Integreating geological and seismological data in point process models for seismical analysis

Marianna Siino, Giada Adelfio
Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli studi di Palermo, Palermo, Italy

April 11, 2017

Abstract

Nowadays in the seismic and geological fields, large and complex data sets are available. This information is a valuable source that can be used for improving the seismic hazard assessment of a given region. In particular, the integration of geologic variables into point process models to study seismic pattern is an open research field that has not been fully explored. In this work, we present several open-access datasets (the catalog of the earthquakes, geological information such as faults, plate boundary and the presence of volcanoes) that are properly treated to describe the seismicity of events occurred in Greece between 2005 and 2014. We use these datasets to fit an advanced spatial point process model for the description of interaction among the points in the presence of larger-scale inhomogeneity.

Keywords: earthquakes; faults; point process; Gibbs model

1 Introduction

Earthquakes are the most unpredictable natural disaster and the level of damages mainly depends on the earthquake magnitude, depth, epicentral distance and on geology in the neighborhood area, which can generate local amplification. An earthquake is a sudden movement of the earth lithosphere, and the two main causes of earthquakes are: volcanic activity and tectonic activity (in this case, events occur along the boundaries of major tectonic plates and active faults).

Commonly, in an observed area, earthquakes can be considered as a realization of a marked space-time point process, where the magnitude is the mark, and a point is identified by its geographical coordinates and time of occurrence (Illian et al.; 2008). The main type of interaction structure in earthquake data is clustering. As a matter of fact, the concentration of earthquakes in space is observed in the neighbourhood of possible sources, such as volcanoes, faults and plate boundaries. Instead, clustering in time can be seen as a significant increase of seismic activity immediately after large earthquakes. In the literature, the model formulation is usually based on self-exiting point processes, such as the ETAS model (Ogata; 1988). Traditionally, there are not used further information than the seismic catalog data and so the analysis is purely based on the distribution of events. When covariate data are also available, it is attractive and motivating the definition and estimation of models to investigate whether the intensity depends on the covariates and to quantify this dependence (Baddeley et al.; 2015). In particular, geologic information, such as the distance from plate boundaries or the nearest fault, can be used as covariates to provide a better estimation of the spatial intensity and to show a correlation between seismic events and geologic data.

In Section 2, we present the seismic catalog and the geological data sets available in the Greek area explaining how the information have been collected, processed and cleaned for the analysis. After some descriptive analysis, the several datasets have been integreated to show a possible usage in point process model. A hybrid of Gibbs point process models is estimated (Baddeley et al.; 2013) to describe the spatial distribution of earthquakes (with a magnitude 4) from 2005 to 2014, accounting for multiscale dependence while also including the effect of the geological covariates (Siino et al.; 2016). In Section 3, the methodological approach is explained, and some results are presented in Section 4. Conclusive remarks will follow.

