ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics. Master Thesis Econometrics & Management Science: Econometrics

An Epidemic Type Aftershock Sequence (ETAS) model on conflict data in Africa

Author:
Gilian van den Hengel
455281gh

Supervisor:
Prof. dr. Ph.H.B.F. Franses
Second assessor
dr. H.J.W.G. Kole

Abstract

The continent of Africa is infamous for its many social conflicts. Research in the dynamics of these conflicts can give us a better understanding and can be useful in preventing them in the future. In this study we model social conflicts in Africa as earthquake occurrences using the spatial-temporal Epidemic Type Aftershock Sequence (ETAS) model. The model parameters are estimated using a simulated annealing algorithm and the results are compared across four major geographical regions in Africa. Several model tests will be applied to study the performance of the model in capturing the dynamics of the social conflicts, as well as to detect similarities between earthquake and conflict catalogs. Finally, the estimated model is used to make a prediction for the amount of future large events and their locations. The results suggest that this model produces useful insights in identifying current, and predicting future conflict regions. In particular, the model predicted a cluster of large events in the Central Africa region which was not expected based on past events and we found that small conflicts can trigger larger conflicts easier in East Africa than in other regions.

Keywords— ETAS model, Epidemic Type Aftershock Sequence, social conflicts, Africa, simulated annealing

Contents

1 Introduction **3**

2 Data **5**
2.1 General description 5
2.2 Summary statistics 6

3 Method **11**
3.1 Model 11
3.2 Estimation method 16
3.3 Model testing 18
3.3.1 Residual analysis 18
3.3.2 Number of events test 19
3.4 Forecasting 19

4 Simulation **21**
4.1 Simulation procedure 21
4.2 Estimation results 21
4.3 Model testing 24
4.3.1 Residual Analysis 24
4.3.2 Number of events test 25
4.4 Forecasting results 26

5 Application to social conflict data **28**
5.1 Data description 28
5.2 Background distributions 29
5.3 Parameter estimates 33
5.4 Model testing 39
5.4.1 Residual analysis 39
5.4.2 Number of events test 41
5.5 Forecasting results 42

6 Conclusion & discussion **46**

7 Appendix **49**
7.1 Magnitude distribution of social conflicts 49
7.2 Model tests 52
7.3 Residual analysis 53

Acknowledgements

I would like to thank prof. Philip Hans Franses for supervising my thesis, for the quick and valuable responses to my questions and sparking my interest in the subject of African conflicts. Furthermore, I would like to thank dr. Erik Kole for being the coreader of my thesis.

1 Introduction

The continent of Africa is infamous for its many social conflicts. In 2014 Cilliers (2015) reported that more than half of conflicts world wide took place in Africa, despite having only 15% of the world population. TIME magazine places the further expansion of conflicts throughout Africa among the top 10 risks in the world in 2018[1]. Understanding the dynamics of conflicts is important in solving them and preventing them in the future. Several studies have been done in trying to predict political or social conflicts. For example, Zammit-Mangion et al. (2012) apply Log-Gaussian Cox models to study the Afghan war, Brandt et al. (2011) use Bayesian VAR models to predict material conflict between the Israelis and Palestinians and Johnson et al. (2017) use self-exciting point processes to predict the amount of political conflict in Venezuela and Colombia. In this study we will investigate whether social conflicts can be modeled as earthquakes.

The reason to model social conflicts as earthquakes is that they show several similarities. First, they both have a self-exciting nature. There can be long periods of non-activity in which there is much tension build-up. Once an event does happen, this tension is released, causing more events, which in turn can cause more events. This triggering increases with the size of the event and decreases over time and distance. Secondly, there is a clustering component. Earthquakes often take place around the edge of tectonic plates and conflicts around country or region borders. Triggered events are more likely to occur near their ancestor events than farther away, so they also tend to cluster around themselves. To study the social conflicts, we estimate an Epidemic Type Aftershock Sequence (ETAS) model on social conflict data in Africa. ETAS models are a special type of a self-exciting point process and are widely used in studying earthquake occurrences. In this model, events are divided into two categories: events that occur spontaneously and events that are triggered by other events. Each event can trigger new events, which in turn can trigger new events. The model then describes the intensity and locations of spontaneous background events and the triggering dynamics. After estimating the model, we will try to predict the number and locations of large events in the future.

Before we estimate the model parameters, we will discuss the social conflict data. A general description about the data and the way it is collected will be given, together with some summary statistics, the distribution of events over space and time and the distribution of the sizes of the conflicts. After that, we will discuss the model, estimation method and several model tests. Furthermore, the method of forecasting future events will be discussed. These procedures will then first be applied to a simulated conflict catalog. In this thesis, the term catalog refers to the dataset of the time, location and size of events. Next, we will apply the model to four different datasets. These are the conflict data from four of the five major geographical regions in Africa: North, West, Central

[1] http://time.com/5083778/2018-top-risks-world/

and East Africa. South Africa is omitted due to the low number of events. The estimation results can then be compared across the four different regions. Finally, we make a forecast for the number and locations of large events in each region.

2 Data

In this section we will discuss the social conflict data. First a general description will be given, together with the distribution of conflicts in time, space and size.

2.1 General description

The conflict data is obtained from the Armed Conflict Location & Event Data (ACLED) Project [2]. The project collects the dates, actors, types of violence, locations, and fatalities of all reported political violence and protest events across Africa since 1997. Because of their complex nature, modeling social conflicts requires a large amount of simplification. For this study, we define a social conflict in the following way:

Definition: A social conflict is an event which consists of four numbers: an event date, location (latitude and longitude) and the number of fatalities.

Hence, we will only look at the date, place and size of conflicts. The model we will use requires an exact time stamp as well, which is often not reported for conflicts. Therefore, every conflict is set to have happened at 12:00 noon.

There is some inherent inaccuracy in the four numbers that define a conflict, since the exact time, location or number of fatalities is not always known. It is therefore important to be aware of the data limitations. The complete description of methodological accuracies and definitions can be found on the ACLED website[3]. We will briefly discuss the most relevant ones. The number of fatalities are estimated as the number reported by source material and are not verified by the ACLED project itself. In order to be most reliable, the ACLED project uses the most conservative estimates. For example, if a news article mentions 'several', 'many' or 'few' fatalities, it is recorded as 10. 'Dozens' is recorded as 12. If a report speaks of 'hundreds of casualties' or a 'massacre' it is recorded as 100. The location of a conflict is recorded as the location of the nearest town or city and if a larger region is described, the provincial capital of that region is used. Furthermore, if a conflict lasts several days, the event is recorded as a new event each day with the average number of fatalities over those days. In this study, we take the data 'as is' and assume that the dates, locations and numbers of fatalities are accurate.

[2] https://www.acleddata.com/data/
[3] https://www.acleddata.com/resources/methodology/

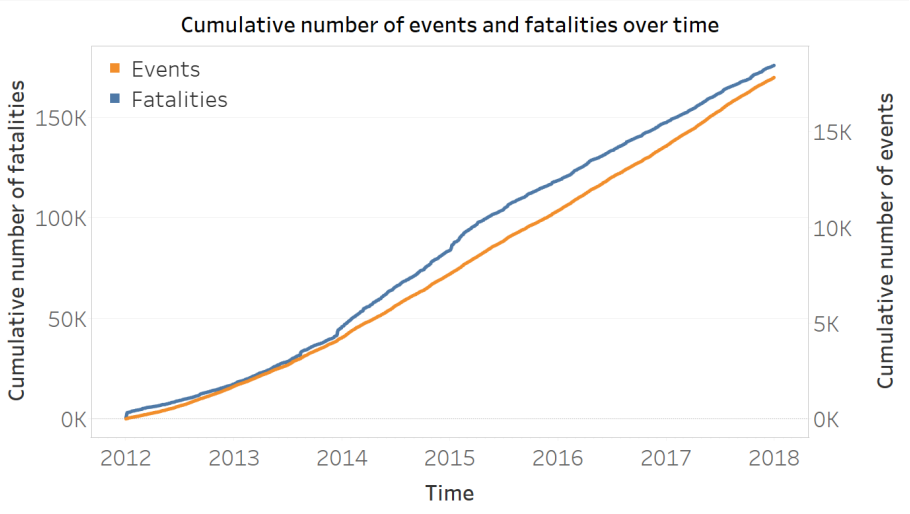
2.2 Summary statistics

We will now discuss the distribution of events over time, space and size.

Time
For this study we will analyze conflicts between 2012 and 2017 with at least two fatalities. This time window and fatality range results in 17,334 events in total with around 2500-7000 events in each of the major regions. This number of events should be large enough to accurately estimate the model in each region, see Lombardi (2017). Using more years of data or including events with one fatality makes estimating the model computationally infeasible.

In Figure 1, the cumulative number of events and fatalities over time are shown. We see that the average number of events per day increases slightly after 2014 but remains relatively constant over the whole period. The number of fatalities per event also remains relatively constant, except for an increase between 2014 and 2015. One possible explanation could be the start of the Second Libyan Civil War in 2014.

Figure 1: Cumulative number of events and fatalities over time.



Space
To explore the spatial distribution of conflicts, all events are displayed on a map in Figure 2. The size of the dots represent the number of fatalities and events with more than one hundred fatalities are colored red. We can see several large clusters of events. For example we see many events with more than a hundred fatalities in the top right corner of Nigeria, an area where the Boko Haram insurgency is active. In the top right corner of Egypt, conflicts resulting from the Arab Spring can be seen. Furthermore, there are large clusters of conflicts

around the Sudan - South-Sudan border resulting from the South Sudanese Civil War. Other notable clusters can be seen in the eastern part of the Democratic Republic of the Congo and in the south of Somalia. Overall, we see that conflicts tend to be located more around country borders than inland, as expected.

Time & Space
To explore the spatial distribution of the events over time, the latitude and longitude of each event are plotted in Figure 3. The red lines represent the latitude and longitude of the geographical center of Africa. We see that there are more events in the northern half of Africa than in the southern half, and more events in the eastern half than in the western half. In the top figure we see that most conflicts are clustered around the equator and that there are relatively fewer events in the south of Africa. In the bottom figure we see that the events are more evenly distributed over longitude. Over time, events tend to stay around the same latitudes and longitudes. One clear exception is the appearance of events around 30 degrees latitude from 2013 onward. This could be due to the the many conflicts resulting from the Arab Spring.

Size
Finally, we will discuss the distribution of the sizes, or magnitudes, of the conflict events. In Table 1 the fatality quantiles are given. It can be seen that the distribution of fatalities is very skewed: most events have a small number of fatalities and relatively few have a large number. Over half of the events have 4 fatalities or fewer, 99% of the events have 100 fatalities or fewer and the largest event has 600 fatalities.

Table 1: Fatality quantiles of events in the ACLED dataset between 2012 and 2017 with at least 2 fatalities.

