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

Yuri Lavinas[1], Claus Aranha[2], Xiucai Ye[3],
Marcelo Ladeira[4], and Tetsuya Sakurai[5]

[1] University of Brasilia, Computer Science Department yclavinas@gmail.com
[2] University of Tsukuba, Graduate School of SIE caranha@cs.tsukuba.ac.jp
[3] University of Tsukuba, Graduate School of SIE yexiucai@mma.cs.tsukuba.ac.jp
[4] University of Brasilia, Computer Science Department mladeira@unb.br
University of Tsukuba, Graduate School of SIE sakurai@cs.tsukuba.ac.jp

Abstract. Earthquake Risk Models describe the risk of occurrence of seismic events on an area based on information such as past earthquakes in nearby regions and the seismic properties of the area under study. These models can be used to help to better understand earthquakes, their patterns and their mechanisms.
In previous work, we showed that Genetic Algorithms can be used to generate earthquake risk models. However, we also noticed some shortcomings in that approach. We propose three improvements to address these shortcomings: A new representation that reduces the search space, a hibridisation process that uses seismic equations to generate the risk model from the GA representation using domain knowledge, and a preprocessing step for the training data using clustering.
We examine each of these changes through simulations using the catalog of Japanese earthquakes between 2000 and 2010, and indicate the contribution of each proposal to the quality of generated models.

1 Introduction

Earthquakes can cause great damage to human society through soil rupture, movement, tsunami, etc. Some recent earthquakes that highlight this destructive potential are the great East Japan Earthquake of 2011 (Figure 1), and the April 2015 earthquake in Nepal. One important tool for the enactment of policies that minimise the consequences of these events are earthquake occurrence models (also called risk models). These models can be used to identify patterns in the seismic mechanisms that generate earthquakes, and are important to increase our understanding of these events.

In our previous work [1], we proposed a way to generate earthquake risk models using a standard Genetic Algorithm (here called the GAModel). The GA model was shown to be competitive with the Relative Intensity (RI) model, while not using any a-priori information about the distribution of earthquake

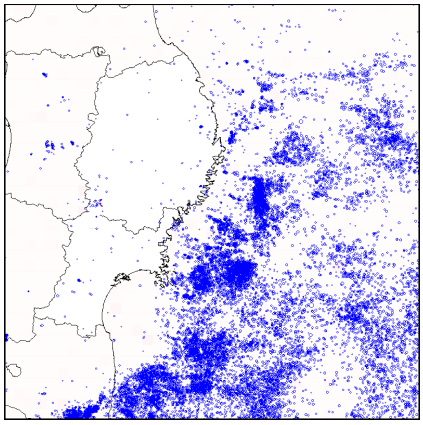

Fig. 1. Seismic Activity in Eastern Japan in 2011. Each blue represents one earthquake

occurrences. In this paper, we identify three key issues with the GAModel, and propose adjustments to the algorithms that address these issues.

The first issue is that the genome representation used by GAModel has too many parameters (over 2000 for regular cases), leading to an unnecessarily large search space. To address this issue, we propose a new genome representation for an earthquake risk model, the *Reduced Representation*. In this representation only those locations with a minimal probability of earthquake are represented as parameters in the evolutionary process.

The second issue is that GAModel does not take into account any sort of domain knowledge, such as the assumption that earthquakes cluster in both time and space. Heuristic search methods usually benefit from the introduction of domain knowledge to the search. Therefore, we propose a hybrid version of the GAModel which incorporates seismic models of earthquake decay. In a two step process, the Genetic Algorithm first generates a set of mainshocks, then uses an adaptation of the Epidemic Type Aftershock Sequence (ETAS) to generate the aftershocks.

The third issue is the examination of "de-clustering" effects in the historical catalog used for generating the risk model. In seismology, de-clustering refers to the act of identifying earthquakes as either main-shocks or aftershocks, and removing all but the main shocks from the catalog, which is considered the representative earthquake for the group. Accordingly, a de-clustered earthquake catalog is considered to be easier to study, given that the de-clustering process removes redundant information [14]. In this work, we generate the de-clustered catalog by grouping earthquakes in space and time using spectral clustering [5].