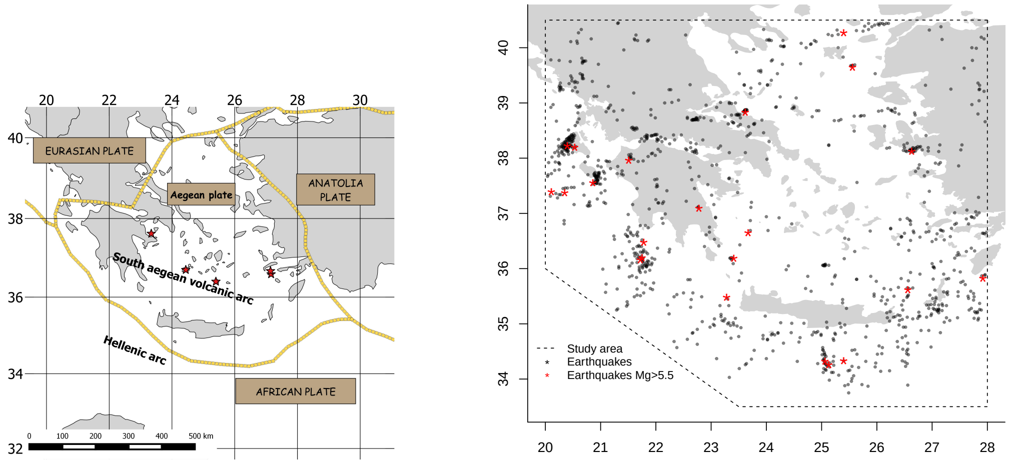
2 The study area and data description

The Hellenic area is the most seismically active area of the European-Mediterranean region having experienced many destructive earthquakes and it is characterised by both tectonic and volcanic seismogenic sources. The area is located at the boundary of the Africa-Eurasia convergence, Figure 1a. The compressional motion between the African and Eurasian plates causes the subduction of the lithosphere forming the Hellenic Arc. About 150 km to the north in the southern Aegean Sea, the Hellenic volcanic arc is located. Most of the volcanoes in the area are not active in terms of eruptions, but nevertheless they are sources of microseismic activity. Moreover, Greece and its surroundings present local active faults.

2.1 The earthquake catalog

A seismic catalog contains focal parameters of earthquakes (latitude, longitude and depth), event magnitude and time of occurrence. Several earthquake catalogs are freely accessible and the data set used in this work comes from the Hellenic Unified Seismological Network, (HUSN, http://www.gein.noa.gr/en/seismicity/earthquake-catalogs) that is nowadays constituted by about 150 seismic stations adequately distributed over the area. The Greek catalog contains events since 1964 and we considered in our analysis the seismic catalogue since 2005, when the network was upgraded and earthquake location was sensibly improved. The study area extends from 33.5 to 40.5 Lat. N and from 20 to 28 Long. E (Figure 1a).

The quality of a catalog is foundamental, since it influences the quality of the seismicity analysis. An earthquake should be considered as a volume rather than a single point, however it is represented by the focal parameters. The accuracy and precision of focal parameters mainly depends on the magnitude and on the number and distribution of the seismic stations. Small magnitude earthquakes are generally recorded from few stations (only the nearest ones) and consequently, focal parameters are not estimated reliably. So earthquake location is affected by errors,



(a) (b)

Figure 1: (a) Tectonic plates in the Hellenic region and red stars indicate the main volcanoes. (b)Study area (W) rapresented by a polygonal window, points are earthquakes occured between 2005 and 2014 with a amgnitude greater than 4 and the red ones are those with a magnitude greater than 5.5.

but we can assume that all the errors associated to all the single events average out each other when they are treated together. A crucial challenge is the evaluation of the catalog completeness, that is defined as the lowest magnitude for which 100% of the earthquakes in a space-time volume are detected. For our subset, it is estimated at 2.4 using the several procedures proposed in Mignan and Woessner (2012). However, the magnitude threshold is setted to 4 since we aim to study the spatial intensity of the main events occurred in the area excluding the micro and the minor events according to a qualitative classification of the magnitude.