Quantile	Fatalities	Quantile	Fatalities
0	2	0.91	20
0.1	2	0.92	22
0.2	2	0.93	25
0.3	3	0.94	28
0.4	3	0.95	31
0.5	4	0.96	38
0.6	6	0.97	47
0.7	8	0.98	58
0.8	10	0.99	96
0.9	19	1	600

The ETAS model we use assumes that the magnitudes of events follow the Gutenberg-Richter law, which is given by

$$N = 10^{a-bM}, \tag{1}$$

where N is the number of events with magnitude M or larger and a and b

are constants. The value of b is typically close to one and in that case, events with one larger magnitude are ten times less likely to happen. To make the conflicts follow this law, we group the events based on their fatalities and assign magnitudes to each group in such a way that the number of events at each increasing magnitude decreases by a factor of ten. The complete derivation of the correspondence between fatalities and magnitudes is given in section 7.1 in the appendix.

Doing this results in the magnitude distribution for the conflict events given in Table 2, where the magnitude range is $\{3.0, 3.1, 3.2, ..., 7\}$. This table only shows the integer magnitudes. The full table is given in Table 21 in the appendix, together with the exact derivation of this distribution. In the left two columns, the correspondence between fatalities and magnitude is shown. The next two columns show the percentage of events expected by the Gutenberg-Richter law at each magnitude, together with the observed percentage. The last two columns show the expected and observed absolute number of events. We see that each magnitude step, the percentage and number of events decreases roughly by a factor of ten.

Table 2: Partial table of the magnitude distribution of the social conflict data. For each magnitude (Magn.), we have the corresponding number of fatalities (Fatal.), the percentage of events expected by the Gutenberg-Richter law (% exp.), the observed percentage of events (% obs.), the expected absolute number of events (N exp.) and the observed absolute number of events (N obs.)

Magn.	Fatl.	% exp.	% obs.	N exp	N obs
3	2	20.57	26.94	3565	4669
4	22	2.06	1.77	357	306
5	100	0.21	0.33	36	57
6	400	0.02	0.00	4	0
7	600	0.00	0.01	0	1

A plot of the expected and observed number of events for each magnitude is shown in Figure 4 (top), together with the log 10 of these numbers (bottom). If the magnitude distribution follows the Gutenberg-Richter law, the blue bars in the figure should follow the red line. Estimating the b value in Equation 1 using maximum likelihood results in a value of $b = 0.95 \pm 0.01$. Hence, this way of assigning magnitudes to fatalities gives a magnitude distribution that resembles the distribution of an earthquake catalog, which is assumed by the model.

Figure 2: Map of social conflicts in Africa in the period 2012-2017 with at least two fatalities. Events with more than 100 fatalities are colored red.

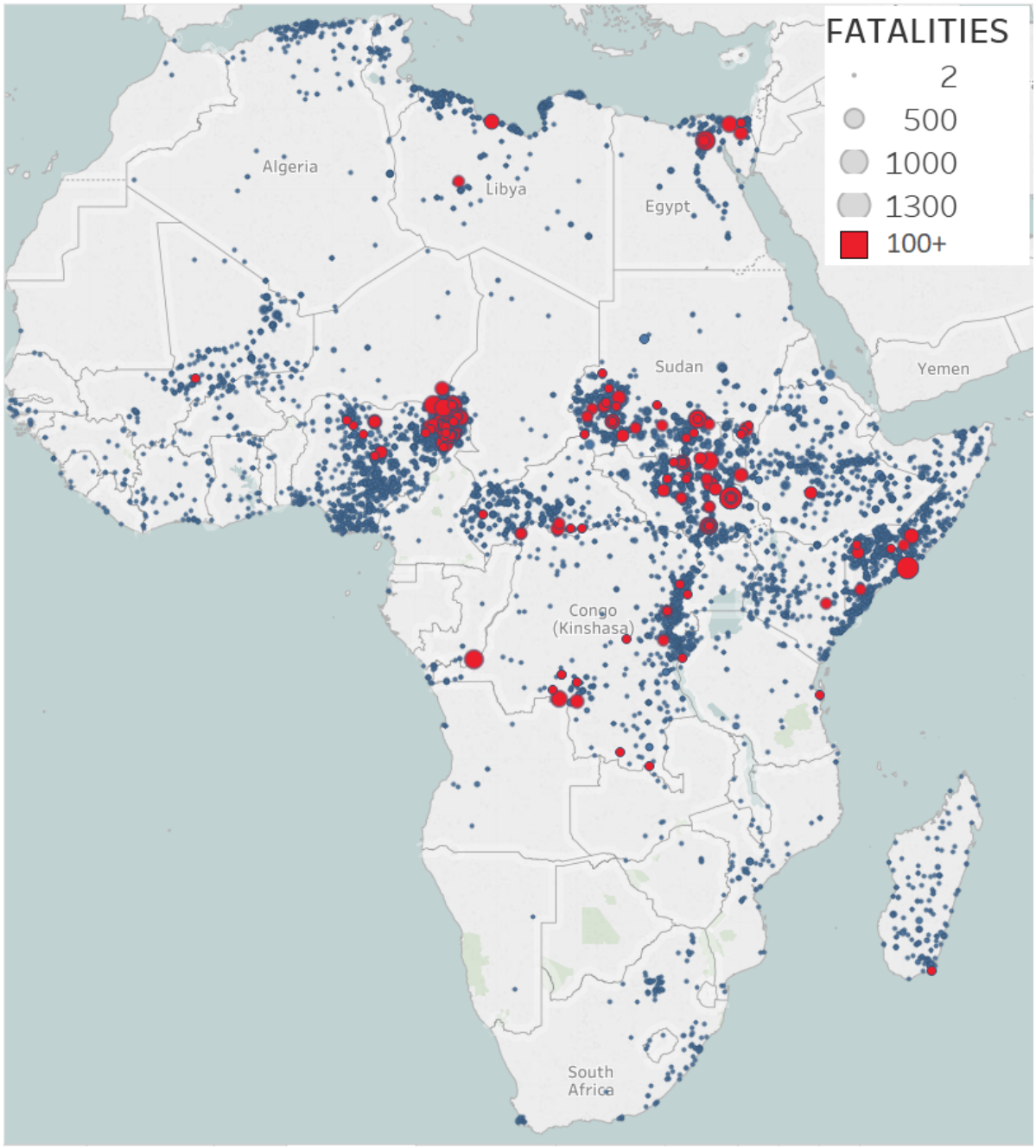


Figure 3: Plot of the latitude and longitude of events over time. The red lines represent the latitude and longitude of the geographical center of Africa. In the latitude plot, points above the red line are more north and points below are more south. In the longitude plot, points above the red line are more east and points below the red line are more west.

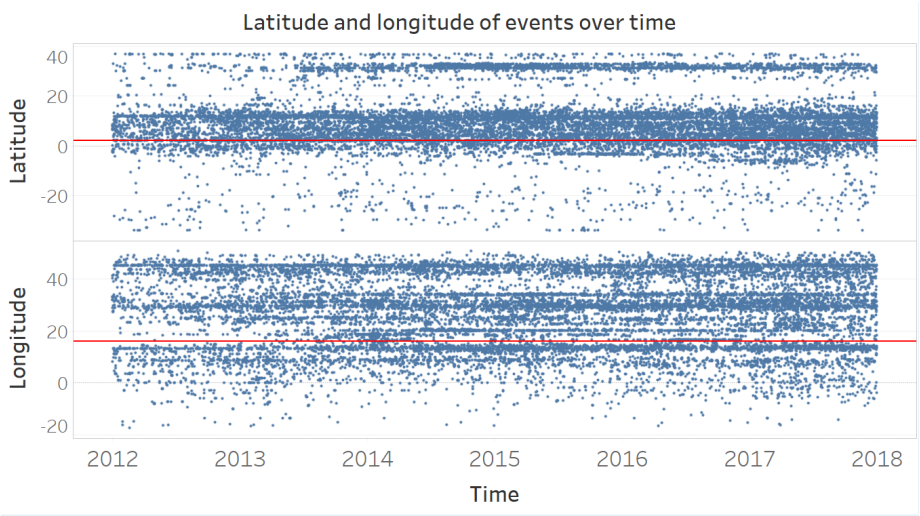
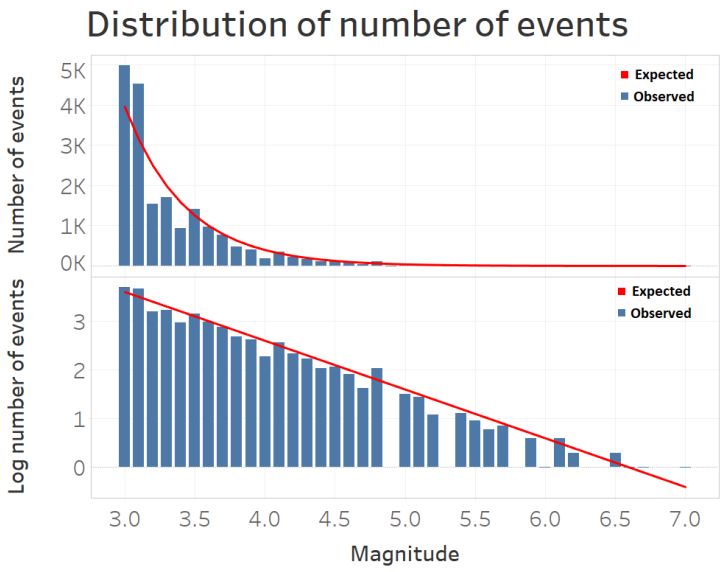


Figure 4: Distribution of the number of events. The blue bars in the top figure show the number of events for each magnitude and the red line is the number of events expected by the Gutenberg-Richter law. In the bottom figure, the logarithm base 10 of the same data is shown. If the magnitude distribution follows the Gutenberg-Richter law, the blue bars in the bottom figure should follow the red line.



3 Method

In this section, we will discuss the ETAS model used to study the social conflicts. This model is estimated using the simulated annealing algorithm proposed by Lombardi (2015). Furthermore, we will discuss several ways of testing the model and describe the way to make forecasts.

3.1 Model

ETAS model features
The model we will use to study the social conflicts originates from the study of earthquake occurrences. These have been modeled using a variety of point process models of which the Epidemic Type Aftershock Sequence (ETAS) models are the most popular. In these models, events are divided into two categories: background events and triggered events. Background events are events that occur spontaneously and triggered events occur as a result of other events. Each event can trigger new events, which in turn can again trigger new events. Zhuang et al. (2002) provide a list of common features of ETAS models regarding the occurrence of these events.

1. The occurrence rate of background events depends on location and magnitude, but is independent of time.
2. The magnitude of a background event is independent of its location.
3. Each event produces offspring events independently and the number of offspring events produced depends on the magnitude of the event.
4. The occurrence time of an offspring event depends only on the time difference from its ancestor and is independent of magnitude.
5. The location of an offspring event depends on the location of its ancestor.
6. The magnitude of an offspring event is independent of the magnitude of its ancestor.

One of the first models was the temporal ETAS model proposed by Ogata (1988) to model the origin times and magnitudes of earthquakes. This model was extended to also include the spatial aspect of earthquake occurrences by Ogata (1998) and Zhuang et al. (2002). This spatial-temporal ETAS model will be used to model the social conflicts. The occurrences of events can often be completely described by the conditional intensity function of the process, which gives the probability of an event with a certain magnitude occurring at a specific place at a specific time. This function often consists of two parts: one part describing the occurrence rate of background events and another part describing the way events trigger new events. The exact conditional intensity functions vary across studies.