These adaptions are described in detail in section 3. We compare the contributions of each adaptation to the generation of models based on the earthquake catalog of the Japanese archipelago, between 2000 and 2010. The set up of this experiment is described in section 4, and the main results are listed in section 5. Our results indicate that the clustering pre-processing step gives the biggest con-

tribution to the model precision, and that the new representation reduces the search space without negatively affecting the model results.

2 Background

An Earthquake Risk model states the probability of earthquake occurrence on a defined area and time period. These models are often based on past occurrence of earthquakes (historical catalogs). They can also make use of seismic properties of the area under study, such as faults, terrain properties, etc.

The "prediction" of earthquakes is a polemic subject, and no research so far has come close to suggesting that individual large scale earthquakes can be predicted. On the other hand, there is value on the study of earthquake mechanisms and the generation of statistical models of earthquake risk [11].

In our previous work [1], we use a Genetic Algorithm (GA) to optimise an Earthquake Risk Model, which is described in the framework proposed by the Collaboratory for the Study of Earthquake Predictability (CSEP).

The CSEP framework defines a model in reference to a geographical region and a time period [18]. The geographical region is divided in a grid, where each cell in the grid is called a bin.

The model defines a number of expected earthquakes for each bin. This number must be a positive integer. A good model is one where the number of estimated earthquakes in each bin corresponds to the actual number of earthquakes that occurs in that bin during the target time interval.

2.1 The GAModel

Using the CSEP framework described in the previous subsection, we proposed the GAModel [1], which uses Genetic Algorithms to generate an earthquake risk model based on earthquake catalog data.

In the GAModel, each individual is treated as a prediction in the CSEP framework. The fitness of each individual will be calculated using the log-likelihood of the catalog data given the individual's prediction.

Genome Representation and Evolutionary Operators. Each individual is represented as real valued array, where each element is a bin, with an associated number of earthquakes. One-point crossover, elitism and polynomial bounded mutation are used as evolutionary operators. The relevant parameters were set as Elite Size = 1, Crossover chance = 0.9, Mutation Chance = 0.1, Polynomial Bounded parameters eta = 1, low = 0, up = 1.

Fitness Function and Selection. Let an individual $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ be a forecast in the CSEP framework. Let the set of earthquake occurrences from the catalog be $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$. The log-likelihood of the catalog data given an individual is calculated as:

$$L(\Omega|\Lambda) = \sum_{i=1}^{n} L(\omega_i|\lambda_i) = \sum_{i=1}^{n} -\lambda_i + \omega_i \log \lambda_i - \log \omega_i!. \quad (1)$$

To avoid overfitting, the period under consideration is divided into sub-periods, the log-likelihood for each sub-period is calculated separately, and the worst value is used as the fitness [1].

2.2 Related Literature

The usage of Evolutionary Computation (EC) in the field of earthquake risk models is somewhat sporadic. Zhang and Wang [19] used Genetic Algorithms to fine tune an Artificial Neural Network (ANN) and used this system to produce a forecast model. Zhou and Zu [20] also proposed a combination of ANN and EC, but their system forecasts only the magnitude parameter of earthquakes. Sadat, in [10], used ANN and GA to predict the magnitude of the earthquakes in North Iran.

There are more works when we discuss EC methods and estimation of parameter values in seismological models. Nicknam et al. [6] simulated some components from a seismogram station and predicted seismograms for other stations. They combined the Empirical Green's Function (EGF) with GA. Kennett and Sambridge [3] used GA and associated teleseisms procedures to determine the Fault Model parameters of an earthquake.

Another popular approach is to use EC methods do calculate the Peak Ground Acceleration (PGA) parameter. The works done by Kerh et al. [4, ?] are a combination of ANN and GA to estimate or predict PGA in Taiwan. Cabalar and Cevik [2] work also aimed to predict the PGA, but their work uses genetic programming (GP) and use strong-ground-motion data from Turkey.

Ramos and Vázques [9] 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.

3 Proposed Changes

In this work, we propose three improvements to the GAModel: A reduced genome representation, Hybridisation with the ETAS empirical model, and the clustering of the earthquake catalog. Each of these changes are described below.