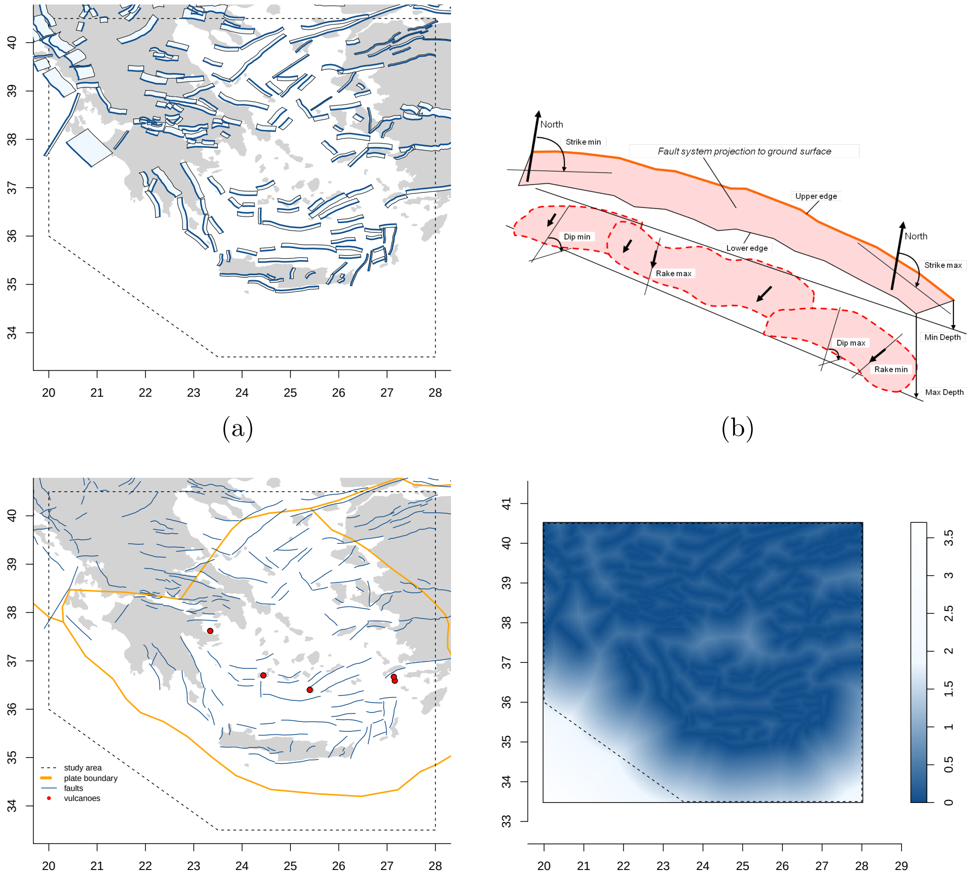
The events in the study window with a magnitude greater or equal than 4 between 2005 and 2014 are 1105. Each seismic event will be considered as a point $\mathbf{u}_i$ in an observed spatial point pattern $\boldsymbol{\upsilon}$ individuated by its two spatial coordinates (latitude and longitude in WGS84 coordinate system). In R, the essential components of a point pattern object $\boldsymbol{\upsilon} = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ are the coordinates of the points and the observation window W that in our case is a polygonal window (see Figure 1b). The main assumption for the analysis of a point patterns are: point locations are measured exactly, there are not multiple points, points are mapped without omission, there are no errors in detecting the presence of points and points could have been observed at any location in the region W (Baddeley et al.; 2015).

Since the goal of the analysis is to describe in a proper way the spatial interactions between points, neglecting time, in the spatial point pattern $\boldsymbol{\upsilon}$ we found some duplication of points. How to deal with coincident points depends on the goals of the analysis, however when the data have replicated points, some statistical procedures can be severely affected. In our case, we do not discard this information and we randomly shift each coincident point by a small random distance in a random direction. For other procedures to deal with this type of data see Baddeley et al. (2015).

2.2 Geologica information

We additionally considered GIS-based open-access geological information to understand dependence between events and the different sources of earthquakes. In Figure 2c, the orange line is the Aegean plate boundary coming from a digitalised global set of present plate boundaries on the Earth (Bird; 2003). The dataset (downloaded from https://github.com/fraxen/tectonicplates) has vector information that represent the fractures of the Earth crust and additional meta-data about the boundaries.

The Global Volcanism Program database gives coordinates and describes the physical characteristics of Holocene volcanoes and their eruptions (http://volcano.si.edu). Moving from the West to the East, the main volcanoes of the Hellenic Volcanic Arc are Methana, Milos, Santorini, Yali and Nisyros (Siebert and Simkin; 2014) and they are represented by points corresponding to the main volcanic craters, see Figure 2c.



(a) (b)

(c) (d)

Figure 2: (a) Composite seismogenetic sources (b) Main geometric (strike, dip, width, depth) and kinematic (rake) parameters that charcaterised a composite source. (c) Faults (upper edges of the composite sources), plate boundary and main volcanoes in the Hellenic area. (d) Immage plot of the spatial covariate distance to the nearest fault ($D_f(\mathbf{u})$).