Conditional intensity function
In this study we will use the model as defined in Lombardi (2017), which is similar to the one in Ogata (1998). The conditional intensity function in this model is given by

$$\lambda(t, m, x, y|H_t) = \left(\mu \cdot u(x, y) + \sum_{T_i<t} g(t, x, y;\ T_i, M_i, X_i, Y_i)\right) f(m) \quad (2)$$

where event i is described by a time T_i, location (X_i, Y_i) and magnitude M_i. It consists of three parts: the spatial background distribution $u(x, y)$, the decay function $g(t, x, y; H_i)$ and the magnitude probability density function $f(m)$. We will discuss them in reverse order.

The magnitude probability density function is based on the Gutenberg-Richter law and is given by

$$f(m) = \frac{\beta \cdot \exp[-\beta(m - M_c)]}{1 - \exp[-\beta(M_{max} - M_c)]}, \quad (3)$$

where $\beta = b \cdot \log(10)$ and M_{max} the magnitude of the largest event. The term M_c is the minimum magnitude for which the catalog is complete, i.e. the magnitude for which all events with magnitude $M \geq M_c$ are recorded.

The rate at which event i can trigger new events decreases with distance and time from event i. This is described by the decay function $g(t, x, y; H_i)$, which is given by

$$g(t, x, y; H_i) = \frac{k \cdot e^{\alpha(M_i - M_c)}}{(t - T_i + c)^p} \cdot \frac{c^i_{d,q,\gamma}}{[r_i^2 + d^2 \cdot e^{2\gamma(M_i - M_c)}]^q}. \quad (4)$$

Here, H_i is the history of events before event i. This function consists of two parts: the first part describes the temporal decay rate and the second part the spatial distribution. The temporal decay is governed by the parameters $\{\alpha, c, k, p\}$. The parameter α measures the influence of excess magnitude $(M_i - M_c)$ in generating offspring events and k is a normalizing constant for the number of offspring events. A small value for α implies that the effect of magnitude on triggering new events is small, such that smaller events can trigger larger events more easily, see Kumazawa et al. (2014). We assume that larger events cause more aftershocks than smaller events so that $\alpha > 0$. Since the number of offspring events cannot be negative, we have that $k > 0$ as well. It is hard to register all aftershocks after a large event and this incompleteness is measured by the parameter c. Finally, the parameter p is the rate of decay over time. We expect $p > 1$ since this implies that each event can only generate a finite number of offspring events.

The second part of Equation 4 determines the spatial probability distribution. The term r_i is the distance to event i and $c^i_{d,q,\gamma}$ is a normalization constant such

that the integral over the whole region equals one. The term $d^2 \cdot \exp[2\gamma(M_i - M_c)]$ measures influence of magnitude on the spatial extent of the aftershock region for event i. Similar to p, the parameter q measures the spatial decay rate. The parameters will be further investigated in a small simulation study later in this section. If we ignore any spatial components in the model and integrate over the whole region, this right term of Equation 4 becomes one and we obtain the decay function for the temporal model in Ogata (1988).

Spatial background distribution

The first term in Equation 2 consists of a parameter μ and the spatial background distribution $u(x, y)$. The parameter μ is the Poissonian rate of background events and the background distribution determines how these events are distributed over the region. The spatial background distribution is estimated following the procedure in Lombardi (2017), which is largely based on the iterative kernel method proposed by Zhuang et al. (2002). For this, the study region is first divided into N_c cells C_i $(i = 1, , 2, ..., N_c)$ to make a discrete approximation. The cells are of size 0.5 x 0.5 degree^2, which corresponds to an area of approximately 55 km x 55 km. The background rate can vary among cells but is assumed to be homogeneous within each cell. We can then write $u(x, y) = u_i/A_i$ where u_i is the probability of a background event occurring in cell i with area A_i.

The probabilities u_i are estimated using the iterative kernel method proposed by Zhuang et al. (2002), which uses the following function for the background distribution.

$$\mu(x, y) = \frac{1}{T} \sum_{j=1}^{N} (1 - \rho_j) \frac{1}{2\pi d_j} \exp\left(- \frac{(x - x_j)^2 + (y - y_j)^2}{2d_j^2} \right). \tag{5}$$

Here, T is the length of the study period, N the total number of events, ρ_j the probability that event j is a triggered event and d_j a variable bandwidth. The probability ρ_j of event j being a triggered event is given by the ratio between the intensity produced by all previous events and the total intensity at that time and place.

$$\rho_j = \frac{\sum_{i=1}^{j-1} g(t_j - t_i, x_j - x_i, y_j - y_i | m_i)}{\lambda(t_j, x_j, y_j | H_{t_j})}. \tag{6}$$

Finally, d_j is a variable bandwidth defined as the smallest disk around event j which includes at least $n_p = 10$ events, as in Zhuang et al. (2002). This way, events happening in rural areas have a farther reaching effect than events happening in dense areas, which makes regions with high and low event density more comparable. Following Lombardi (2015), we can use Equation 5 to estimate the probabilities u_i by

$$u_i = \frac{\mu(X_i, Y_i)}{\sum_{j=1}^{N_c} \mu(X_j, Y_j)}, \quad i = 1,\ 2,\ ...,\ N_c, \tag{7}$$

These probabilities then give the spatial background distribution, where the sum over all cells equals one. Together with the average background rate μ, the decay function $g(t, x, y; H_i)$ and the magnitude distribution $f(m)$ we have described the conditional intensity function of the process. This function will be used for calculating the log likelihood to estimate the parameters of the model.

Log Likelihood function
The log likelihood function is given by Vere-Jones (1970),

$$\begin{aligned} \text{Log L}(\theta|H_t) = &\sum_{i=1}^{N} \ln[\lambda(T_i, X_i, Y_i|H_{T_i})] \\ &- \int_T \iint_R \lambda(t, x, y|H_t) dt\, dx\, dy \end{aligned} \tag{8}$$

Here, T is the time range of the study period, R is the region the events take place and N is the total number of events in the catalog. For the calculation of the log likelihood, the background probabilities u_i from Equation 7 are needed for the cells in which an event has taken place. Hence, the parameters of the model are then $\{\mu, k, p, c, \alpha, d, q, \gamma; u_i, i = 1, ..., N_c'\}$, where N_c' is the subset of nonempty cells. These parameters can be estimated by maximizing the log likelihood function using the simulated annealing algorithm, which will be discussed in the next section.

Simulation study
Due to the high dimensionality of the likelihood function, there is a chance that a parameter might not be identified and becomes arbitrarily large or small during the estimation. Therefore, we will impose certain parameter restrictions on the 8 parameters in the model. These restrictions are based on the restrictions used for estimating earthquake occurrences and on values that are physically desirable, see Ogata (1998). These restrictions are given in Table 3

Table 3: Parameter restrictions in estimating the ETAS model. These restrictions are based on the restrictions used for estimating earthquake occurrences and on values that are physically desirable Ogata (1998).

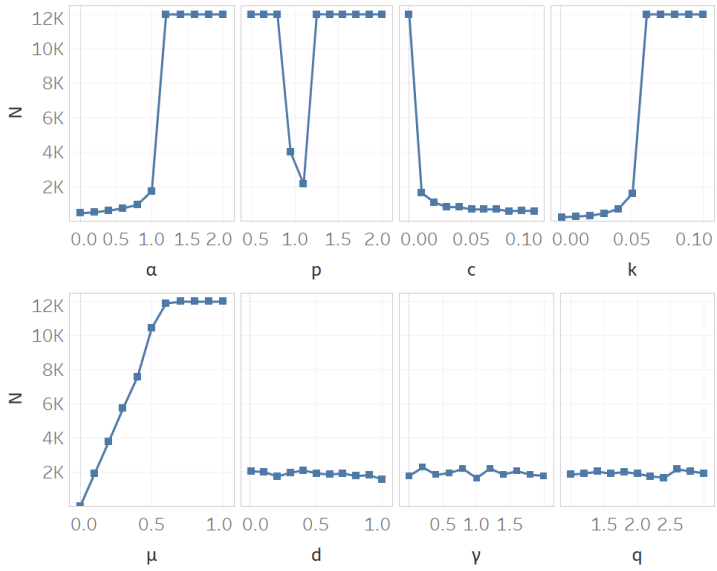
Parameter	Lower bound	Upper bound
μ	0	1
k	0.001	0.1
p	0.5	2
c	1.00e-5	0.1
α	0	2
d	0.01	1
q	1	3
γ	0	2

The branching ratio of the process gives the average number of events triggered by an event and is given by Lombardi (2017)

$$\phi = \int_0^{\infty} \int_{M_c}^{M_{max}} \frac{k \cdot e^{\alpha(M_i - M_c)}}{(t+c)^p} f(m)\, dm\, dt \tag{9}$$

If $\phi > 1$ or $p < 1$ events are generated faster than they die out and the process becomes explosive. To investigate the effects of the parameters, we will perform a small simulation study. For this, we first set the parameter values to $\mu = 0.1$, $k = 0.05$, $p = 1.1$, $c = 0.01$, $\alpha = 1$, $d = 1$, $q = 1.5$ and $\gamma = 1$, which are typical values in earthquake studies as in Ogata (1998). We will vary one parameter at a time and simulate 10 event catalogs and plot the median number of events against the parameter that is varied. Since the number of events can become explosive, an upper bound of 12,000 events is set during the simulations. For each parameter, we will increase its value from its lower to its upper bound in 10 steps. The results are given in Figure 5. We see that the four parameters in the temporal component, α, p, c and k, influence the number of events in the simulated catalogs. Furthermore, the parameter μ has a positive linear effect on the number of events and the parameter that determine the spatial decay, d, γ and q, have no effect on the number of events. This is as expected, since the four temporal parameters fully determine the branching ratio and the number of events generated. The spatial parameters only describes the distribution of these events over the region. We see that a higher value for α and k lead to more events and higher value for c leads to fewer events. For the temporal decay rate p there is a small region in which the process is not explosive. However, the size of this region varies depending on the values of the other parameters.

Figure 5: Results of the simulation study to investigate the effect of the parameters on the number of events. The parameters are varied one at a time and $N_{SIM} = 10$ catalogs are simulated based on those parameter values. The median number of events is the plotted against the parameter values. The parameters are varied from their lower to upper constraint in 10 steps. Since the number of events can get infinite, we set an upper bound of 12,000 events during the simulations.



3.2 Estimation method

The parameters will be estimated using a simulated annealing algorithm , which is a method for approximating the global optimum of a given function. It originates from annealing in metallurgy, describing the physical process of reducing defects by heating and then slowly cooling the material. Even though this method is unlikely to find the optimal solution, it can often find a very good approximation of the optimum. In particular, it can be preferable over (quasi) Newton methods in situations with a large number of independent variables. Since the algorithm finds a good solution in a relatively short amount of time, we can run it many times, providing a probability distribution for the parameters and the background distribution. These can be used to evaluate model uncertainties. For each dataset, we will run the algorithm $N_{SIM} = 100$ times.

