

SEDA a software package for the Statistical Earthquake Data Analysis: a tutorial application to the 2009 L'Aquila and the 2012 Emilia (Italia) sequences

Anna Maria Lombardi¹

¹ Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

Article history

Received September 6, 2016; accepted November 30, 2016.

Subject classification:

Earthquake interactions and probability, Statistical analysis, Algorithms and implementation.

ABSTRACT

The main purpose of this paper is to provide a tutorial application of SEDAv1.0, the first version of a software package, recently designed for the statistical analysis of earthquake data. SEDAv1.0 consists of a user-friendly Matlab-based interface, to facilitate the interaction with the application, and of a computational core of Fortran codes, to guarantee fast running times. The main part of SEDAv1.0 is devoted to the ETAS modeling. For the first time, an almost complete set of consistent tools based on ETAS models is collected in a single, free software. Moreover, SEDA guarantees the research reproducibility, which is becoming an increasingly major concern among scientists. The peculiarities of some routines of SEDAv1.0 are discussed in this paper, by the application to two important recent seismic sequences occurred in Italy. Specifically, the paper illustrates how using SEDAv1.0, to estimate the completeness magnitude and the b-value, to set and test the ETAS model and, finally, to identify the earthquakes sequences, basing on causal connections.

1. Introduction

This paper describes the use of the first version of SEDA (Statistical Earthquake Data Analysis, SEDAv1.0), a new software designed for the statistical analysis of earthquake data, by the application to two recent sequences occurred in Italy.

The tools collected in SEDAv1.0 are classified in two main topics (Lombardi, 2016). The first class, called Catalog Analysis, allows the descriptive analysis of an earthquake catalog, the selection of its subsets and the estimation of the magnitude distribution. This set of tools includes original, but not innovative, codes and supports the user, in managing the database and in evaluating its homogeneity and magnitude completeness. The second group of tools, called ETAS Model, is designed for the analysis of an earthquake database by the ETAS (Epidemic Type Aftershock Se-

quence) modeling (Ogata, 1988; 1998). It is the core of SEDAv1.0 and contains original and partially innovative Fortran codes.

The design of software as SEDAv1.0 is requested by the code share policy, which is a main point for the reproducibility of published research results and highly recommended by the most important scientific journals (Nature Editors, 2014a). Sharing a code means that the source or the executable code is freely accessible to the public. This allows replication of results, which is a key concept in science, and ensures that the scientific community can apply the methodology to their own data, without the need of re-implementing the algorithms. Moreover, the free distribution of a code allows the evaluation of its performance and helps the comparison of different methodologies.

SEDAv1.0 is freely provided via the Zenodo open access platform (<https://zenodo.org/>), a service that allows deposit and DOI assignment to software, besides of ensuring an easy and stable access. Please refer to <https://zenodo.org/record/55277> to download the first version of SEDA for the MAC operating system (a Windows version will be available soon), including the User Manual. Some technical details about SEDAv1.0 are discussed also in Lombardi (2016).

This paper illustrates the use of SEDAv1.0, by mean of an application to two important sequences, occurred in Italy in the last years, following the L'Aquila (April 6, 2009, ML5.9) and the Emilia (May 20, 2012, ML5.9) earthquakes (see Figure 1). Firstly, I estimate the completeness and the b-value for the magnitude distribution. Second, the ETAS model, implemented in SEDAv1.0, is applied and tested on both sequences. Finally, the procedure for sequences identification is presented and discussed.

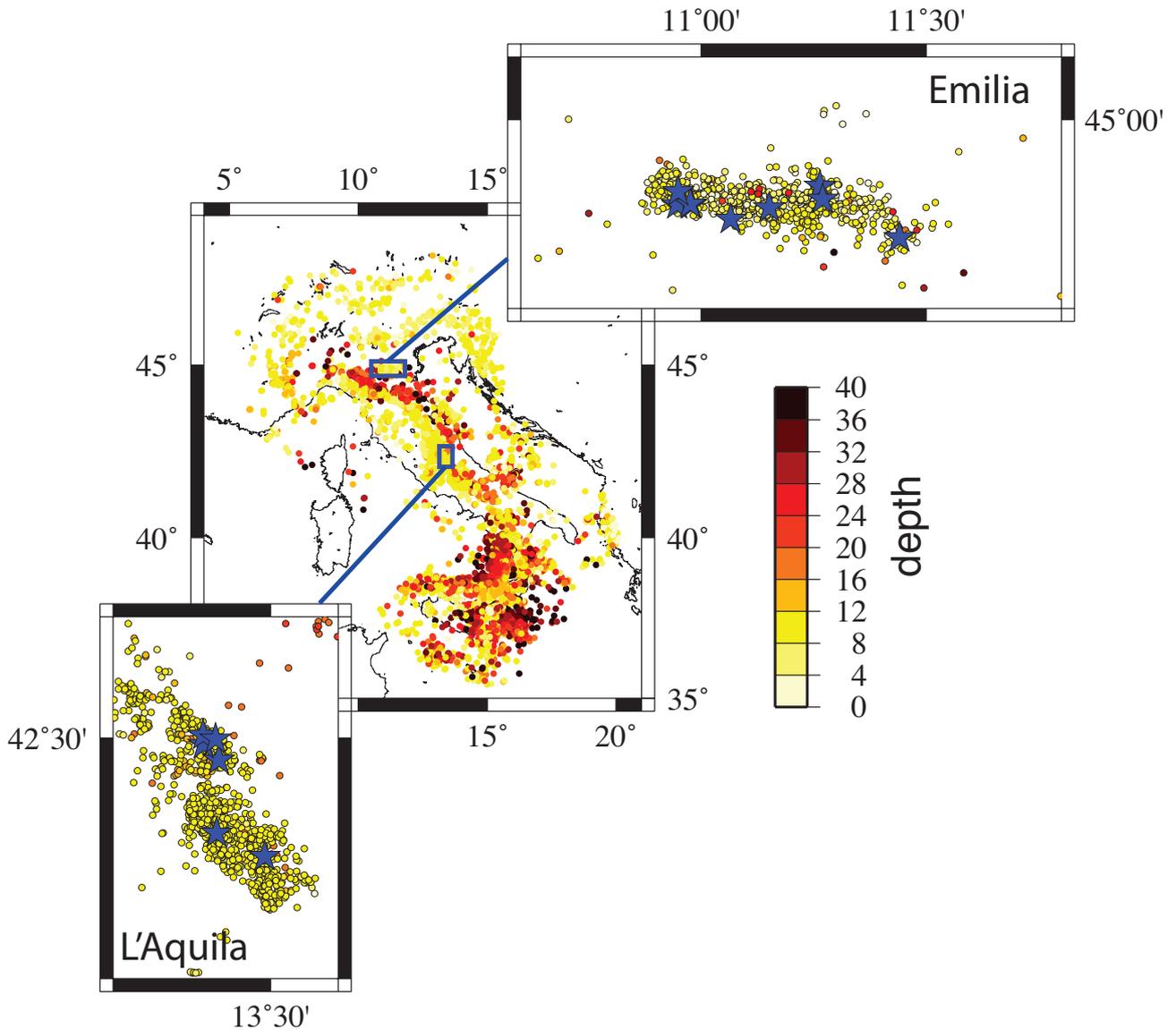


Figure 1. Map of Italian seismicity, occurred from 04/16/2005 to 01/05/2016, with magnitude above ML2.5 and depth above 40kms. The blue rectangles identify the areas of the two most important sequences of the last years: the April 6, 2009 L'Aquila (ML5.9) and the May 20, 2012 Emilia (ML5.9) sequences, of which seismicity is zoomed in the smaller panels. Blue stars mark the events with magnitude above ML5.0.