3.1 Reduced Genome Representation

In the GAModel, problem is represented as a vector X where each bin corresponds to an element in the vector. As the number of bins in a region numbers into the thousands, this representation leads to a huge search space to be explored.

We have observed that in many cases, the vector of catalog earthquakes is sparse. To use this fact to decrease the search space, we propose a "reduced" representation of a risk Model.

The reduced representation is a vector V of ordered pairs. The first element of this pair is the integer index that identify a bin in the model. The second element of the pair is the number of earthquake occurrences estimated for this bin.

The size of the vector V is calculated as the number of bins in the historical catalog that contain at least one earthquake. For each element in V, the bin index and the estimated number of occurrences are drawn randomly from a uniform distribution.

To generate a model from the reduced representation, we need to go two intermediate steps. The first one is to transform the reduced representation into a regular representation. This is achieved by copying the estimated value of an element to the bin indicated by stored index for that element. Bins that are not indicated by any element in the vector are set to zero estimated earthquakes. The second step is to apply the inverse Poisson on the estimated values to retrieve the number of earthquakes.

Using the reduced representation requires us to change the mutation operator. The mutation operator selects one element in the vector and draw new values for the index and estimation parameter from a uniform distribution.

3.2 Hybridisation with ETAS

The GAModel produces risk models without using any sort of domain knowledge, other than the difference between the individual being evaluated and the earthquake catalog data.

However, one simple observation that could be added to the GAModel is that earthquakes cluster in space and time. Large earthquakes are usually followed by a wave of smaller earthquakes, these pairings being commonly known as *mainshocks* and *aftershocks* [12].

Therefore, we introduce this idea by modifying the proccess which generates a model from an individual into a two step procedure, named *EMP-GA*.

In the first step, we use the GAModel to generate a set of mainshock earthquakes, which we will refer to as *synthetic mainshock data*. In the second step, we use seismic empirical equations to obtain the aftershocks from the synthetic mainshock data, and add them to the model.

The total number of earthquakes in a bin is given as

$$\Lambda(t, x, y|\Upsilon_t) = [\mu(x, y) + \sum_{t_i \in t} K(M_i)g(t - t_i)P(x, y)]J(M). \tag{2}$$

In this equation, t is the target time for the model, x, y are latitude/longitude coordinates within the target area, Υ_t is the number of mainshocks derived in the first step, $\mu(x, y)$ is the expected number of earthquakes at the (x, y) bin, t_i is a time interval within t, $K(M_i)$ is the total amount of triggered events, $g(\Delta t)$

is the probability density form of the modified Omori law, $P(x, y)$ is a function that distributes aftershocks in space nearby the mainshock, and $J(M)$ is the ETAS simulated magnitude. Let us explain each of these components below.

Omori's Law and Triggered Events The Omori law, which is considered to be an empirical seismic formula which has withstood the test of time [15, 8], is a power law that relates the magnitude of an earthquake with the decay of aftershock activity over time. It can estimate the number of aftershocks based on the mainshocks in the synthetic data generated by one individual in the EMPGA. For this approach, we use the probability density function (PDF) form of the modified Omori law [21], defined as

$$g(\Delta t) = \frac{(p-1)}{c(1+\frac{t}{c})^{-p}}. \tag{3}$$

In this equation, p and c are constants. Utsu [15], summarise the studies of this formula for the Japan case, and describe a range for these variables using the Davidon-Fletcher-Powell optimisation procedure. These ranges, used in ETAS, are 0.9 to 1.4 for p, and 0.003 and 0.3 for c.

Also, Δt is the time interval for how long a mainshock may influence or cause an aftershock. A value too short will lead to a small number of aftershocks, while a value too long might confound aftershocks and background activity. Previous work suggests the values $p = 1.3$, $c = 0.003$, and $\Delta t = 30$ days [16].

