 5

[17] Nobuo Mimura, Kazuya Yasuhara, Seiki Kawagoe, Hiromune Yokoki, e So Kazama. Damage from the great east japan earthquake and tsunami-a quick report. *Mitigation and Adaptation Strategies for Global Change*, 16(7):803–818, 2011. 1, 4

[18] K. Z. Nanjo. Earthquake forecasts for the csep japan experiment based on the ri algorithm. *Earth Planets Space*, 63:261–274, 2011. 33

[19] Ahmad Nicknam, Reza Abbasnia, Yasser Eslamian, Mohsen Bozorgnasab, e Ehsan Adeli 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. 14

[20] The Institute of Statistical Mathematics. Package 'sapp'. `https://cran.r-project.org/web/packages/SAPP/SAPP.pdf`, June 2016. [Online; acessed: 27-07-2016]. 24

[21] Yosihiko Ogata e Jiancang Zhuang. Space–time etas models and an improved extension. *Tectonophysics*, 413(1):13–23, 2006. 23

[22] Fusakichi Omori. On the after-shocks of earthquakes. 1895. 23

[23] Josafath I. Espinosa Ramos e Roberto A. Vázques. Locating seismic-sense stations through genetic algorithms. In *Proceedings of the GECCO'11*, pages 941–948, Dublin, Ireland, July 2011. ACM. 15

[24] Negar Sadat, Soleimani Zakeri, e Saeid Pashazadeh. Application of neural network based on genetic algorithm in predicting magnitude of earthquake in north tabriz fault (nw iran). *Current Science (00113891)*, 109(9), 2015. 13

[25] A. Saegusa. Japan tries to understand quakes, not predict them. Nature 397, 284, 1999. 3, 6

[26] Bahram Saeidian, Mohammad Saadi Mesgari, e Mostafa Ghodousi. Evaluation and comparison of genetic algorithm and bees algorithm for location-allocation of earthquake relief centers. *International Journal of Disaster Risk Reduction*, 2016. 15

[27] D. Schorlemmer, M. Gerstenberger, S. Wiemer, D. Jackson, e D. A. Rhoades. Earthquake likelihood model testing. *Seismological Research Letters*, 78(1):17–29, 2007. 7, 8, 9, 10, 11

[28] Danijel Schorlemmer, J Douglas Zechar, Maximilian J Werner, Edward H Field, David D Jackson, Thomas H Jordan, e RELM Working Group. First results of the regional earthquake likelihood models experiment. *Pure and Applied Geophysics*, 167(8-9):859–876, 2010. 10, 11, 17

[29] Mark Simons, Sarah E Minson, Anthony Sladen, Francisco Ortega, Junle Jiang, Susan E Owen, Lingsen Meng, Jean-Paul Ampuero, Shengji Wei, Risheng Chu, et al. The 2011 magnitude 9.0 tohoku-oki earthquake: Mosaicking the megathrust from seconds to centuries. *science*, 332(6036):1421–1425, 2011. 1, 4

[30] Tokuji Utsu e Yosihiko 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. 23

[31] Thomas van Stiphout, Jiancang Zhuang, e David Marsan. Seismicity declustering. *Community Online Resource for Statistical Seismicity Analysis*, 10, 2012. 1, 4, 30

[32] S Wilkinson, G Chiaro, Rama Mohan Pokhrel, T Kiyota, Toshihiko Katagiri, Keshab Sharma, e K Goda. The 2015 gorkha nepal earthquake: Insights from earthquake damage survey. 2015. 1, 4

[33] Yoshiko Yamanaka e Kunihiko 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. 23, 29

[34] J Douglas 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. 8, 17

[35] Qiuen Zhang e Cheng 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. 13

[36] Feiyan Zhou e Xiaofeng Zhu. Earthquake prediction based on lm-bp neural network. In Xiaozhu Liu e Yunyue Ye, editors, *Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 1*, volume 270 of *Lecture Notes in Electrical Engineering*, pages 13–20. Springer Berlin Heidelberg, 2014. 13

[37] Jiancang Zhuang, Yosihiko Ogata, e David Vere-Jones. Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research: Solid Earth*, 109(B5), 2004. 17, 23, 26