2. Cases Study: the 2009 ML5.9 L'Aquila and the 2012 ML5.9 Emilia sequences

The L'Aquila and the Emilia sequences are located in areas with different tectonic structure and have specific peculiarities. The L'Aquila region is inside the Central Apennines belt, with prevalent normal faulting, whereas the Emilia area covers alluvial lowland, with thrust faulting. The L'Aquila mainshock was preceded by at least 3 months of moderate-size seismicity and struck an highly hazardous seismic zone of Italy (Stucchi et al., 2010). The Emilia sequence, occurred in a relatively low seismic hazard area, was characterized by a migration of the seismicity towards the E and NE and its strongest aftershock (May 29, 2012, ML5.8) has a magnitude close to the mainshock.

The data of the Italian seismicity are downloaded from the site of the official INGV bulletin, www.iside.rm.ingv.it. I select the events occurred from April 16, 2005 up to April 30 2016. The starting data marks the start-up of a new seismic network, causing a significant improvement of the earthquakes detection (Bono and Badiali, 2005; Schorlemmer et al., 2010).

The areas interested by the L'Aquila and the Emilia sequences are [13.15-13.65E, 42.10-42.70N] and [10.60-11.80E, 44.70-45.10N], respectively (see Figure 1). The INGV bulletin collects 29,787 events (ML from 0.1 to 5.9 and depth above 28km), inside the first area, and 3,014 earthquakes (ML from 1.2 to 5.9 and depth above 37km), in the Emilia region. The lower minimum magnitude of the L'Aquila dataset is due to

the dense seismic network in this zone, whereas the deeper events of the Emilia region are generated from the deep thrusting structures, well-recognized in this zone (Vannoli et al., 2015).

In the following sections I present and discuss the data analysis. The title of each of them contains, in the brackets, the reference to the SEDAv1.0 tools, used for the specific application discussed inside. All the results refer to a retrospective analysis: the parameters of the ETAS model are set on relocated earthquake data, including those of the L'Aquila and Emilia sequences. This strongly differs from the purely prospective, real-time earthquake forecast experiments, made during the L'Aquila and the Emilia seismic emergencies, using provisional data and models independently set (Marzocchi and Lombardi, 2009; Marzocchi et al., 2012).

3. Completeness Magnitude and b-value estimation (Catalog Analysis tool "B-value analysis")

This first part of SEDAv1.0 is devoted to a quick descriptive analysis of an earthquake catalog, including the subsets selection and the magnitude completeness and b-value analysis (Lombardi, 2016).

In SEDAv1.0 the Gutenberg-Richer Law is adopted for magnitudes; it has the following probability density function

$$f(m) = \beta \cdot \exp[-\beta \cdot (m - Mc)] \quad (1)$$

where $\beta = b \cdot \ln(10)$ is a parameter and Mc is the completeness magnitude of the database. Further magnitude distributions will be added in future versions.

SEDAv1.0 assumes a magnitude step of 0.1 and uses two methods to estimate b and Mc : the Mc and B-value Stability method (MBS; Cao and Gao, 2002; Woessner and Wiemer, 2005) and the Goodness of Fit Test method (GFT; Wiemer and Wyss, 2000). This last is performed in SEDAv1.0 both at 90% and at 95% confidence levels. Moreover, SEDAv1.0 fixes a magnitude range equal to 0.5 to calculate the b-value means for the MBS method (see Woessner and Wiemer, 2005, for details).

I run the SEDAv1.0 "B-value analysis" tool on the L'Aquila and Emilia databases. Previous studies, based on the National Seismic Network detection, defined $Mc = 2.5$ as a reasonable completeness magnitude for most of the Italian territory, starting from April 16 2005 (Schorlemmer et al., 2010). By applying the SEDAv1.0 "B-value analysis" tool on the L'Aquila region, I find a completeness local magnitude $Mc = 1.2, 1.5$ and 1.8 by mean of the GFT(90%),

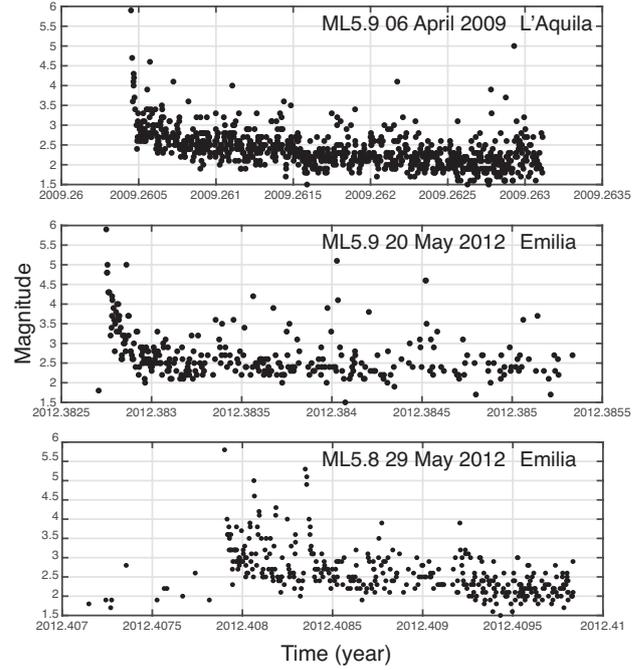


Figure 2. Time-magnitude plot for the events occurred in the first day, after the strongest events, of the L'Aquila and the Emilia sequences. The plot shows an increase of the detection in the first hours after their occurrence.

GFT(95%) and MBS methods, respectively (Table 1). The related b-values are $b = 0.85, 1.0$ and 1.1 . I find larger, more consistent, values of $Mc = 2.1, 2.2, 2.2$ for the Emilia area, while the related b-values are $b = 0.9, 0.97, 0.97$. The larger completeness magnitude of the Emilia region, respect to L'Aquila area, is due to the larger depth of the events and to the lower detection of the National Seismic Network. All the above results are consistent with Schorlemmer et al. (2010).

Previous estimation does not take into account any possible temporal variations of the completeness magnitude. Figure 2 shows the time/magnitude plots of the seismicity, soon after the occurrence of the strongest events of both sequences (April 06 2009, ML5.9, for the L'Aquila sequence; May 20 2012, ML5.9, and May 29 2012, ML 5.8, for the Emilia sequence). They reveal an increase of the minimum detection magnitude up to ML2.5. The values of Mc and b , estimated on the seismicity of the first 12 hours after the strongest events, are reported in Table1. For both sequences there is a significant increase of Mc , up to ML2.5, respect to the overall sequence. Moreover, the b-values decrease down to 0.7, in the Emilia region.

For all above said, I select the events above $Mc = 2.5$ for both sequences. The inferred b-values are 1.1 and 1.0 for the L'Aquila (1153 events) and the Emilia (964 events) regions, respectively (see Table1 and Figure 3).