The total amount of events triggered by a mainshock is represented in equation 2 as $K(M_i)$. To calculate this value, we count the number of aftershocks within a given area A from the mainshock, using the formula

$$K(M_i) = A\ exp([\alpha(M_i - M_c)]). \tag{4}$$

Where $M_c = 3.0$ is the magnitude threshold and $\alpha(M)$ is defined as the inverse of the magnitude, according to Ogata [7]. The area A is obtained using the equation from Yamanaka [16]

$$A = e^{(1.02M-4)}. \tag{5}$$

Using $K(Mi)$ and $g(t)$, it is possible to calculate the total number of earthquakes generated from a mainshock, by iterating over t_i:

$$\sum_{t_i \in t} K(M_i) g(t - t_i) \tag{6}$$

The resulting aftershocks need to be spread on bins near the mainshock position. The $P(x, y)$ component of equation 2 calculates the position of the aftershocks based on the position of the original mainshock. It simply places each aftershock either north, south, east or west of the mainshock, getting further from the origin after each iteration, until there are no more events to be placed.

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

3.3 Clustering the Catalog Data

The third adaptation to the GAModel is the pre-processing of the earthquake catalog data using Spectral Clustering. This pre-processing step aims to remove redundant information from the catalog, by clustering together earthquakes which are closely related in a mainshock/aftershock relationship [14]. Because it is difficult to determine exactly when two earthquakes should be clustered together, we choose a non-supervised method to generate the clusters.

Spectral Clustering involves constructing a similarity matrix of the elements to be clustered, finding the k-Nearest-Neighbours graph (KNN) based on the similarity matrix, calculating the Laplacian matrix of the KNN graph, and performing k-means clustering on the eigenvectors of this matrix.

One of the main characteristics of Spectral Clustering that make it interesting for this problem is that it can be very computationally efficient [17]. This is very important for the clustering of earthquake data, since each data set can contain tens of thousands of earthquakes.

Spectral Clustering Implementation Let the earthquake catalog data be represented as a vector $X = \{x_1, x_2, \ldots, x_n \; in \Re_d\}$, where n is the number of earthquakes in the catalog, d is the number of attributes that characterise an earthquake in the catalog, and K is the desired number of clusters. The clusters are calculated following algorithm 1

Algorithm 1 Spectral Clustering

Construct the similarity matrix S
for i,j in X **do**
  IF i and j are connected in the KNN graph THEN $s_{i,j} = \exp(-||x_i - x_j||^2/2\sigma^2$
  ELSE $s_{i,j} = 0$
end for
Matrix $D = nxn$ diagonal matrix where $d_{i,i} = \sum_{j=1}^{n} s_{ij}$
Compute Matrix $L = D - S$ and the K smallest eigenvectors of L
Compute matrix $V = (v_{ij})_{nxK}$, using these eigenvactors as columns.
Compute matrix $U = (u_{ij})_{nxK}$, normalising the rows of V such as $u_{i,j} = v_{i,j}/\sqrt{\sum_j v_{ij}^2}$
Let each row in U represent a data point, and cluster these points using k-means
FOR each point x_i in X DO Assign the cluster of u_i to x_i

In this algorithm, $||x_i - x_j||$ is the Euclidean distance between data points x_i, x_j. We use the number of nearest neighbours equal to five, and σ, the kernel parameter, equal to 100.

We cluster the earthquakes based on their latitude, longitude, time (in minutes), and depth. By observing the distributions of the eigenvectors, we defined the weight of each dimension in the algorithm: *latitude and longitude: 150, time: 7, Depth: 0.5*

4 Experiment Design

In this paper we propose three improvements for the GAModel algorithm that generates earthquake risk models using GA: A reduced representation that limits the search space of the algorithm, a hybrid model generation that uses domain knowledge, and the pre-processing of the data using Spectral Clustering.

We are interested in determine what effect these improvements have on the generation of earthquake risk models. To achieve this, we execute a series of simulation experiments. In these experiments, we generate earthquake models for each combination of the above modifications, using historical earthquake catalog data from the Japanese archipelago.

4.1 Experiment Design

Our experiment has a factorial design with three factors: Using the reduced representation, Using the hybrid model, and clustering the data set. For each combinations of these factors, we generate 10 models for two target regions and 6 five-year periods, for a total of 120 models per combination.

We use ANOVA to test whether any of the combinations shows a significant deviation in terms of model accuracy, represented as the log-likelihood between the model and the catalog data. If this is indicated, we compare each combination with the original algorithm using Tukey HSD. We repeat this procedure for each of the two areas separately as well. In each of these tests, we set $\alpha = 0.05$.