Furthermore, we used the Greek Database of Seismogenic Sources (GreDaSS) that concerns tectonic and active-fault data in Greece and its surroundings (Caputo et al.; 2013). The database consists of several layers (graphical and metadata) on seismogenic sources that are categorised into two types: individual and composite. In our analysis, we considered the spatial shape file of the Composite Seismogenic Sources (CSS) since they are much more appropriate for investigating large-scale processes, Figure 2a. A composite source represents a complex fault system with an unspecified number of aligned individual seismogenic sources that cannot be separated spatially. The database provides geometric (strike, dip, width, depth) and kinematic (rake) parameters and descriptive information (e.g. comments, latest earthquakes) associated to each source, Figure 2b. In particular for the analysis of the point process model, we consider the upper edges of the composite sources that for the sake of simplicity are named faults, blue lines in Figure 2c.

We transform all the previous geological information into spatial variables, $Z(\mathbf{u})$ defined at all locations $\mathbf{u} \in W$ (Baddeley et al.; 2015), $D_f(\mathbf{u})$ distance to the nearest fault, $D_v(\mathbf{u})$ distance to the nearest volcano and $D_{pb}(\mathbf{u})$ distance to the plate boundary. These spatial variables are treated as covariates since the research questions are whether the the intensity depends on the geological information, and whether, after accounting for the influence of geological data, there is evidence of spatial clustering of the earthquakes. Giving an example, the spatial data about the faults in Figure 2c are converted into a pixel images, and the pixel value represents the distance from that pixel to the nearest fault, and Figure 2d represents $D_f(\mathbf{u})$.

3 Methodology

A spatial point pattern $\boldsymbol{\upsilon} = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ is an unordered set of points in the region $W \subset \mathbb{R}^2$ where $n(\boldsymbol{\upsilon}) = n$ is the number of points, $|W| < \infty$ (Baddeley et al.; 2015). A point process model assumes that $\boldsymbol{\upsilon}$ is a realisation of a finite point process $\mathcal{X}$ in W without multiple points. The first-order property of $\mathcal{X}$ is described by the intensity function, $\rho(\mathbf{u}) = \lim_{|d\mathbf{u}|\to 0} \mathbb{E}(Y(d\mathbf{u}))/|d\mathbf{u}|$ where $d\mathbf{u}$ is an infinitesimal region that contains the point $\mathbf{u} \in W$, $|d\mathbf{u}|$ is its area and $\mathbb{E}(Y(d\mathbf{u}))$ denotes the expected number of events in $d\mathbf{u}$. When the intensity is constant the process is called homogeneous. In our case, we aim to describe the spatial arrangement as a function of environmental information using the spatial variable $Z(\mathbf{u})$ defined is Section 2.2.

Several functional summary statistics are used to study the second-order characteristics of a point pattern and so measuring dependence between events. Two widely used summary statistics are the Ripley's K-function and the G-function.

The G-function is the cumulative distribution function of the nearest-neighbour distance at a typical point of $\mathcal{X}$. It is a function of $d_i = d(\mathbf{u}_i, \boldsymbol{\upsilon}\backslash\mathbf{u}_i)$ that indicates the shortest distance from $\mathbf{u}_i$ to the pattern $\boldsymbol{\upsilon}\backslash\mathbf{u}_i$ consisting of all points of $\boldsymbol{\upsilon}$ except $\mathbf{u}_i$. It is defined as $G(r) = \mathbb{P}\{d(\mathbf{u}, \mathcal{X}\backslash\mathbf{u} \leq r)/\mathcal{X}$ has a point in $\mathbf{u}\}$ for any distance $r \geq 0$ and any location $\mathbf{u}$. Under the homogeneous Poisson assumption this relation holds $G(r) = 1 - \exp(-\rho\pi r^2)$. Visual inspection of the G-function with the theoretical summary statistic of the homogeneous Poisson process is used to study correlation in a descriptive analysis of the data.