The algorithm starts by generating a random solution, after which it searched for a new point in that neighborhood. This search is based on a probability distribution that depends on the so called 'temperature' of the process. The method accepts all points that raise the objective function (in case of maximization) and also accepts points that lower the objective, with a certain probability. That way it avoids being trapped in local maxima. By decreasing the temperature,

the probability of accepting a worse solution is lowered as the solution space is explored.

More formally, the simulated annealing procedure used to estimate the ETAS models is the following, as described in Lombardi (2015), which consists of an initialization, a loop and a cooling scheme.

Initialization
Set the count $k = 0$ and generate a random starting point $\vec{\theta_{opt}}$ and set the initial temperature $T_0 > 0$. The starting temperature must be high enough that any solution can be selected and is calculated based on the function to be optimized, see Lin et al. (1993).

Loop
The loop is a random search for a better solution around the local maximum. First, set $\vec{\theta_k} = \vec{\theta_{opt}}$ and generate the next candidate $\vec{y}$ from a multi-dimensional Cauchy distribution G.

$$\vec{y} = G(\vec{\theta_k})$$

We move to this new point with a certain probability. Sample $p \in [0,\ 1]$ uniformly and move to the newly generated point if $p \leq A(\vec{\theta_{opt}}, \vec{y}, T_k)$, where the acceptance function A is the Metropolis criterion in Sen and Stoffa (2013)

$$A(\vec{\theta_{opt}}, \vec{y}, T_k) = min \left\{ 1, \exp(\frac{f(\vec{y}) - f(\vec{\theta_{opt}})}{T_k}) \right\} \tag{10}$$

If the log-likelihood at this new point is larger, we set this new point as the current optimum.

$$\text{If } \ \text{LogL}(\vec{y}|H_t) > \text{LogL}(\vec{\theta_{opt}}|H_t) \rightarrow \vec{\theta_{opt}} = \vec{y}$$

Cooling scheme
Once a new optimum is found, the temperature is lowered according to the cooling schedule in Ingber (2000) and Lombardi (2017).

$$T_k = T_0 \exp\left[-13.8 \cdot \exp(-\frac{3.4}{D}) \cdot k^{1/D} \right],$$

where $D = 8$ is the number of parameters in the model. The two constants -13.8 and 3.4 were found to provide a fast algorithm by a simulation study in Lombardi (2017). After decreasing the temperature, the loop is repeated until convergence. Events that took place before the study period can still have an effect on events inside the study period. We will therefore use a one year burn-in period when estimating the model. Hence, the estimation period consists of five years from 2012 to 2016, where 2012 will be used as a 'learning' period and the four other years, 2013-2016, will be used as the study period. The data of the last year, 2017, will be used for evaluating forecast performance.

3.3 Model testing

The performance of the model will be tested by performing a residual analysis and a number-of-events test, which will be discussed in this section.

3.3.1 Residual analysis

If the model describes the dynamics of the conflicts well, the residuals are expected to follow a stationary Poisson process with unit rate. The residuals are obtained by a transformation of the time axis, as in Ogata (1998).

$$\tau = \Lambda(t) = \int_0^t \iint_R \int_{M_c}^{M_{max}} \lambda(t', x, y, m|H_{t'}) dt'\, dx\, dy\, dm \tag{11}$$

These residuals give the expected number of events with magnitude larger than M_c up to time t and in the region R. If the residuals follow a stationary Poisson process with rate one, the inter-event times follow an exponential distribution. This can be tested using a one-sided Kolmogorov-Smirnov test (KS-test). We will also perform a Wald− Wolfowitz runs test, which tests the null hypothesis that the inter-event times are stationary and not autocorrelated.

Furthermore, we can perform a visual check by plotting the number of events expected by the model against the observed number of events. This will be done for the transformed times, all events, background events and triggered events. The expected number of events are calculated by integrating the intensities over time, magnitude and space. The expected number of total events is given by

$$N_{exp}^{All}[\theta|H_t] = \int_T \iint_R \lambda(t, x, y|H_t) dt\, dx\, dy\,, \tag{12}$$

the expected number of background events is given by

$$N_{exp}^{Back}[\theta|H_t] = \int_T \iint_R \mu \cdot u(x, y) dt = \mu(T_2 - T_1), \tag{13}$$

and the expected number of triggered events by

$$\begin{aligned} N_{exp}^{Trig}[\theta|H_t] &= N_{exp}^{All}[\theta|H_t] - N_{exp}^{Back}[\theta|H_t] \\ &= \int_T \iint_R \sum_{T_i<t} \frac{k \cdot e^{\alpha(M_i - M_c)}}{(t - T_i + c)^p} \cdot \frac{c^i_{d,q,\gamma}}{[r_i^2 + d^2 \cdot e^{2\gamma(M_i - M_c)}]^q} dt\, dx\, dy \end{aligned} \tag{14}$$

Here, T is the time window of events, R the region and T_1 and T_2 are the start and end times of the study period. Note that we also integrate over

the magnitude distribution, but since the magnitudes are independent of the parameters, this integral equals one.

The observed number of all events (background + triggered) can be obtained directly from the catalog. Since we cannot know for certain whether an event is a background or triggered event, we cannot count them. Therefore, the number of observed background and triggered events are calculated as the sum of the probabilities of being a background or triggered event. The probability that event i is a background event is given by Lombardi (2017)

$$pr_i^B = \frac{\mu \cdot u(X_i, Y_i)}{\lambda(T_i, X_i, Y_i | H_{T_i})} \tag{15}$$

and the probability that event i is a triggered event is given by $pr_i^T = 1 - pr_i^B$. The sum of these probabilities over all events determine the 'observed' number of background and triggered events.

3.3.2 Number of events test

The number-of-events test, or N-test, compares the number of events in the conflict catalog with the number of events expected by the model. The expected number of events and its distribution are obtained by simulating a large number of event catalogs and calculating the number of events in each of them. Specifically, this is done in the following way.

Number of events test

1. Simulate $N_{SIM} = 100$ catalogs based on the estimated ETAS model.
2. Fit a normal distribution to the number of events in each simulated catalog.
3. Calculate the median and the 95% confidence bounds.
4. Compute the probability that we observe more events than in our dataset.

The simulation method is described briefly in the next section. We then compare the number of events in the conflict dataset with this distribution. If they differ too much, we then reject the model.

3.4 Forecasting

Finally, we will test the performance of the model by making a forecast for the number of events and their location in the year after the study period. We focus on events with magnitude $M_f \geq 4$, which corresponds to 22 or more fatalities. For this, we simulate $N_{SIM} = 100$ catalogs using the estimated ETAS model and calculate the expected number of events with magnitude $M_f \geq 4$ in each cell of the study region in the year 2017. The expected number of events in simulated catalog i in cell C_j is given by Lombardi (2017)

$$f_{ij} = \int_T \iint_{C_j} \int_{M_f}^{M_{max}} \lambda(t, x, y, m|H_t^i) dm\, dx\, dy\, dt, \tag{16}$$

where H_t^i is the history of the process up to time t for catalog i. For each cell we then take the median number of events of all simulated catalogs as the forecast. The total number of events can then be calculated as the sum of the median number of events over all cells. The forecast number of events and locations will then be compared with their observed number and locations.

4 Simulation

4.1 Simulation procedure

Before we apply the model to the conflict data, we will first estimate a simulated catalog. The simulation for the time of events is based on the thinning method proposed by Zhuang et al. (2002) and the locations and magnitudes are simulated following the procedure by Lombardi (2017). This procedure will be used to simulate the data in this section, as well as providing simulated catalogs for the N-test.

4.2 Estimation results

We successfully estimated the spatial-temporal ETAS model on a simulated catalog. For this, the model was estimated $N_{SIM} = 100$ times. The parameter estimates are given in Table 4, together with their true values and distributions. The estimated spatial background distribution is shown in Figure 6, which also includes a map with the locations of the conflicts in the catalog. This is the background distribution for the best model, i.e. the run with the highest log likelihood. The study area is divided in $N_C = 1804$ cells with size 0.5 x 0.5 degree^2, which corresponds to an area of approximately 55 x 55 km. In each cell, the probability of a background event occurring in that cell is given on a logarithmic scale. When comparing this spatial distribution with the map of actual events taken place, we see that areas with high probability correspond to areas with a high event density. This suggests that the spatial background distribution gives a good indication of where events are likely to take place.

The parameter estimates in Table 4 show that they are close to the true parameter values and that the true parameters are well within the 95% confidence bounds. Moreover, the median value of the $N_{SIM} = 100$ runs is also close. This suggests that the model performs well in estimating the parameters. Furthermore, the distributions tend to be symmetrical, except for the parameter d, which is close to its upper bound of one. In this section, we will only focus on whether the parameter estimates are close to their true values. More in-depth discussion about the parameter interpretations is given when we apply the model to real conflict data.

Figure 6: Map of the estimated spatial background distribution (left) and the simulated event catalog (right). In the left figure the spatial background distribution corresponding to the run with the highest log likelihood is shown. The study area is divided into $N_c = 1804$ cells of 0.5 $degree^2$, which corresponds to cells of size 55 x 55 km. Each cell gives the probability of a background event occurring in that cell. In the right figure, the events of the simulated catalog are shown, where the size of the dots correspond to the magnitude of the events. If the spatial background distribution is estimated correctly, the red areas should overlap with areas where more events take place.

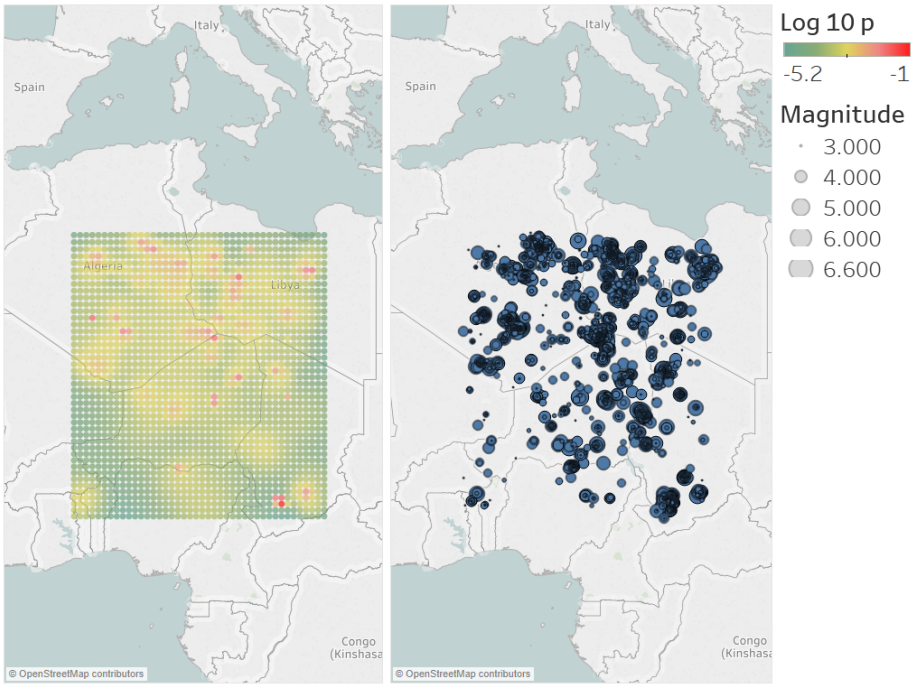
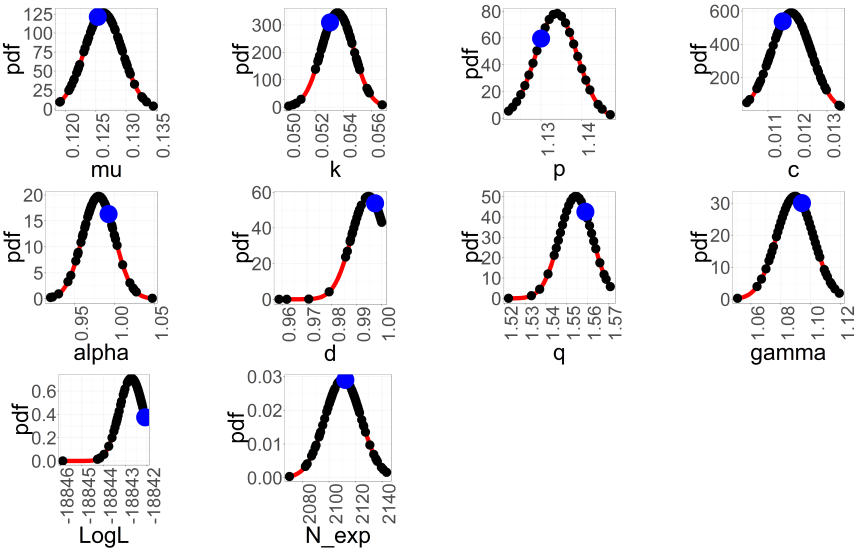