4.2 Data Sets

We use the earthquake catalog made available by the *Japan Meteorological Agency* (JMA) webpage. From this catalog, we use the following earthquake data: time of occurrence, magnitude, latitude, longitude and epicentre depth.

From this catalog, we focus on two areas for our study.

Kanto is the region around metropolitan Tokyo. In this work we define it as the area within latitude 34.8N to 37N and longitude 138.8W to 141W. It is divided into 2025 bins of approximately 25km^2.

East Japan region covers the east coast of Japan, including a large ocean area. In this work we define it as the area within latitude 37N to 41N, and 140W to 144W. It is divided into 1600 bins of approximately 100km^2.

We use earthquakes from the JMA catalog that happen between 2000 and 2010 [5]. This is divided in 11 five-year overlapping periods (2000–2005, 2001–2006, and so on), which are identified by the last year in the period.

For each period, we filter earthquakes from the catalog using the following rules:

[5] We deliberately avoid using the 2011 catalog, as the occurrence of the Great East Japan earthquake caused an anomalous number of aftershocks, more than the past five years added together. We are interested in adding this event to our analyses in the future

– Depth of the earthquake must be under below 100km, as shallow earthquakes are considered to be more independent and easier to analyse [16].
– Magnitude of the earthquake must be above 3. While the catalog lists some earthquakes with magnitude under 3, often such earthquakes escape detection, and thus introduce incompleteness to the catalog.

5 Results

Table 1 summarizes the results of the experiments comparing the eight combinations over the different scenarios.

Table 1. Log Likelihood Values for each scenario and combination. Higher values correspond to better models. Values in bold are the two highest log likelihood for a scenario. Each value is the average of 10 runs.

	No Pre-processing				Spectral Clustering			
Scenario	GA	Red	EMP	EMP-Red	GA	Red	EMP	EMP-Red
Kanto 2005	-2291.4	-2354.1	-2345.1	-2557.8	**-2202.7**	-2233.3	**-2203.7**	-2355.0
Kanto 2006	-2269.2	-2317.3	-2350.7	-2520.0	**-2173.3**	-2203.1	**-2175.8**	-2313.5
Kanto 2007	-2204.9	-2235.7	-2293.3	-2449.5	**-2104.1**	-2125.0	**-2110.3**	-2213.9
Kanto 2008	-2203.0	-2273.2	-2277.7	-2501.0	**-2097.9**	-2124.3	**-2010.3**	-2245.7
Kanto 2009	-2375.6	-2418.5	-2463.6	-2630.7	**-2279.1**	-2299.1	**-2282.4**	-2382.0
Kanto 2010	-2203.6	-2296.6	-2294.9	-2534.4	**-2099.5**	-2125.0	**-2104.0**	-2249.8
EastJapan 2005	-2442.8	-2394.9	-2633.6	-2588.4	**-2099.6**	**-2150.2**	-2177.4	-2300.6
EastJapan 2006	-2211.1	-2191.7	-2408.9	-2390.9	**-1896.7**	**-1960.4**	-1965.7	-2131.8
EastJapan 2007	-2112.2	-2100.5	-2305.1	-2294.9	**-1821.9**	**-1889.4**	-1914.4	-2070.0
EastJapan 2008	-4139.7	-4288.6	-4301.3	-4424.8	**-3942.5**	**-3989.1**	-4034.9	-4156.8
EastJapan 2009	-2281.2	-2221.2	-2498.9	-2416.5	**-1948.5**	**-1087.4**	-2043.7	-2164.5
EastJapan 2010	-2577.7	-2579.1	-2783.9	-2783.9	**-2232.7**	**-2291.3**	-2296.9	-2455.2

To better understand these results, we perform an ANOVA analysis...

* Anova in all areas

Because the anova indicated a significant difference, we use Tukey's HSD to see which combination showed this difference ...

* HSD Tukey against GAModel

To get a better intuition about what these results mean in concrete terms, we show a selection of the actual models... (explain heat map)

* Heat Maps

6 Discussion

Clustering is good. It is okay if Reduced GA is not worse, because "lower search space"

7 Conclusion

We did many choices of the target regions and datas that made the models easier to analyse. In future work, we would like to see the results of removing these constraints from the data set, and how to deal with this harder problem.