3.1 Gibbs and Hybrid of Gibbs point process models

The class of Gibbs processes $\mathcal{X}$ is determined through a probability density function $f : \mathcal{X} \to [0, \infty)$, where $\mathcal{X} = \{\boldsymbol{\upsilon} \subset W : n(\boldsymbol{\upsilon}) < \infty\}$ is a set of point configurations contained in W. In the literature several Gibbs models have been proposed such as the area-interaction, Strauss, Geyer, hard core processes. However Gibbs processes have some drawbacks when points have a strong clustering and show spatial dependence at multiple scales (Illian et al.; 2008; Baddeley et al.; 2015). Baddeley et al. (2013) proposes hybrid models as a general way to generate multi-scale processes combining Gibbs processes. Given m unnormalized densities $f_1, f_2, \dots, f_m$, the hybrid density is defined as $f(\boldsymbol{v}) = f_1(\boldsymbol{v}) \times \dots \times f_m(\boldsymbol{v})$, where the components have to respect some properties (Baddeley et al.; 2013).

For example the density of the stationary hybrid process obtained considering m Geyer components (with interaction ranges $r_1, \dots r_m$ and saturation parameters $s_1, \dots, s_m$) is

$$f(\boldsymbol{v}) = \beta^{n(\boldsymbol{v})} \prod_{i=1}^{n(\boldsymbol{v})} \prod_{j=1}^{m} \gamma_j^{min(s_j, t(\mathbf{u}_i, \boldsymbol{v} \backslash \mathbf{u}_i; r_j))} \tag{1}$$

where $t(\mathbf{u}_i, \boldsymbol{v} \backslash \mathbf{u}_i; r_j) = \sum_i \{\mathbf{1} \|\mathbf{u} - \mathbf{u}_i\| \leq r_j\}$. This density indicates that the spatial interaction between points changes with the distances r_j and the parameters that capture this information are the interaction parameters γ_j. When the s_j is set to infinity, the corresponding component f_j reduces to the Strauss process. Instead if s = 0, the component reduces to the Poisson point process. If s is a finite positive number, then the interaction parameter γ_j may take any positive value. The f_j Geyer component indicates inhibitive interaction when $\gamma_j \leq 1$, and clustered when $\gamma_j > 1$. To consider inhomogeneity in (1), the parameter β is replaced by a function $\beta(\mathbf{u}_i)$ that expresses a spatial trend and it can be a function of the coordinates of the points and of spatial covariates, $Z(u)$.

Gibbs models are fitted to data by pseudo-likelihood that is function of the Papangelou conditional intensity (Baddeley et al.; 2015). The models are compared and assessed in terms of AIC, and graphical diagnostic plots based on the spatial raw residuals. Furthermore, the diagnostic plots based on the residual K- and G-functions are used to decide which component has to be added at each step to the hybrid model. Indeed, these graphs show for which spatial distances the current model has a lack of fit in describing the interaction between points. For example, the residual G-function for a fitted model, evaluated at a given r, is the score residual used for testing the current model against the alternative of a Geyer saturation model with saturation parameter 1 and interaction radius r. We use the the `spatstat` package (Baddeley and Turner; 2005) of R (R Development Core Team; 2005), for fitting, prediction, simulation and validation of Hybrid models.

4 Some results of the analysis

The point pattern of seismic events that is described in Section 2.1 shows a spatial inhomogeneity and multi-scale interactions between points, (Figure 1b). Figure 3 shows the non-parametric empirical G-function and its corresponding envelope: the observed pattern is not a Poisson process since the empirical G-function is outside the shadow region.