Table 4: Parameter estimates for the simulated catalog.
Top: Parameter estimates for the 8 parameters of the simulated catalog, the log likelihood and the number of events. **Bottom:** Parameter distributions. The black dots are the estimated parameters in each of the $N_{SIM} = 100$ runs of the simulated annealing algorithm. The blue dot represent the estimated parameters of the best model, i.e. the run with the highest log likelohood. The red curve is a normal distribution fitted to the data for visual comparison of the results.

Parameter	Best value	True value	Median	95% Confidence Interval
μ	$1.25 \cdot 10^{-1}$	$1.00 \cdot 10^{-1}$	$1.23 \cdot 10^{-1}$	$[\ 1.20 \cdot 10^{-1}\ ,\ 1.33 \cdot 10^{-1}\ ]$
k	$5.31 \cdot 10^{-2}$	$5.00 \cdot 10^{-2}$	$5.50 \cdot 10^{-2}$	$[\ 5.07 \cdot 10^{-2}\ ,\ 5.58 \cdot 10^{-2}\ ]$
p	1.13	1.10	1.14	$[\ 1.12\ ,\ 1.15\ ]$
c	$1.15 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	$1.21 \cdot 10^{-2}$	$[\ 1.06 \cdot 10^{-2}\ ,\ 1.30 \cdot 10^{-2}\ ]$
α	$9.92 \cdot 10^{-1}$	1.00	$9.60 \cdot 10^{-1}$	$[\ 9.30 \cdot 10^{-2}\ ,\ 1.02\ ]$
d	$9.97 \cdot 10^{-1}$	1.00	$9.95 \cdot 10^{-1}$	$[\ 9.71 \cdot 10^{-1}\ ,\ 1.00\ ]$
q	1.56	1.50	1.55	$[\ 1.54\ ,\ 1.57\ ]$
γ	1.09	1.00	1.08	$[\ 1.07\ ,\ 1.11\ ]$
LogL	$-1.88 \cdot 10^4$	-	$-1.88 \cdot 10^4$	$[-1.88 \cdot 10^4\ ,-1.88 \cdot 10^4\ ]$
N_{events}	$2.12 \cdot 10^3$	$2.11 \cdot 10^3$	$2.12 \cdot 10^3$	$[\ 2.08 \cdot 10^3\ ,\ 2.14 \cdot 10^3\ ]$



4.3 Model testing

In this section we will discuss the residual analysis and the N-test.

4.3.1 Residual Analysis

To analyze the performance of the model, we will perform the residual analysis as discussed in section 3.3.1. The results of the KS-test and the runs test are given in Table 5: they both fail to reject their null hypothesis. This suggests that the transformed times are drawn from a Poisson distribution and that the model describes the temporal dynamics of the data well. Furthermore, the runs test suggests that the inter-event times are stationary and not autocorrelated.

Table 5: Results of the one-sided Kolmogorov-Smirnov (KS-1) test and the Wald-Wolfowitz RUNS test. The p-values of each test are given.

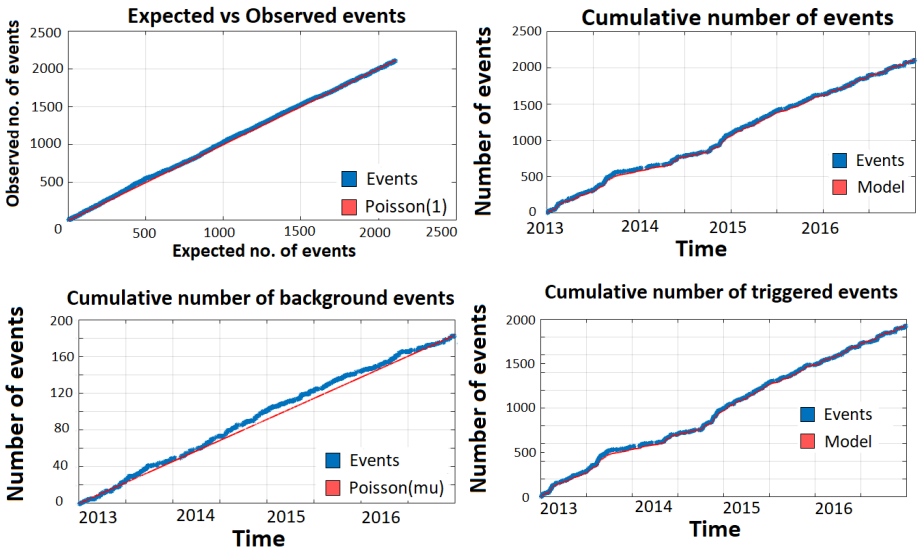
Catalog	KS-test	RUNS test
Simulated	0.65	0.88

The observed number of events are compared to the expected number of events by the model. This is done for all events, background events and triggered events. The total observed and expected numbers are given in Table 6, where we can see that all numbers lie very close to each other. The number of events over time is displayed in Figure 7. In the top left figure the observed number of events is plotted against the transformed times from equation 11. We see that the transformed times lie very close to the red line, which follows a Poisson process with unit rate. In the top right figure, the observed number of events (blue) is plotted against the expected number of events by the model (red). These are calculated using Equation 12. Furthermore, in the bottom two figures, the observed number of background and triggered events are compared to their expected number. The expected number of background and triggered events are calculated using Equation 13 and 14 respectively. We see that also over time, the expected and observed number of events remain very close. These results suggest that the temporal behavior of the data is well described by the model.

Table 6: Total number of expected and observed events, background events and triggered events for the simulated catalog. The expected numbers of events are calculated using the equations in section 3.3.1.

	All		Background		Triggered	
	exp	obs	exp	obs	exp	obs
Sim	2112	2105	183	183	1930	1922

Figure 7: Results of the residual analysis for the simulated catalog. In the top left figure the observed number of events is plotted against the transformed times from equation 11. In the top right figure, the observed number of events (blue) is plotted against the expected number of events by the model (red). These are calculated using Equation 12. In the bottom two figures, the observed number of background and triggered events (blue) are compared to the number of events expected by the model (red).



4.3.2 Number of events test

For the N-test, $N_{SIM} = 100$ catalogs were simulated and for each catalog the number of events is calculated. The median and distribution are given in Table 7, together with the observed number of events in our catalog and the probability of observing more events than expected by the model. We see that the median lies close to the observed number of events and that the observed number lies well within the 95% confidence bounds. The probability that we observe a higher number of events than in our catalog is 0.63, which indicates the number of events in the catalog is close to what we would expect from this model.

Table 7: Results for the Number of Events test, or N-test, for the simulated catalog. The estimated ETAS model is used to simulate $N_{SIM} = 100$ catalogs. The median and 95% confidence interval for are calculated for the simulated catalogs, as well as the relative difference (% dif.). Furthermore, the probability (Prob.) that we observe a number of events equal to or higher than the event catalog is given. A low probability means that we would expect more events from this model and a high probability means we would expect less events from this model.

			N-test			
Catalog	Median	Observed	% dif.	95% Confidence Interval		Prob.
Simulated	2305	2105	9.5%	[1166,	3710]	0.63

4.4 Forecasting results

Finally, we will test the performance of the model by making a forecast of the number and location of events in the year following the study period. We will focus on events with magnitude $M_f \geq 4$ in the year 2017. For this, $N_{SIM} = 100$ catalogs are simulated and in each cell of the study region, the median expected number of events with magnitude $M_f \geq 4$ is calculated using Equation 16. The total expected number of events in 2017 is calculated as the sum of the expected number of events of all cells. The forecast number of events with magnitude $M_f \geq 4$ are given in Table 8. This is compared to two naive forecasts: the average number of events with $M \geq 4$ over the past five years and the number of events in the last year before the forecasting period. Furthermore, the actual number of events is given. We see that the model gives a good estimate, even though it is slightly above the actual number. In this case, the naive forecasts and the model forecast lie very close to each other.

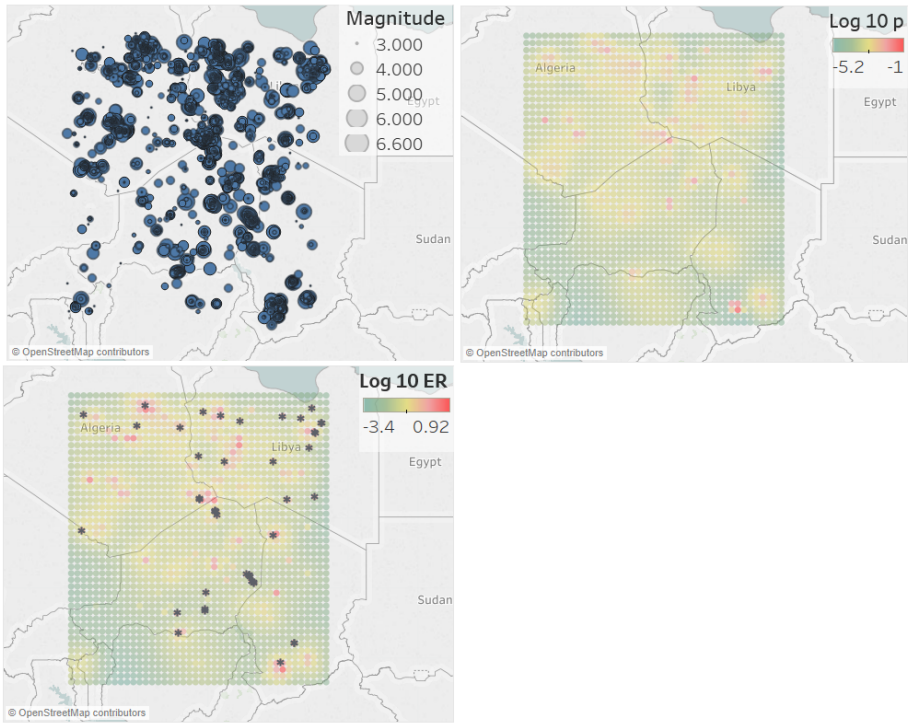
Table 8: Forecasts for the number of events with magnitude $M_f \geq 4$ in 2017. The average number of these events in the period 2012-2016 is given, the number in 2016, the number predicted by the model and the observed number of events in 2017.