	L'Aquila				Emilia					
	Overall		April 06 2009 ML 5.9 Firts 12h (454 events)		Overall		May 20 2012 ML 5.9 Firts 12h (220 events)		May 29 2012 ML 5.8 Firts 12h (239 events)	
	Mc	b	Mc	b	Mc	b	Mc	b	Mc	b
GFT 90%	1.2	0.85 (0.005)	2.3	0.97 (0.04)	2.1	0.91 (0.02)	2.1	0.59 (0.03)	2.3	0.64 (0.04)
GFT 95%	1.5	1.00 (0.008)	2.5	1.10 (0.07)	2.2	0.97 (0.02)	NAN	NAN	NAN	NAN
MBS	1.8	1.09 (0.01)	2.5	1.10 (0.07)	2.2	0.97 (0.02)	2.3	0.69 (0.05)	2.5	0.71 (0.05)
FIXED MAGNITUDE (2.5)	2.5	1.08 (0.03)	2.5	1.10 (0.07)	2.5	0.99 (0.04)	2.5	0.75 (0.06)	2.5	0.71 (0.05)

Table 1. Estimation of the Gutenberg-Richter distribution for the L'Aquila and Emilia overall seismicity and for the first hours after the occurrence of the strongest events.

4. ETAS Model

The core of SEDAv1.0 is a set of tools related to the temporal-magnitude (TM) and the time-magnitude-space (TMS) ETAS modeling. A so comprehensive set of tools, based on ETAS, is collected in a single, free software, for the first time. Some packages are been developed and made available in the past, but they refer to

different versions of the ETAS model and allow only a partial analysis of an earthquake catalog (see Lombardi, 2016, for a comprehensive list).

SEDAv1.0 collects two classes of tools, based on the ETAS model. Specifically, the ETAS Basic tools are the most important algorithms concerning an ETAS model. They allow the estimation of the model (parameters and background), the log-likelihood calculation, the declustering of the catalog, the testing of an ETAS model on data, forecasting calculations and the simulations of earthquakes databases. The ETAS Additional tools are designed to deep the investigation of a catalog. They allow the calculation of the background/triggering probabilities for all events, the selection of sequences and the computation of retrospective forecasts.

The conditional intensities of the TM and TMS ETAS models, implemented in SEDA are, respectively:

$$\lambda_{TM}(t, m | H_t) = \left\{ \mu + \sum_{T_i < t} \frac{k \cdot \exp[\alpha \cdot (M_i - Mc)]}{(t - T_i + c)^p} \right\} f(m)$$

$$\lambda_{TMS}(t, m, x, y | H_t) = \left\{ \mu \cdot u(x, y) + \sum_{T_i < t} \frac{k \cdot \exp[\alpha \cdot (M_i - Mc)]}{(t - T_i + c)^p} \cdot \frac{c_{d,q,y}}{[r_i^2 + d^2 \cdot e^{2\gamma(M_i - Mc)}]^q} \right\} f(m) \quad (2)$$

where $f(m) = \frac{\beta \cdot \exp[-\beta \cdot (m - Mc)]}{1 - \exp[-\beta \cdot (Mmax - Mc)]}$ is the magnitude probability density function and $Mmax$ is the maximum magnitude allowed;
 H_t is the history of the process up the time t ;

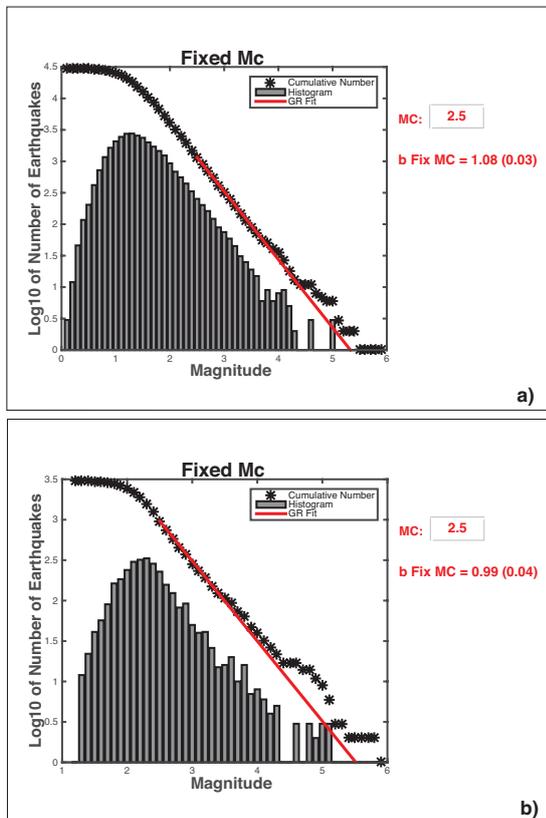


Figure 3. B-value estimation for the L'Aquila and the Emilia databases, obtained by SEDAv1.0, fixing $Mc=2.5$.

$\{u(x,y), (x,y) \in R\}$ is the spatial density function of the background events in the region R of interest. This is assumed uniform in each of the N_c cells C_j (of area A_j) of a regular grid (in degrees), recovering R , so that

$$u(x,y) = \sum_{j=1}^{N_c} \frac{u_j}{A_j} 1_{\{(x,y) \in C_j\}} \quad \text{with} \quad \sum_{j=1}^{N_c} u_j = 1. \quad (3)$$

$\{\mu, k, p, c, \alpha\}$ and $\{\mu, k, p, c, \alpha, d, q, \gamma\}$ are the 5/8 parameters for the TM/TMS model to be estimated;

r_i is the distance (in kms) between the location (x,y) and the epicenter of the i -th event (X_p, Y_i) ;

$c(d, q, \gamma)$ is a normalization constant so that

$$\iint \frac{c_{d,q,\gamma}}{[r_i^2 + d^2 \cdot e^{2\gamma(M_i - M_c)}]^q} = 1 \forall i.$$

SEDA gives the opportunity of including a precursory period (an interval time before the target period, on which the model is estimated or applied) into calculations. In this way, there is a better evaluation of the interactions, since the triggering effect of precursory earthquakes on the target period is taken into account.

The following subsections describe and discuss the application of some of the SEDAv1.0 ‘‘ETAS Model’’ tools on the L’Aquila and Emilia sequences.

4.1 ETAS Parameters Estimation (Basic ETAS tool ‘‘Estimation of Parameters’’)

One of the most innovative codes in SEDAv1.0 is a Simulated Annealing (SA) algorithm for the Maximum Likelihood Estimation of the ETAS model (Lombardi, 2015). Unlike the Newton methods, largely used by the other published codes (Harte, 2010; Ogata, 2006), the method implemented here allows an exhaustive evaluation of the model uncertainties, by mean of multiple runs. In this way, SEDAv1.0 both provides a probability distribution for each parameter (including the background rate) and identifies possible correlations among the parameters, causing a multimodal distribution of the maximum log-likelihood (Lombardi, 2016). Finally, this new algorithm rules out any dependence on the starting values of parameters, with the consequent risk of finding local maxima of the log-likelihood (Lombardi, 2015).