The first attempt in model estimation is to fit an inhomogeneous Poisson model with the following parametric log-linear intensity $\rho(\mathbf{u}) = exp\{\beta_0 + g(\mathbf{u}; \boldsymbol{\beta}) + h(D_v(\mathbf{u}), D_{pb}(\mathbf{u}, D_f(\mathbf{u}); \boldsymbol{\alpha})\}$, where $g(\mathbf{u}; \boldsymbol{\beta})$ is a second order polynomial in the spatial coordinates and $h(D_v(\mathbf{u}), D_{pb}(\mathbf{u}), D_f(\mathbf{u}); \boldsymbol{\alpha})$ is a function of the spatial covariates defined in Section 2.2. However, the assumption of an inhomogeneous Poisson point process model is inappropriate since there is an unexplained interaction

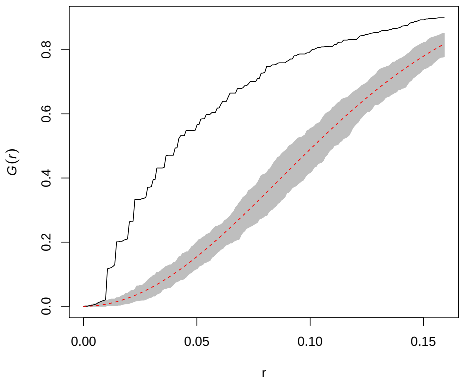


Figure 3: The envelope of the G-function (shaded region) to test the Complete Spatial Randomness (CSR) assumption. The black curve is the estimated G-function instead the red one is the function under the Poisson assumption.

	Estimate	γ_j		Zval
(Intercept)	8.7863			0.45
$Geyer_1$ (r_1=5.6km)	0.0841	1.09	***	12.06
$Geyer_2$ (r_2=6.7km)	0.4627	1.59	***	8.06
$Geyer_3$ (r_3=10km)	0.1193	1.13	***	12.22
$Geyer_4$ (r_4=16km)	0.1968	1.22	***	6.51
D_{pb}	-0.2656		***	-6.25
$I(D_f < \phi_1)D_f$	-1.0196		***	-4.15
$I(\phi_1 \leq D_f < \phi_2)D_f$	-0.4963		*	-2.38
$I(\phi_2 \leq D_f < \phi_3)D_f$	-0.5362		***	-4.39
$I(\phi_3 \leq D_f < \phi_4)D_f$	-1.1964		*	-2.54
$I(D_f \geq \phi_4)D_f$	-1.0059		***	-4.97
x	-1.1664		*	-2.42
y	0.5004			0.63
x^2	0.0273		***	5.48
xy	-0.0045			-0.55
y^2	-0.0065			-0.75

Note: *p<0.1; **p<0.05; ***p<0.01

Table 1: Estimated parameters of the hybrid model with four Geyer components. The log-intensity has a second-order polynomial term in x and y. The variable D_f is inserted considering a segmented linear relationship, where $\{\phi_1, \phi_2, \phi_3, \phi_4\} = \{0.43, 0.60, 1.21, 1.32\}$

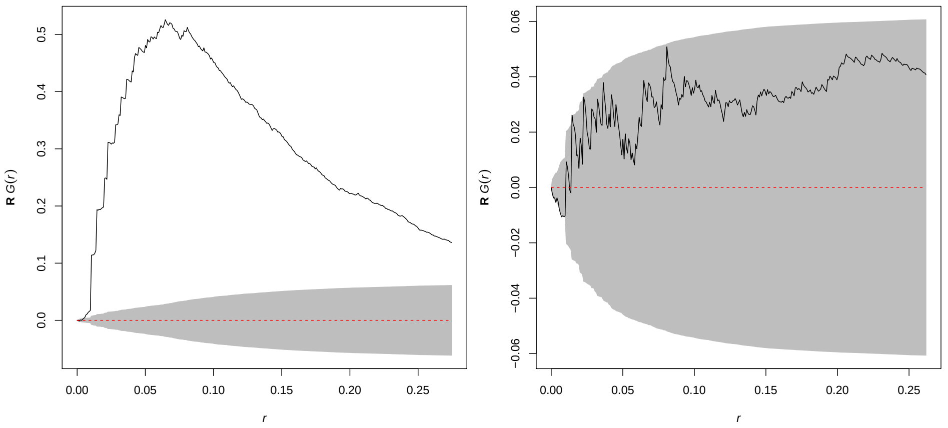
.