Region	Average	Last year	Model	Observed
Simulated	45.4	49	56	47

The predicted locations are given in Figure 8. The top two figures are the event catalog and the spatial background distribution as in Figure 6. These will be used as the naive forecast. In the bottom left figure, the expected number of events in each cell is given for the year 2017, along with the actual locations of events that have taken place. We see that the forecast provides a good indication of where the actual events have taken place. The similarity between the forecast map and the spatial background distribution suggests that this background distribution plays a large role in making the forecast.

The results from estimating the simulated catalog show that the model and estimation method can correctly estimate the model parameters. Furthermore, the tests suggest that the model is a good fit and can be used to make forecasts

Figure 8: Three maps for the simulated catalog. Top left: simulated event catalog for the period 2012-2016. Top right: spatial background distribution estimated from the data from 2012-2016. Bottom left: forecast for the expected number of events with $M \geq 4$ in each cell in the year 2017, together with the locations of the actual conflicts that have taken place.



for future events. In the next section, we will apply the model to real conflict data and compare the results between the major geographical regions in Africa.

5 Application to social conflict data

In this section, we will apply the model to real conflict data in Africa and make predictions of future events.

5.1 Data description

The continent of Africa is often divided in five major geographical regions. These are North-, West-, Central-, East- and South Africa. These regions are displayed in Figure 9. The number of events in each region for the learning period, study period, forecasting period and in total, is given in Table 9. Some summary statistics regarding the number of fatalities and magnitudes are given in Table 10. This last table shows that the size of events is quite similarly distributed in each region, with some difference regarding the largest event. Since the number of events in the region South Africa is too small to reasonably estimate the model, we will restrict the study to the other four major regions.

Figure 9: The five major regions of Africa: North, West, Central, East and South. Source: Wikimedia Commons, the free media repository.

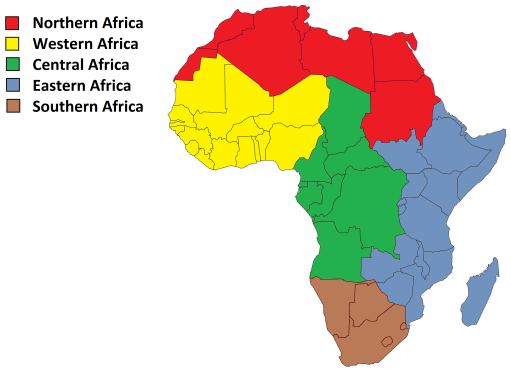


Table 9: Number of conflict events in each region for the learning period (2012), study period (2013-2016) and forecasting period (2017) and in total (2012-2017).

	North	West	Central	East	South	Total
N learning period	291	378	203	708	40	1620
N study period	3518	2164	1834	4617	117	12250
N forecasting period	612	617	589	1630	16	3464
N total	4421	3159	2626	6955	173	17334

Table 10: Summary statistics of the magnitudes and number of fatalities for each region. The columns are the minimum (Min.), the first quantile (1st Qu.), the median (Med.), Mean, third quantile (3rd Qu.) and the maximum (Max.).

	Region	Min.	1st Qu.	Med.	Mean	3rd Qu.	Max.
Magnitude							
	North	3.0	3.0	3.2	3.4	3.5	6.1
	West	3.0	3.1	3.2	3.4	3.7	7
	Central	3.0	3.1	3.2	3.3	3.5	6.1
	East	3.0	3.0	3.2	3.3	3.5	6.7
Fatalities							
	North	2	2	4	10	10	411
	West	2	3	5	13	12	600
	Central	2	3	4	9	10	420
	East	2	2	4	9	10	598

Since the model assumes the magnitude probability density function based on the Gutenberg-Richter law in Equation 3, it is good to check whether this assumption still more or less holds for the catalogs in each region. If it deviates too much, we might assume the wrong magnitude distribution for the data. For this, we estimate the parameter b in Equation 1 for each region, which are given in Table 11. Since the value of b is still close to one for each region, the magnitude probability density function in the conditional intensity function in Equation 2 still holds.

Table 11: Maximum likelihood estimation of the parameter b in the Gutenberg-Richter law in Equation 1. The b column is the estimate with SE the standard error.

Region	b	SE
North	0.956	0.025
West	0.854	0.024
Central	0.990	0.034
East	1.006	0.022

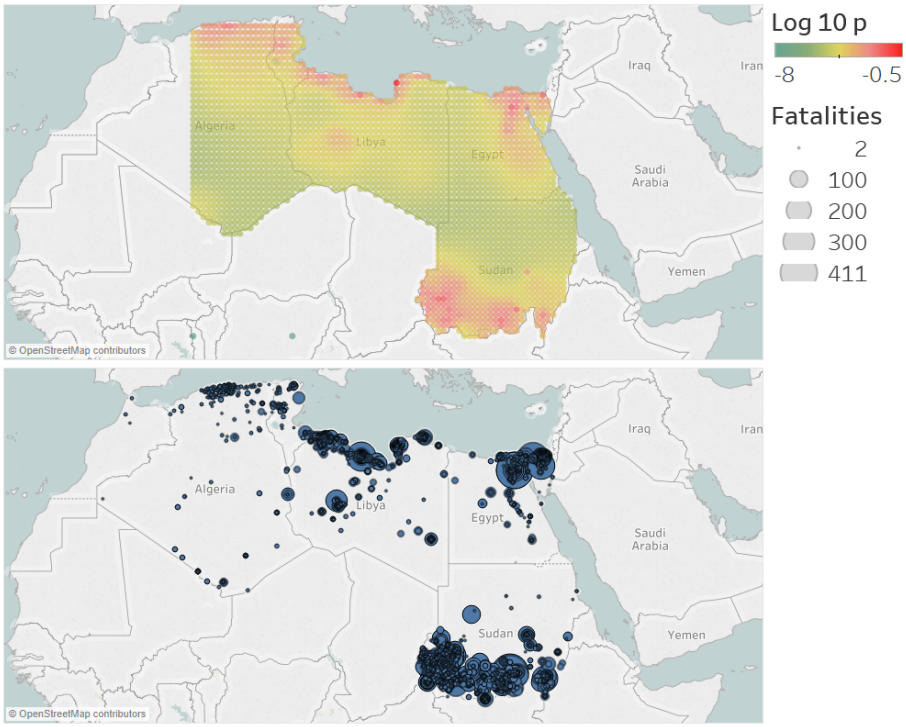
Due to a bug in the software, events with a negative longitude cause an error. We will therefore restrict our study region to the right side of the prime meridian. This means that out of a total of 17,829 events, 495 events will be omitted in North- and West Africa. This should not affect the estimation results in a large way since this number of events is relatively small.

5.2 Background distributions

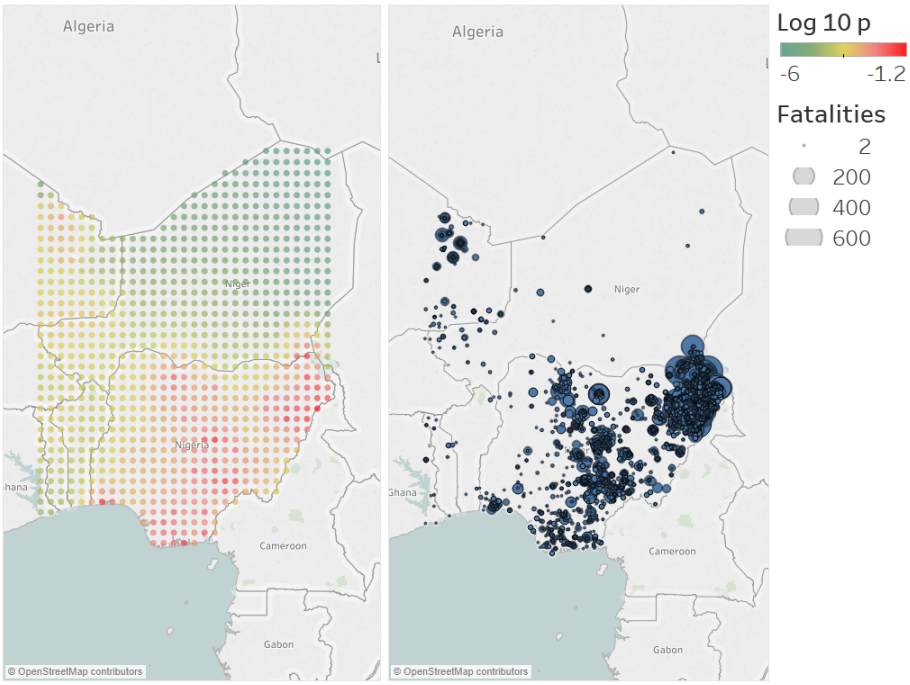
For each region, the study area is determined by the smallest rectangular area covering all events. The spatial background distribution is estimated for this area. Since this means that it can cover areas that are not in the study region,

such as other countries or parts of the ocean, the cells outside the study region are removed after estimation. The effect on the total probability should be minimal, since these areas have a very low probability due to the lack of events. The estimated spatial background distributions are given in Figure 10 for North and West Africa and in Figure 11 for Central and East Africa. Furthermore, a map with the actual conflicts in each region is given next to the map of the background distribution. We see that the estimated distributions are in line with what we would expect from the catalogs, since areas with high probability correspond with areas with high event density.

Figure 10: Spatial background distributions of the models with the largest log-likelihood for the regions North- and West Africa. Each study area is divided into $N_c = 4070$ and $N_c = 1102$ number of cells for the North and West region respectively. In these cells, the background intensity is assumed to be homogeneous. The cells give the probability of a background event occurring in that cell. First, the background distribution is estimated for the smallest rectangle covering all events. Then, cells outside the study region are removed. These cells have very little impact on the total probability, so the sum of all cells nearly equals one.