Acknowledgments. The authors would like to thank the Japan Meteorological Agency for making available the earthquake catalog used in this study.

References

1. Aranha, C., Lavinas, Y.C., Ladeira, M., Enescu, B.: Is it possible to generate good earthquake risk models using genetic algorithms? In: Proceedings of the International Conference on Evolutionary Computation Theory and Applications. pp. 49–58 (2014)
2. Cabalar, A.F., Cevik, A.: Genetic programming-based attenuation relationship: An application of recent earthquakes in turkey. Computers and Geosciences 35, 1884–1896 (October 2009)
3. Kennet, B.L.N., Sambridge, M.S.: Earthquake location — genetic algorithms for teleseisms. Physics of the Earth and Planetary Interiors 75(1–3), 103–110 (December 1992)
4. Kerh, T., Gunaratnam, D., Chan, Y.: Neural computing with genetic algorithm in evaluating potentially hazardous metropolitan areas result from earthquake. Neural Comput. Appl. 19(4), 521–529 (Jun 2010), http://dx.doi.org/10.1007/s00521-009-0301-z
5. Luxburg, U.V.: A tutorial on spectral clustering. Statistics Computing 17(4), 395–416 (2007)
6. Nicknam, A., Abbasnia, R., Eslamian, Y., Bozorgnasab, M., Mosabbeb, E.A.: 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)
7. Ogata, Y., Zhuang, J.: Space–time etas models and an improved extension. Tectonophysics 413(1), 13–23 (2006)
8. Omori, F.: On the after-shocks of earthquakes (1895)
9. Ramos, J.I.E., Vázques, R.A.: Locating seismic-sense stations through genetic algorithms. In: Proceedings of the GECCO'11. pp. 941–948. ACM, Dublin, Ireland (July 2011)
10. Sadat, N., Zakeri, S., Pashazadeh, S.: 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)
11. Saegusa, A.: Japan tries to understand quakes, not predict them. Nature 397, 284 (1999)
12. Schorlemmer, D., Zechar, J.D., Werner, M.J., Field, E.H., Jackson, D.D., Jordan, T.H., Group, R.W.: First results of the regional earthquake likelihood models experiment. Pure and Applied Geophysics 167(8-9), 859–876 (2010)
13. of Statistical Mathematics, T.I.: Package 'sapp'. https://cran.r-project.org/web/packages/SAPP/SAPP.pdf (Jun 2016), [Online; acessed: 27-07-2016]

14. van Stiphout, T., Zhuang, J., Marsan, D.: Seismicity declustering. Community Online Resource for Statistical Seismicity Analysis 10 (2012)
15. Utsu, T., Ogata, Y.: The centenary of the omori formula for a decay law of aftershock activity. Journal of Physics of the Earth 43(1), 1–33 (1995)
16. Yamanaka, Y., Shimazaki, K.: 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)
17. Ye, X., Sakurai, T.: Robust similarity measure for spectral clustering based on shared neighbors. ETRI Journal 38(3), 540–550 (2016)
18. Zechar, J.D.: Evaluating earthquake predictions and earthquake forecasts: A guide for students and new researchers. Community Online Resource for Statistical Seismicity Analysis pp. 1–26 (2010)
19. Zhang, Q., Wang, C.: Using genetic algorithms to optimize artificial neural network: a case study on earthquake prediction. In: Second International Conference on Genetic and Evolutionary Computing. pp. 128–131. IEEE (2012)
20. Zhou, F., Zhu, X.: Earthquake prediction based on lm-bp neural network. In: Liu, X., Ye, Y. (eds.) Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 1, Lecture Notes in Electrical Engineering, vol. 270, pp. 13–20. Springer Berlin Heidelberg (2014), http://dx.doi.org/10.1007/978-3-642-40618-8_2
21. Zhuang, J., Ogata, Y., Vere-Jones, D.: Analyzing earthquake clustering features by using stochastic reconstruction. Journal of Geophysical Research: Solid Earth 109(B5) (2004)