SA algorithms are random iterative procedures that generate a candidate point, that optimizes a function, and move to this point or stay at the current one, based on a stochastic mechanism. This latter is controlled by a parameter T , called temperature, which decreases during the algorithm, according to a ‘‘cooling’’ scheduled. The temperature and the cooling schedule are of critical importance to the success of SA. When T decreases, the search of new points becomes more directive; there-

fore, a too low T might cause an overly restricted search, whereas a high T or a slow cooling schedule might cause an overly high computational time and a possible unsuccessful search. Finally, a too fast cooling schedule can trap the algorithm in a local maximum.

In SEDAv1.0 a new cooling schedule for the temperature is implemented, respect to what proposed by Lombardi (2015). Specifically, SEDAv1.0 adopts the schedule proposed by Ingber (1996)

$$T_j = T_0 \exp\left(l \cdot j^{\frac{1}{D}}\right) \quad (4)$$

where: a) T_j is the temperature at the j -th iteration and T_0 is the initial temperature; b) D is the number of parameters; c) l has the form $l = -\delta \cdot \exp(-\nu/D)$.

If one imposes that after $J = \exp(\nu)$ iterations the values of the temperature is \bar{T} , then one can estimate the parameters δ and ν by the formulas

$$\begin{cases} \nu = \ln(J) \\ \bar{T} = T_j = T_0 \exp(-\sigma) = -\ln\left(\frac{\bar{T}}{T_0}\right) \end{cases} \quad (5)$$

Some applications of the algorithm on simulated data, with J varying from 10 to 100 and \bar{T} varying from 10^{-8} to 10^{-4} , indicate as best parameters $J = 30$ and $\bar{T} = 10^{-5}$, giving $\delta = 3.4$ and $\nu = 13.8$. Therefore, the cooling schedule adopted in SEDAv1.0 is

$$T_j = T_0 \exp\left[-13.8 \cdot \exp\left(-\frac{3.4}{D}\right) \cdot j^{\frac{1}{D}}\right]. \quad (6)$$

I estimate the parameters of the TMS ETAS model on the L’Aquila and Emilia regions, with the learning period April 16 2005, 00:00:00 - January 01

	L’Aquila	Emilia
μ	0.024 (0.022, 0.27)	0.01 (0.009, 0.011)
k	0.047 (0.042, 0.052)	0.050 (0.047, 0.055)
p	1.29 (1.24, 1.31)	1.33 (1.31, 1.36)
c	0.04 (0.03, 0.05)	0.028 (0.023, 0.037)
α	1.45 (1.36, 1.53)	1.14 (1.08, 1.19)
d	1.0 (0.9, 1.2)	1.0 (0.9, 1.2)
q	2.7 (2.4, 2.9)	2.5 (2.2, 2.7)
γ	0.50 (0.46, 0.53)	0.51 (0.47, 0.55)

Table 2. TMS ETAS parameters for the L’Aquila and the Emilia sequences, estimated by SEDAv1.0. The 95% confidence bounds are reported in the brackets.

2007, 00:00:00 and the testing period January 01 2007, 00:00:00 - May 01 2016, 00:00:00. The step of the background grid is equal to 0.01° .

Table 2 lists the values of parameters, together with the 95% confidence bounds, obtained by 100 runs of the simulated annealing algorithm (Lombardi, 2015), implemented in SEDAv1.0. The two sets of parameters are surprisingly very similar, considering the different regional tectonic settings of two sequences. The significantly different background rates μ are due to the different sei-

smic potential of two areas. The smaller α value of the Emilia sequence is likely due to the presence of stronger aftershocks, respect to the L'Aquila sequence (see Figure 3). Figures 4 and 5 show the distribution of the 8 parameters $\{\mu, k, p, c, \alpha, d, q, \gamma\}$ and of the maximum log-likelihood for both sequences. The maximum log-likelihoods given in all runs are very close and mostly recover a range of 2 values, showing a good performance of the estimation procedure. This result is due also to correlation of some parameters, so that different combi-

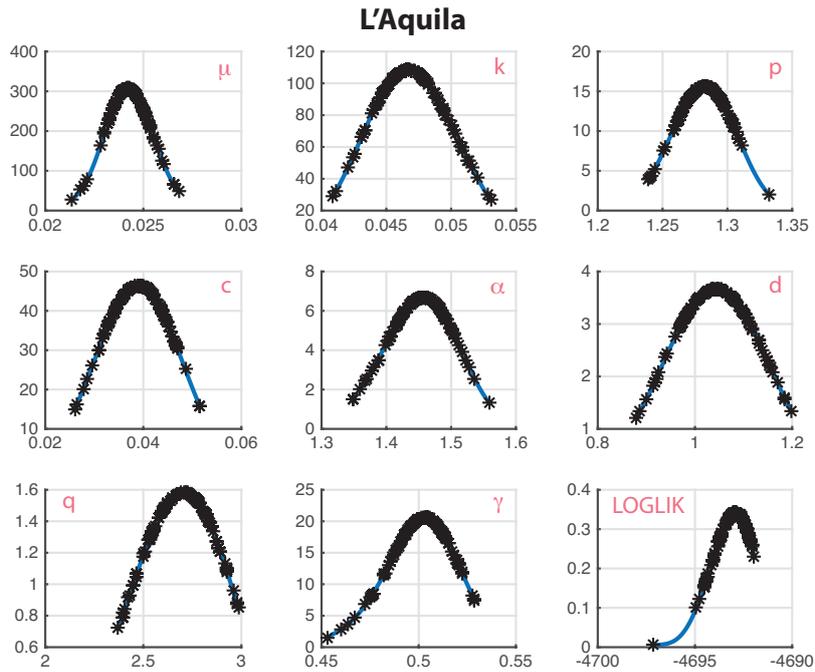


Figure 4. Distribution of the 8 parameters $\{\mu, k, p, c, \alpha, d, q, \gamma\}$ estimated for the L'Aquila sequence, by running 100 times the SEDAv1.0 tool. The last panel shows the distribution of the maximum log-likelihood.

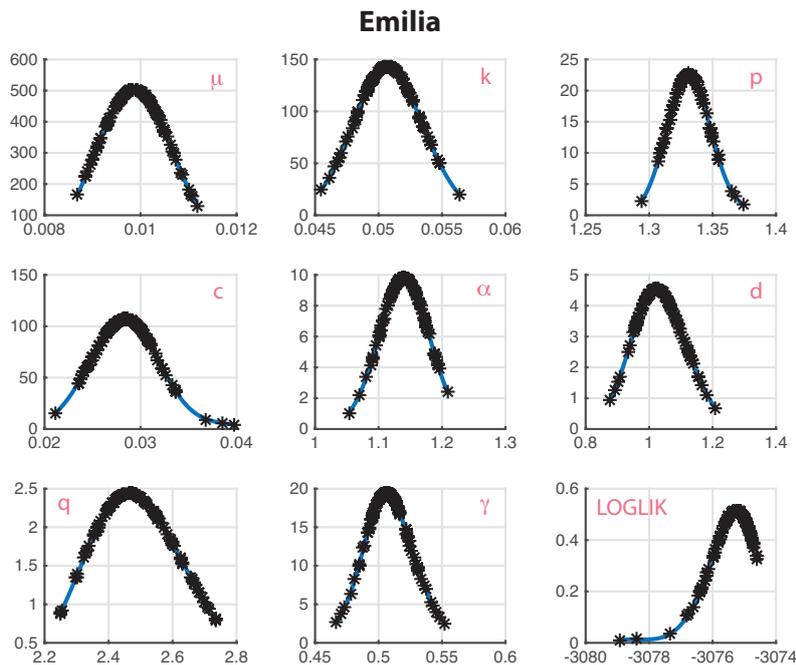


Figure 5. The same of Figure 4, but for the Emilia region.