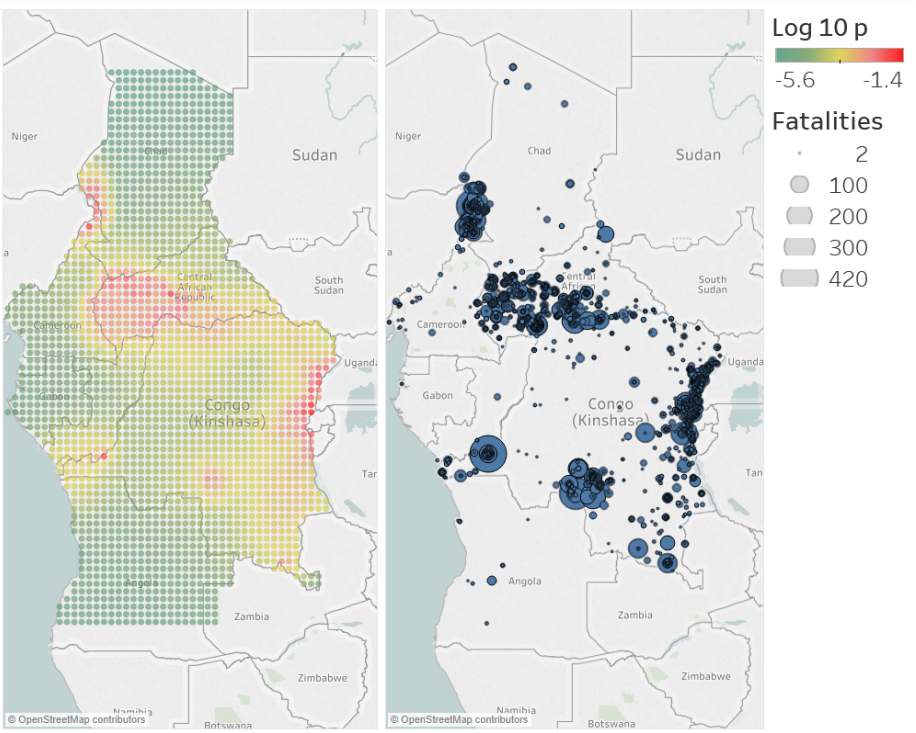


(a) Spatial background distribution for the region North Africa.

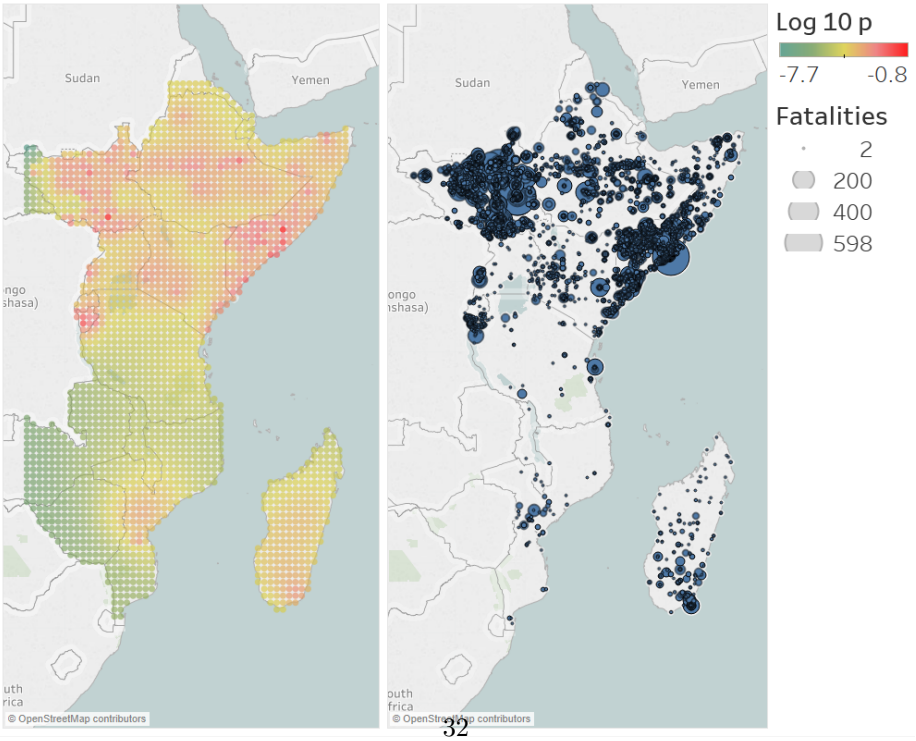


(b) Spatial background distribution for the region West Africa.

Figure 11: Spatial background distributions of the models with the largest log-likelihood for the regions North- and West Africa. Each study area is divided into $N_c = 3330$ and $N_c = 4368$ number of cells for the Central and East region respectively. In these cells, the background intensity is assumed to be homogeneous. The cells give the probability of a background event occurring in that cell. First, the background distribution is estimated for the smallest rectangle covering all events. Then, cells outside the study region are hidden. These cells have very little impact on the total probability, so the sum of all cells nearly equals one.



(a) Spatial background distribution for the region Central Africa.



(b) Spatial background distribution for the region East Africa.

5.3 Parameter estimates

For each region we have estimated the 8 parameters of the spatial-temporal ETAS model $N_{SIM} = 100$ times. The results for the North, West, Central and East Africa catalog are given in Tables 13, 14, 15 and 16 respectively. In these tables, the parameter estimates of the best model, their median and distributions are given. Estimates with a '$*$' are estimates that lie on, or are very close to, the boundary of their constraint. We see that this is the case for the parameters c, d, q and γ in each region, where c, d and γ go to their lower bound and q goes to its upper bound. The parameter c measures the incompleteness of the catalogs after a large event. Since the conflicts are reported on a daily basis, the next triggered event occurs the next day or later. It then makes sense that there are no undetected aftershocks right after an event. Furthermore, a value of $c = 0$ would not change the shape of the likelihood function by much. The term $d^2 \cdot \exp[2\gamma(M_i - M_c)]$ measures influence of magnitude on the spatial extent of the aftershock region for event i. Since d goes to its lower bound and γ tends to become arbitrarily small, it suggests that magnitude has little influence on the region of aftershocks. If we decrease the lower bound of the constraint for parameter d it also tends to become arbitrarily small. In that case, the only influence on the spatial decay is the distance from the event. This is a very counter intuitive result, since we would expect that a very large conflict has farther reaching influence than a small conflict.

The other parameters, μ, k, p and α, vary across the regions and can be used to make a comparison. For easy comparison, they are displayed together in Table 12. What is most notable is that $p \leq 1$ in all regions. This implies that the process is explosive: each event generates an infinite amount of triggered events when $t \to \infty$. However, this effect is not very noticeable at this small time scale. Simulating catalogs with these values for p produces a finite number of events, which is comparable to the number of events in the conflict catalogs.

The value for the background rate μ in each region corresponds to the number of background events, which is given in Table 17 in the next section. When we divide the number of background events by the number of days, we obtain an average of 0.47, 0.48, 0.36 and 0.86 background events per day for the regions North, West, Central and East respectively. Hence, the estimates for μ resemble the number of background events in the catalog. However, this might also be the case because μ is used for calculating the number of background events in Equation 15. The parameter k governs the expected number of direct aftershocks caused by an event. These results suggest that events cause the most aftershocks in the North region and the least in the West region by comparison. The parameter α measures the efficiency of magnitude in generating aftershocks. A small value indicates that the triggering rate is less dependent on magnitude. Hence, smaller events can trigger larger events more easily if α is small, see Kumazawa et al. (2014). This suggests that in the East region, small conflicts can trigger larger conflicts easier compared to the other three regions.

Table 12: Parameter estimates for the parameters μ, k, p and α of the ETAS model. The values are the values from the 'Best value' columns in tables 13, 14, 15 and 16. Furthermore, the number of events in the study period is given.

Region	μ	k	p	α	N events
North	0.460	0.0376	0.746	0.139	3518
West	0.502	0.0226	0.701	0.246	2164
Central	0.379	0.0352	0.795	0.303	1834
East	0.838	0.0307	0.714	0.0893	4617

Table 13: Parameter estimates North Africa.
Top: Estimation results for the 8 parameters of the ETAS model, the log likelihood and the number of events for the North Africa conflict catalog. Estimates marked with a $'*'$ are at, or very close to, their constraint values. The best estimate, i.e. the estimate of the run with the highest log likelihood, is given for each parameter. Furthermore, the median and the 95% confidence intervals are given. **Bottom:** Parameter distributions for the 8 parameters of the ETAS model, the log likelihood and the number of events. The black dots are the estimated parameters in each of the $N_{SIM} = 100$ runs of the simulated annealing algorithm. The blue dot represent the estimated parameters of the best model, i.e. the run with the highest log likelohood. The red curve is a normal distribution fitted to the data for visual comparison of the results.

Parameter	Best value	Median	95% Confidence Interval
μ	$4.60 \cdot 10^{-1}$	$5.05 \cdot 10^{-1}$	$[\ 4.17 \cdot 10^{-1}\ ,\ 5.45 \cdot 10^{-1}\ ]$
k	$3.76 \cdot 10^{-2}$	$3.58 \cdot 10^{-2}$	$[\ 2.57 \cdot 10^{-2}\ ,\ 6.13 \cdot 10^{-1}\ ]$
p	$7.46 \cdot 10^{-1}$	$7.43 \cdot 10^{-1}$	$[\ 6.79 \cdot 10^{-1}\ ,\ 8.53 \cdot 10^{-1}\ ]$
c	$1.00 \cdot 10^{-5}$ *	$1.04 \cdot 10^{-5}$	$[\ 1.00 \cdot 10^{-5}\ ,\ 9.91 \cdot 10^{-2}\ ]$
α	$1.39 \cdot 10^{-1}$	$2.48 \cdot 10^{-1}$	$[\ 5.21 \cdot 10^{-3}\ ,\ 5.37 \cdot 10^{-1}\ ]$
d	$1.00 \cdot 10^{-2}$ *	$1.00 \cdot 10^{-2}$	$[\ 1.00 \cdot 10^{-2}\ ,\ 1.00 \cdot 10^{-2}\ ]$
q	3.00 *	3.00	$[\ 2.98\ ,\ 3.00\ ]$
γ	$6.94 \cdot 10^{-4}$ *	$3.05 \cdot 10^{-3}$	$[\ 1.58 \cdot 10^{-5}\ ,\ 9.04 \cdot 10^{-3}\ ]$
LogL	$1.33 \cdot 10^{3}$	$1.31 \cdot 10^{3}$	$[\ 6.36 \cdot 10^{2}\ ,\ 1.33 \cdot 10^{3}\ ]$
N_{events}	$3.56 \cdot 10^{3}$	$3.63 \cdot 10^{3}$	$[\ 3.35 \cdot 10^{3}\ ,\ 3.75 \cdot 10^{3}\ ]$

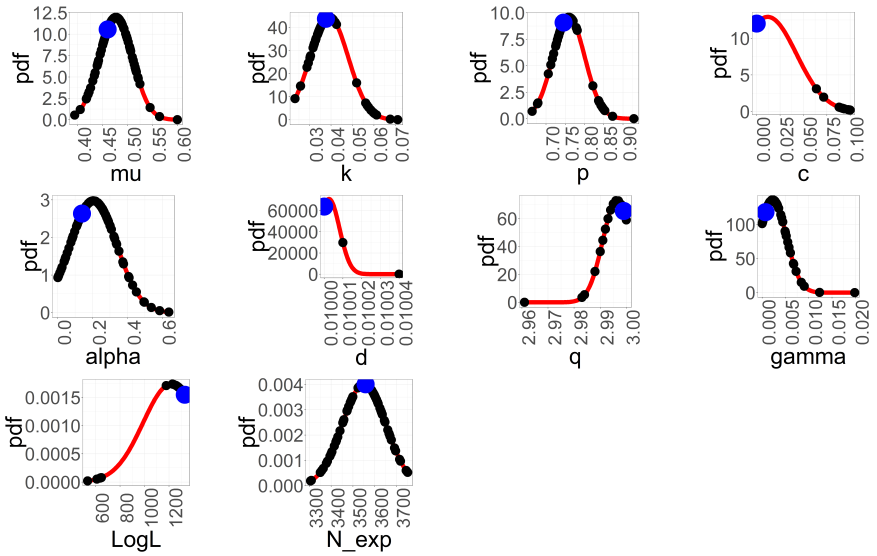


Table 14: Parameter estimates West Africa.
Top: Estimation results for the 8 parameters of the ETAS model, the log likelihood and the number of events for the West Africa conflict catalog. Estimates marked with a $'*'$ are at, or very close to, their constraint values. The best estimate, i.e. the estimate of the run with the highest log likelihood, is given for each parameter. Furthermore, the median and the 95% confidence intervals are given. **Bottom:** Parameter distributions for the 8 parameters of the ETAS model, the log likelihood and the number of events. The black dots are the estimated parameters in each of the $N_{SIM} = 100$ runs of the simulated annealing algorithm. The blue dot represent the estimated parameters of the best model, i.e. the run with the highest log likelohood. The red curve is a normal distribution fitted to the data for visual comparison of the results.