between points. The residual G-function has a specific trend that is far from zero for short distances indicating clustering behavior between events (Figure 4a), so, we considered hybrid models. We fitted several combinations of hybrid models and we compared nested models in terms of residual deviance for variable selection. The final selected model is an inhomogeneous hybrid model with four Geyer components that have interaction parameters (γ_j) greater than 1 (Table 1). We identify a multi-scale clustering between points for interpoint distances approximately less than 16 km. The variables $D_f(\mathbf{u})$ and $D_{pb}(\mathbf{u})$ are both significant and negatively related to the log spatial intensity. We found that the variable that indicates the distance to the nearest volcano is not significant in explaining the spatial intensity, in fact the volcanic Hellenic arc area is mostly characterised by microseismic activity. Moving from the inhomogeneous Poisson process to the

	Inhomogeneous model	
	Poisson	Hybrid
AIC	-5247.94	-7121.84
Range of raw residuals	[-5.77;6.14]	[-1.62;2.29]

Table 2: AIC and range of the spatial raw residuals of the fitted inhomogeneous Poisson model and the final selected hybrid model.

hybrid formulation, there is an improvement in terms of AIC, it decreases by 1874. Moreover, there is a sensible reduction of the range of the spatial raw residuals (Table 2). Finally, the residual G-function oscillates around zero indicating that the interaction structure between the earthquakes is well described by the hybrid model (Figure 4b).



(a) (b)

Figure 4: (a) The residual G-function for the inhomogeneous Poisson process model. (b) The residual G-function for the inhomogeneous hybrid model with four Geyer components.

5 Remarks

We describe the available seismic and geological information in the Greek area. Fitting point process models, we obtain parameters that relate the seismic information to the geological one (Table 1). These parameters are relevant for the stakeholders of the analysis because indicate how the intensity changes in the neighbourhood of the several earthquake sources. In fact, we can drawn for the study region these general conclusions: there is a more rapid reduction of the intensity closer to the faults than in the neighborhood of the plate boundary.

A drawback of this analysis is that we cannot do prediction since time dimension is neglected. A way to improve the analysis could be to consider a spatio-temporal model formulation for the catalog data, such as the log-Gaussian Cox model (Møller and Waagepetersen; 2003), in which the covariate geological information are included.

Another important aspect is that the spatial geological information are transformed into spatial covariates considering the distance of a generic point to the sources of earthquakes. It could be interesting to consider in the analysis also the several additional meta-data information that at the moment is not used.

References

Baddeley, A., Rubak, E. and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*, London: Chapman and Hall/CRC Press.

Baddeley, A. and Turner, R. (2005). Spatstat: An r package for analyzing spatial point patterns, *Journal of Statistical Software* **12**(i06).

Baddeley, A., Turner, R., Mateu, J. and Bevan, A. (2013). Hybrids of gibbs point process models and their implementation, *Journal of Statistical Software* **55**(11): 1–43.

Bird, P. (2003). An updated digital model of plate boundaries, *Geochemistry, Geophysics, Geosystems* **4**(3).

Caputo, R., Chatzipetros, A., Pavlides, S. and Sboras, S. (2013). The greek database of seismogenic sources (gredass): state-of-the-art for northern greece, *Annals of Geophysics* **55**(5).

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*, Vol. 70, John Wiley & Sons.

Mignan, A. and Woessner, J. (2012). Estimating the magnitude of completeness for earthquake catalogs, community online resource for statistical seismicity analysis, doi: 10.5078/corssa-00180805.

Møller, J. and Waagepetersen, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*, Chapman and Hall/CRC, Boca Raton.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *Journal of the American Statistical Association* **83**(401): 9–27.

R Development Core Team (2005). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Siebert, L. and Simkin, T. (2014). Volcanoes of the world: an illustrated catalog of holocene volcanoes and their eruptions, *Smithsonian Institution, Global Volcanism Program Digital Information Series, GVP-3* .

Siino, M., Adelfio, G., Mateu, J., Chiodi, M. and D'Alessandro, A. (2016). Spatial pattern analysis using hybrid models: an application to the hellenic seismicity, *Stochastic Environmental Research and Risk Assessment* (DOI: 10.1007/s00477-016-1294-7).