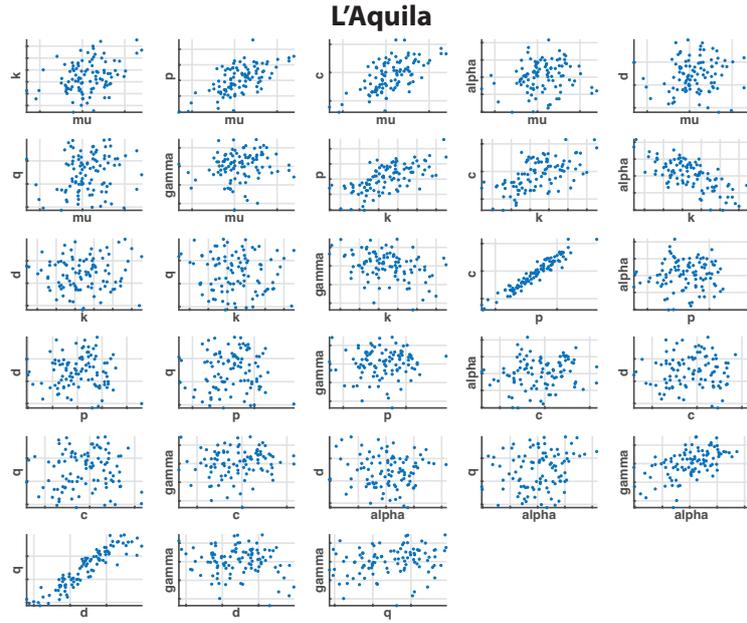


Figure 6. Plot of all possible couples of TMS ETAS parameters, estimated by the 100 runs of the simulated annealing algorithm, implemented in SEDAv1.0, for the L'Aquila sequence.

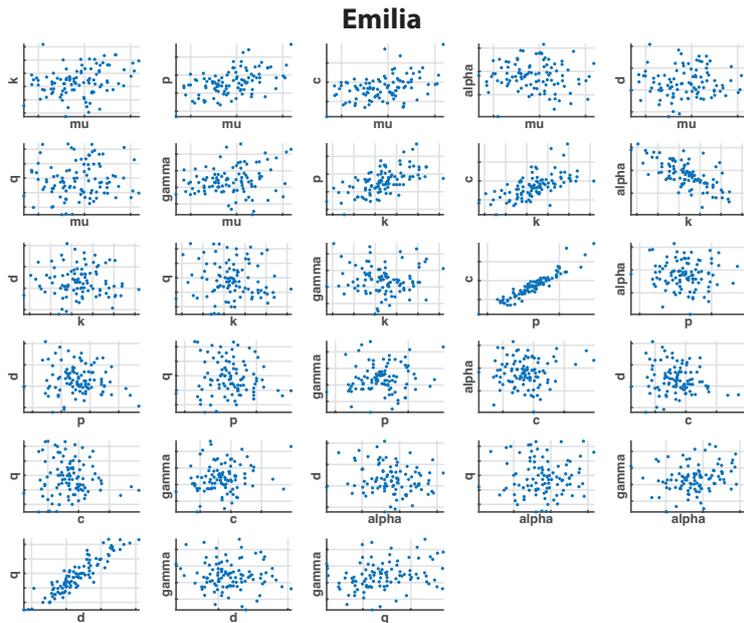


Figure 7. The same of Figure 6, but for the Emilia sequence.

nations of them give very close log-likelihood values and, therefore, indistinguishable models (see Figures 6 and 7). Specifically, the sets of parameters $\{k, p, c, \alpha\}$ and $\{d, q, \gamma\}$ are clearly correlated, whereas the background rate μ seems to be uncorrelated from any other parameter, revealing an univocal way to distinguish the background and the triggered activity.

4.2 Model testing (Basic ETAS tools “Testing the Model/Residual Analysis” and “Testing the Model/Number of events Test”)

A further novelty of SEDAv1.0 concerns the te-

sting of the ETAS model. Actually, SEDA has three test options: the residuals analysis, the Number of events test and the Log-Likelihood test. Whereas the residuals analysis is a well-known procedure (Ogata, 1988), the other two tests are a revised version of the analogue tests adopted by the CSEP laboratories (Schorlemmer et al., 2007). Specifically, the Number of events and the Log-Likelihood tests consist in comparing the size and the log-likelihood of a catalog with the related theoretical distributions, inferred by the ETAS model. These last are obtained computing the size and the log-likelihood on a certain number of ETAS simulated cata-

logs (see Lombardi, 2014, for details). More detailed tests, concerning also the spatial distribution of the events, will be implemented in SEDAv1.0, in the future.

Firstly, I test the inferred ETAS models on both L'Aquila and Emilia regions, by mean of the residual analysis (Ogata, 1988). Specifically, SEDAv1.0 checks the hypothesis of a Poissonian distribution of residuals, by mean of two tests: the Runs test, which examines whether there is a temporal trend in the inter-event transformed times, and the one-sample Kolmogorov-Smirnov (KS1) test, which determines if the inter-event transformed times follow an exponential distribution (Lombardi and Marzocchi, 2007). SEDAv1.0 provides a p-value equal to 0.1 and 0.6, for the Runs and the KS1 tests, respectively, for the L'Aquila sequence. For the Emilia region the correspondent p-values are 0.3 (Runs) and 0.001 (KS1). The inconsistencies between model and data, mainly limited to the first hours after the occurrence of the stronger shocks, are not significant for the L'Aquila sequence. The difference between the observed and the expected number of events (see Figure 8) becomes significant for the Emilia sequence, after the occurrence of the strongest aftershock (May 29 ML5.8).

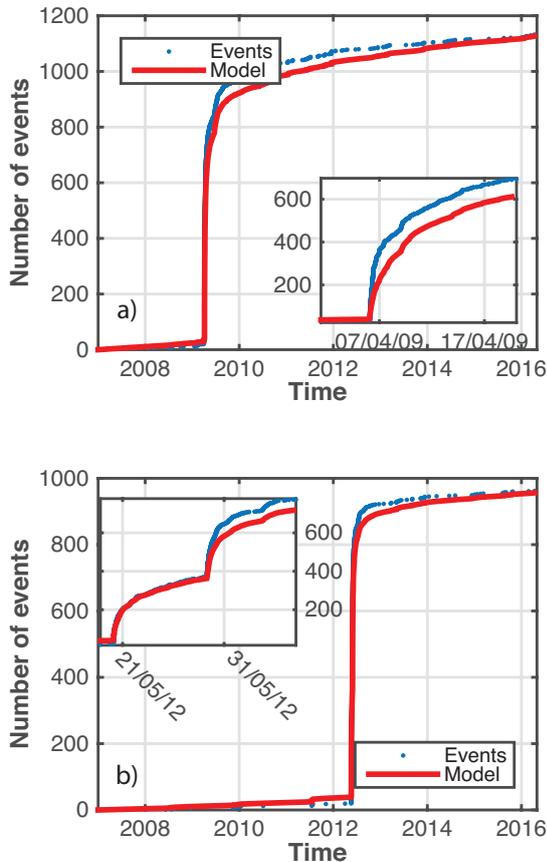


Figure 8. Residuals analysis for the L'Aquila and the Emilia sequences. a) Cumulative plot of observed events (blue points) in the L'Aquila region, together with the expected cumulative distributions (red line). The small panel shows a zoom around the occurrence of the mainshock. b) The same of panel a), but for the Emilia sequence.

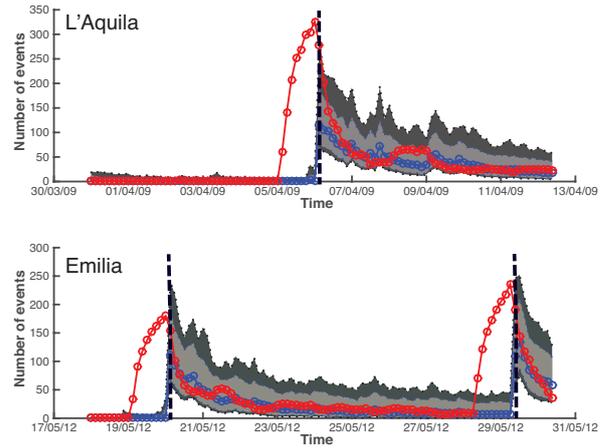


Figure 9. Number of events test for the L'Aquila and the Emilia regions. The expected and observed daily numbers of events are compared. The forecast times are updated each 3 hours. The red and the blue lines mark the observed and expected number of events, respectively. The forecasts are obtained as the median of 1000 ETAS simulations. The light and dark gray shaded areas mark the 95% and 99% confidence interval of forecasts, respectively. The vertical black dashed lines mark the first forecast made with included the strongest events in the history (April 06 2009 ML5.9 for the L'Aquila sequence; May 20 2012 ML5.9 and May 29 2012 ML 5.8 for the Emilia sequence).

I perform a daily Number of events test, which tacks into account the aleatoric uncertainty of the model (Lombardi, 2014), to better evaluate the consistency between model and data. Specifically, I compare the observed daily number of events with the distribution obtained from 1000 ETAS simulations. The daily forecasting windows are updated each 3 hours, together with the history of events actually occurred before. Figure 9 shows a good agreement between forecasts and observations, since these last are inside the 95% or the 99% confidence intervals of forecasts. Some inconsistencies concern the day before the occurrence of the strongest events (April 06 2009, ML5.9, for the L'Aquila sequence; May 20 2012, ML5.9, and May 29 2012, ML 5.8, for the Emilia sequence); this reveals the unpredictability of their occurrence and, therefore, of the following significant increases of the seismic rate.

The sensible agreement between model and data shows that the inconsistencies founded during the real-time forecasting experiments (Marzocchi and Lombardi, 2009; Marzocchi et al., 2012) are likely due to the provisional nature of the earthquake data and of the ETAS model. There is, therefore, no need to turn to nonstationary versions of the ETAS models, for describing these sequences.

4.3 Sequences identification (ETAS Additional tool "Identify sequences")

One of the most important ETAS additional tools of SEDAv1.0 is designed for the identification of se-

quences and is based on the stochastic reconstruction, proposed by Zhuang et al. (2004). Specifically, the tool “Identify sequences” applies the random method of Zhuang et al. (2004) $NRUN$ times and assigns a set of integers $\{isi, 1, \dots, isi, NRUN\}$ to each event E_i , where $is_{i,k}$ is a number, that identifies the sequence of E_i , for the k -th run. Then, the user may choose a target event \bar{E}_j and select all the events belonging to the same sequence of \bar{E}_j , with a probability above a chosen probability level PL . The probability $Pi(\bar{E}_j)$ that E_i and \bar{E}_j , belong to the same sequence is computed as

$$Pi(\bar{E}_j) = \frac{\sum_{k=1}^{NRUN} 1_{\{is_{i,k}=is_{j,k}\}}}{NRUN} \quad (8)$$

The tool is run, by choosing the mainshocks as target event \bar{E}_j , for both sequences. SEDAv1.0 identifies a sequence of 817 events, from 30/03/2009 to 12/08/2009, for 2009 L’Aquila sequence, by mean of $NRUN=1,000$ stochastic reconstructions, with a threshold probability $PL=0.99$. For $PL=0.95$, SEDAv1.0 identifies 855 events, from 17/02/2009 to 14/06/2010 (see Figure 10). For the Emilia sequence, SEDAv1.0 selects 843 events, from 19/05/2012 to 05/09/2012, for $PL=0.99$ and 875 events, from 19/05/2012 to 26/10/2012, for $PL=0.95$.

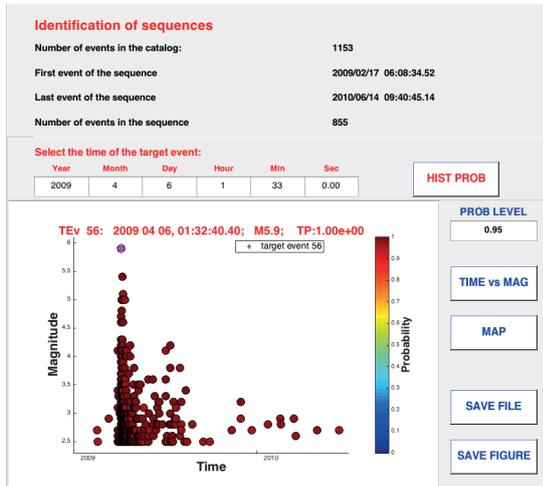


Figure 10. Screenshot from SEDAv1.0 showing the results of the L’Aquila sequence identification for $PL=0.95$, choosing as target event the mainshock (April 6 2009, ML5.9; see text for details). The figure shows the time-magnitude plot of the identified events.

If one runs the tool more times, then the number of events and the temporal limits of the sequences may slightly change, due to the stochastic nature of the method. Figure 11 shows the temporal limits of both sequences, together with the number of selected events, as function of PL . For both sequences, the temporal limits significantly vary for $PL>95\%$ and are more stable for smaller values of PL , down to about

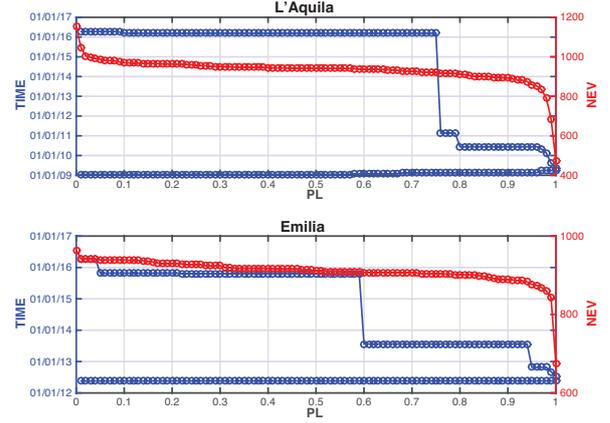


Figure 11. Temporal limits (blue line) and number of events (red line) as functions of the probability threshold PL for both the L’Aquila and the Emilia sequences. The target events are the mainshocks (April 06 2009 ML 5.9 for the L’Aquila and May 20 2012 ML 5.9 for the Emilia). The probability threshold PL goes from 0 to 1, with a step of 0.01 (see text for details).

75% for L’Aquila and 60% for Emilia. The remaining values of PL are probably too small for an unequivocal assignment of the events to the sequences, given also the limited size of both catalogs, due to the relatively high value of Mc .

4.4 Retrospective forecast (ETAS Additional tool “Retrospective Forecast”)

The retrospective forecast gives a further possibility to test a model. The “Retrospective Forecast” SEDAv1.0 tool was implemented in a module apart from the other tests, among the ETAS additional tools, since it is not still a formalized test. It currently consists in comparing the overall expected number of events above a magnitude $MF \geq Mc$, given by the catalog, with the analogue values, given by a certain number of simulated catalogs. Specifically, SEDAv1.0 computes the forecasts fr_i^{TM} (TM model) and fr_{ij}^{TMS} (TMS model) (i.e. the number of expected events, conditional on the past) for the i -th simulated catalog, by the formulae

$$fr_i^{TM} = \int_{T_1}^{T_2} \int_{MF}^{M_{\max}} \lambda_{TM}(t, m | H_t^i) dt dm \quad (7)$$

$$fr_{ij}^{TMS} = \int_{T_1}^{T_2} \iint_{C_j} \int_{MF}^{M_{\max}} \lambda(t, x, y, m | H_t^i) dt dx dy dm$$

where H_t^i is the history up time t , collected in the i -th simulated catalog. These quantities represent a probability distribution for the ETAS forecasts and give the probabilities to have larger values of the “observed” expected values \bar{fr}_i^{TM} or \bar{fr}_j^{TMS} , obtained by replacing the actual observed history \bar{H}_t , in formulas (7).

The retrospective forecast can provide an impor-

tant feedback, regarding the performance of the ETAS model and may be applied in different way, by varying the forecast time period or the threshold forecast magnitude MF .

Two specific applications are shown in the following. The first is an evaluation of the probability to have an event above ML5.0, in the L'Aquila area, during the day April 06 2009, when the mainshock actually occurred. In the second example, I compute the probability to have an event above ML5.5, during the day May 29 2012, in which the strongest aftershock occurred (see Figure 12).

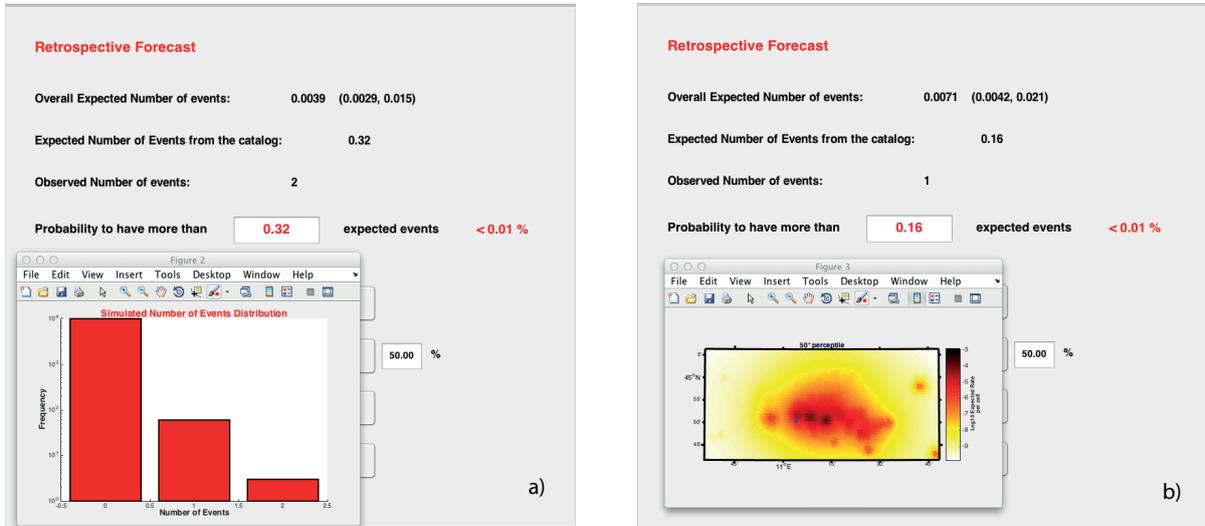


Figure 12. Screenshots from SEDAv1.0 showing the results of the retrospective forecasting for the L'Aquila and the Emilia sequences. a) Retrospective forecasting of the seismicity occurred from April 06 2009, 00:00:00 to April 07 2009, 00:00:00, in the L'Aquila area, with $MF=5.0$ and 10,000 simulations. The histogram shows the distribution of the number of events above 5.0 for the simulations. The observed number is 2. b) Retrospective forecasting of the seismicity occurred from May 29 2012, 00:00:00 to May 30 2012, 00:00:00, in the Emilia area, with $MF=5.5$ and 10,000 simulations. The map shows the spatial distribution of the expected median number of events ($\text{median} \{ \sum_{j=1}^{N_c} f_{ij}^{TMS}; i=1, \dots, 10000 \}$) above 5.5 for each cell C_j (of size $0.01^\circ \times 0.01^\circ$).

In the first example, I set as learning period April 16 2005, 00:00:00 - April 06 2009, 00:00:00 and I compute the forecasts f_{ij}^{TMS} by mean of 10,000 simulations. The value of MF is 5.0 and the forecast period is April 06 2009, 00:00:00 - April 07 2009, 00:00:00. The overall median expected rate $\text{median} \{ \sum_{j=1}^{N_c} f_{ij}^{TMS}; i=1, \dots, 10000 \}$ is equal to 0.0039 (the 95% confidence interval is [0.0029 0.015]). The observed value $\sum_{j=1}^{N_c} \bar{f}_j^{TMS}$ is equal to 0.32 that is outside the inferred distribution. 3 on 10,000 simulated catalogs have 2 events above ML5.0, as actually occurred in the forecast day.

In the second example, related to the Emilia sequence, the forecast period is May 29 2012, 00:00:00 – May 30 2012, 00:00:00 and $MF=5.5$. SEDAv1.0 gives an overall median expected rate equal to 0.0071 and a 95% confidence interval [0.0042 0.021]. The observed value is 0.16 that is outside the inferred distribution.

90 simulated catalogs on 10,000 have 1 event above ML5.5, at least, as actually observed.

The comparison of these last results with those obtained by the model testing (see subsection 4.2) shows a good ability of the ETAS model to reproduce the overall seismic rate, but also the “unpredictability” of the time occurrence of specific strong shocks.

5. Conclusions: the future of SEDA

The program SEDA, presented in this paper, has been developed with the main aim of allowing an easy use of the tools. Specifically, the Matlab GUI facilitates

the user control and allows the display of the results. Moreover the modular design of the GUI allows an easy upgrade of the actual version and the inclusion of new modules.

For the first time, the most important operations concerning the ETAS model are collected in a single, free, user-friendly, software. From a scientific point of view, the most important novelties are the estimation method: this ensures the reaching of the overall best solution and gives an effective evaluation of the model uncertainties. Particular care has been taken in developing the algorithms for the model testing and for the identification of sequences.

SEDA has been developed in order to guarantee the reproducibility of research. The computational methods used in many published papers are often not fully explained or described. This is due both to con-

straints imposed by the traditional research papers and to the reluctance to share intellectual property. As a consequence, it is difficult to reproduce research results, to verify their correctness and to build on them in future research and applications. Reproducible research has become a prominent issue in several academic fields (see, for example, Liu et al., 2015; Irving, 2015, among many others published papers). Journals as *Science* and *Nature* have published numerous articles about reproducibility and editorials in favor of policies that promote open science. Finally the code sharing should at least lead to better quality of codes themselves (Nature Editors, 2014a; 2014b; McNutt, 2014; Geoscientific Model Development Editors, 2013).

The main aim of this paper is to provide a tutorial example for the use of SEDAv1.0, but it gives also important information on the ETAS modeling. The analysis reported here clearly shows that the two sets of ETAS parameters, estimated on the L'Aquila and the Emilia regions, are surprisingly very similar, considering the different regional tectonic settings of two sequences. Significant differences are found only for the μ and α parameters (see Table 2). The first parameter is a proxy of the seismic potential of the zone and the estimated values reveal the lower hazard of the Emilia region. The α parameter is related to the magnitude distribution and to the presence of large aftershocks in major sequences. The magnitude distribution of the Emilia dataset shows a clear bump for magnitudes above ML4.5. The L'Aquila sequence has 5 aftershocks above ML5.0, with a magnitude ranging from ML5.0 to ML5.4, whereas the Emilia sequence has 8 aftershocks above ML5.0, having a maximum magnitude equal to ML5.8. This may be the reason of the lower α value estimated for the Emilia dataset.

The tests of the ETAS models and the retrospective forecast calculations reveal a good agreement with observations (see Figures 8 and 9). Specifically, the inferred models are able to predict the temporal evolution of the overall rate, without the inclusion of time-varying parameters. Nevertheless, they do not have good skill in forecasting specific large earthquakes and the following sudden increase of the seismic rate (Figure 12).

The future of SEDA is somewhat unclear and will be driven by research interests and by the collaborations, besides of suggestions and criticism. Anyway, some advances and improvements are already scheduled (Lombardi, 2016). Great effort has been devoted to test all the tools of SEDAv1.0, nevertheless some undetected errors may exist and some features may remain

cryptic to many users. Please, send an e-mail at the address annamaria.lombardi@ingv.it, for questions, suggestions or bugs to report.

References

- Bono, A. and L. Badiali (2005). Pwl personal wavelab 1.0, an object-oriented workbench for seismogram analysis on windows system, *Comput. Geosci.*, 31, 55-64.
- Cao, A.M. and S.S. Gao (2002). Temporal variation of seismic b-values beneath northeastern Japan island arc, *Geoph. Res. Lett.*, 29(9), doi:10.1029/2001GL013775.
- Geoscientific Model Development Editors (2013). Editorial: The publication of geoscientific model developments v1.0, *Geosci. Model Dev.*, 6, 1233-1242, doi:10.5194/gmd-6-1233-2013.
- Harte, D. (2010). PtProcess: An R Package for Modeling Marked Point Processes Indexed by Time, *J. Stat. Softw.*, 35(8).
- Ingber, L. (1996). Adaptive simulated annealing (asa): Lessons learned. *Control Cybern.*, 25(1):33-54.
- Irving, D. (2015). A minimum standard for publishing computational results in the weather and climate sciences, *Bull. Am. Met. Soc.*, doi:10.1175/BAMS-D-15-00010.1.
- Liu, L., S. Peng, C. Zhang, R. Li, B. Wang, C. Sun, Q. Liu, L. Dong, L. Li, Y. Shi, Y. He, W. Zhao and G. Yang (2015). Importance of bitwise identical reproducibility in earth system modeling and status report. *Geosci. Model Dev.*, 8, 4375-4400, doi:10.5194/gmdd-8-4375-2015.
- Lombardi, A.M. and W. Marzocchi (2007). Evidence of clustering and nonstationarity in the time distribution of large worldwide earthquakes, *J. Geoph. Res.*, 112, B02303; doi:10.1029/2006JB004568.
- Lombardi A.M. (2014). Some reasoning on the RELM-CSEP Likelihood-Based Tests, *Earth Planets Space*, 66(4).
- Lombardi A.M. (2015). Estimation of the parameters of ETAS models by Simulated Annealing, *Sc. Rep.*, Feb 12;5:8417. doi: 10.1038/srep08417.
- Lombardi A.M. (2016). SEDA a software package for the Statistical Earthquake Data Analysis, *Sc. Rep.*, accepted.
- Marzocchi, W. and A.M. Lombardi (2009). Real-time forecasting following a damaging earthquake, *Geoph. Res. Lett.*, 36, L21302; doi:10.1029/2009GL040233.
- Marzocchi, W., M. Murru, A.M. Lombardi, G. Falcone and R. Console (2012). Daily earthquake forecast

- during the May-June 2012 Emilia earthquakes sequence (Northern Italy), *Ann. Geophys.*, 55(4), 561-567, doi: 10.4401/ag-6161.
- McNutt, M. (2014). Journals unite for reproducibility, *Science* 346 (6210), 679, doi:10.1126/science.aaa1724.
- Nature Editors (2014a). Code share. Papers in Nature journals should make computer code accessible where possible, *Nature*, 514, 536, doi:10.1038/514536a.
- Nature Editors (2014b). Journals unite for reproducibility. Consensus on reporting principles aims to improve quality control in biomedical research and encourage public trust in science, *Nature*, 515 doi:10.1038/515007a.
- Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.*, 83, 9-27.
- Ogata, Y. (1998). Space-Time Point-Process models for Earthquake Occurrences, *Ann. I. Stat. Math.*, 50, 379-402.
- Ogata, Y. (2006). *ISM Computer Science Monographs*, No. 33, The Institute of Statistical Mathematics, Tokyo, Japan.
- Schorlemmer, D., M.C. Gerstenberger, S. Wiemer, D.D. Jackson and D.A. Rhoades (2007). Earthquake Likelihood Model Testing, *Seism. Res. Lett.*, Vol. 78, No. 1, 17-29.
- Schorlemmer, D., F. Mele and W. Marzocchi (2010). A Completeness Analysis of the National Seismic Network of Italy, *J. Geophys. Res.*, 115, B04308; doi: 10.1029/2008JB006097.
- Stucchi, M., C. Meletti, A. Rovida, V. D'Amico and A.A. Gomez Capera (2010). Historical earthquake and seismic hazard of the L'Aquila area, *Progettazione Sismica*, 3, 23-34.
- Vannoli, P., P. Burrato, and G. Valensise (2015). The Seismotectonics of the Po Plain (Northern Italy): Tectonic Diversity in a Blind Faulting Domain, *Pure Appl. Geophys.*, 172, 1105-1142.
- Wiemer, S. and M. Wyss (2000). Minimum magnitude of complete reporting in earthquake catalogs: examples from Alaska, the Western United States, and Japan, *Bull. Seism. Soc. Am.*, 90, 859-869.
- Woessner, J. and S. Wiemer (2005). Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty, *Bull. Seism. Soc. Am.*, 95 (2), 684-698.
- Zhuang, J., Y. Ogata and D. Vere-Jones (2004). Analyzing earthquake clustering features by using stochastic reconstruction, *J. Geophys. Res.*, 109.

*Corresponding author: Anna Maria Lombardi,
Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy
email: annamaria.lombardi@ingv.it