Parameter	Best value		Median	95% Confidence Interval
μ	$5.02 \cdot 10^{-1}$		$5.21 \cdot 10^{-1}$	$[\ 4.13 \cdot 10^{-1}\ ,\ 5.40 \cdot 10^{-1}\ ]$
k	$2.26 \cdot 10^{-2}$		$2.24 \cdot 10^{-2}$	$[\ 1.75 \cdot 10^{-2}\ ,\ 3.02 \cdot 10^{-2}\ ]$
p	$7.01 \cdot 10^{-1}$		$7.22 \cdot 10^{-1}$	$[\ 6.61 \cdot 10^{-1}\ ,\ 7.52 \cdot 10^{-1}\ ]$
c	$1.00 \cdot 10^{-5}$	*	$1.04 \cdot 10^{-5}$	$[\ 1.00 \cdot 10^{-5}\ ,\ 1.15 \cdot 10^{-2}\ ]$
α	$2.46 \cdot 10^{-1}$		$4.76 \cdot 10^{-1}$	$[\ 4.45 \cdot 10^{-2}\ ,\ 5.62 \cdot 10^{-1}\ ]$
d	$1.00 \cdot 10^{-2}$	*	$1.00 \cdot 10^{-2}$	$[\ 1.00 \cdot 10^{-2}\ ,\ 1.00 \cdot 10^{-2}\ ]$
q	3.00	*	3.00	$[\ 2.98\ ,\ 3.00\ ]$
γ	$1.97 \cdot 10^{-4}$	*	$2.15 \cdot 10^{-3}$	$[\ 2.54 \cdot 10^{-5}\ ,\ 6.74 \cdot 10^{-3}\ ]$
LogL	$-6.16 \cdot 10^{3}$		$-6.17 \cdot 10^{3}$	$[-6.20 \cdot 10^{3}\ ,-6.16 \cdot 10^{3}\ ]$
N_{events}	$2.16 \cdot 10^{3}$		$2.26 \cdot 10^{3}$	$[\ 2.04 \cdot 10^{3}\ ,\ 2.31 \cdot 10^{3}\ ]$

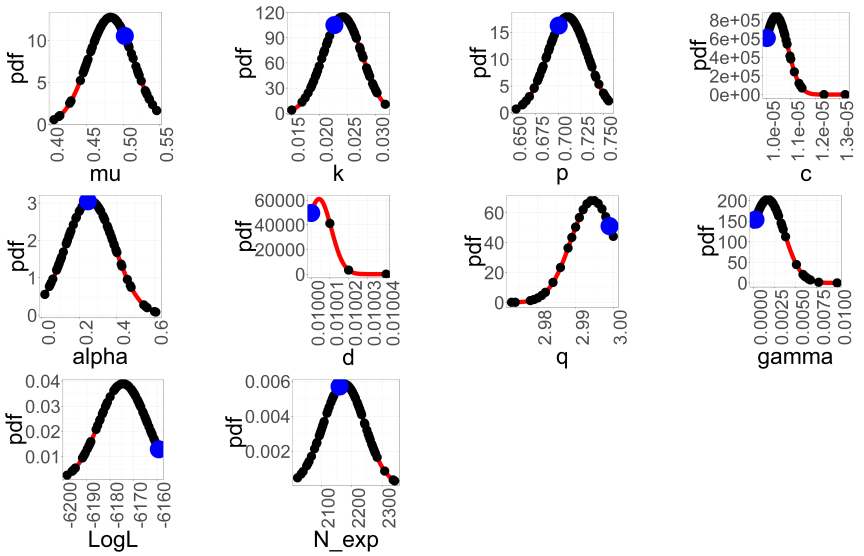


Table 15: Parameter estimates Central Africa.
Top: Estimation results for the 8 parameters of the ETAS model, the log likelihood and the number of events for the Central Africa conflict catalog. Estimates marked with a $'*'$ are at, or very close to, their constraint values. The best estimate, i.e. the estimate of the run with the highest log likelihood, is given for each parameter. Furthermore, the median and the 95% confidence intervals are given. **Bottom:** Parameter distributions for the 8 parameters of the ETAS model, the log likelihood and the number of events. The black dots are the estimated parameters in each of the $N_{SIM} = 100$ runs of the simulated annealing algorithm. The blue dot represent the estimated parameters of the best model, i.e. the run with the highest log likelohood. The red curve is a normal distribution fitted to the data for visual comparison of the results.

Parameter	Best value		Median	95% Confidence Interval
μ	$3.79 \cdot 10^{-1}$		$3.59 \cdot 10^{-1}$	$[\ 3.01 \cdot 10^{-1}\ ,\ 4.34 \cdot 10^{-1}\ ]$
k	$3.52 \cdot 10^{-2}$		$3.80 \cdot 10^{-2}$	$[\ 2.02 \cdot 10^{-2}\ ,\ 4.29 \cdot 10^{-2}\ ]$
p	$7.95 \cdot 10^{-1}$		$7.96 \cdot 10^{-1}$	$[\ 7.48 \cdot 10^{-1}\ ,\ 8.28 \cdot 10^{-1}\ ]$
c	$1.00 \cdot 10^{-5}$	*	$1.00 \cdot 10^{-5}$	$[\ 1.00 \cdot 10^{-5}\ ,\ 1.15 \cdot 10^{-5}\ ]$
α	$3.03 \cdot 10^{-1}$		$1.86 \cdot 10^{-1}$	$[\ 4.49 \cdot 10^{-2}\ ,\ 8.27 \cdot 10^{-1}\ ]$
d	$1.00 \cdot 10^{-2}$	*	$1.00 \cdot 10^{-2}$	$[\ 1.00 \cdot 10^{-2}\ ,\ 1.01 \cdot 10^{-2}\ ]$
q	3.00	*	2.99	$[\ 2.96\ ,\ 3.00\ ]$
γ	$3.21 \cdot 10^{-5}$	*	$5.65 \cdot 10^{-3}$	$[\ 8.02 \cdot 10^{-5}\ ,\ 2.39 \cdot 10^{-2}\ ]$
LogL	$-3.13 \cdot 10^{3}$		$-3.15 \cdot 10^{3}$	$[-3.30 \cdot 10^{3}\ ,-3.13 \cdot 10^{3}\ ]$
N_{events}	$1.89 \cdot 10^{3}$		$1.90 \cdot 10^{3}$	$[\ 1.68 \cdot 10^{3}\ ,\ 2.00 \cdot 10^{3}\ ]$

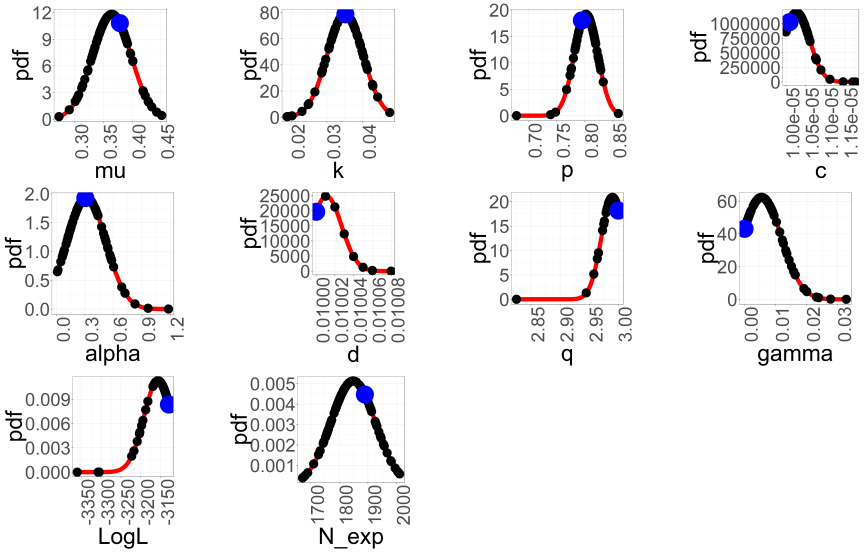
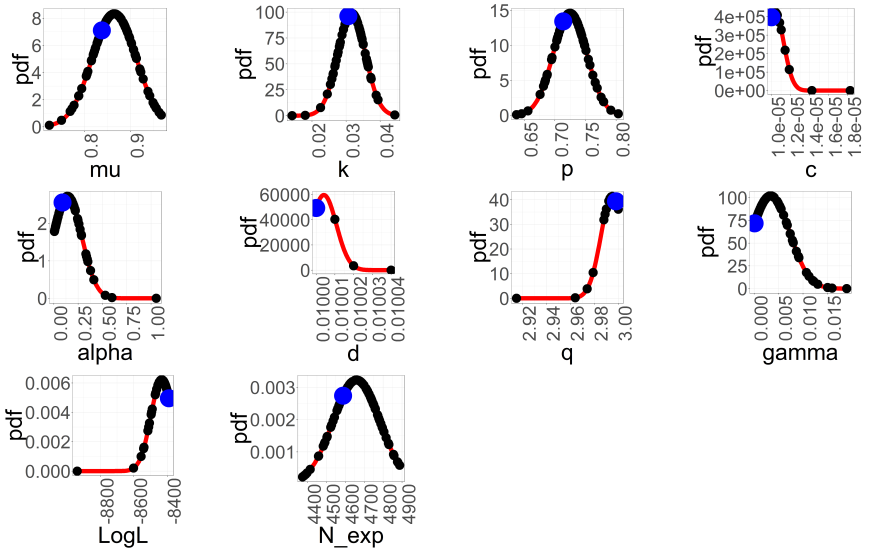


Table 16: Parameter estimates East Africa.
Top: Estimation results for the 8 parameters of the ETAS model, the log likelihood and the number of events for the East Africa conflict catalog. Estimates marked with a $'*'$ are at, or very close to, their constraint values. The best estimate, i.e. the estimate of the run with the highest log likelihood, is given for each parameter. Furthermore, the median and the 95% confidence intervals are given. **Bottom:** Parameter distributions for the 8 parameters of the ETAS model, the log likelihood and the number of events. The black dots are the estimated parameters in each of the $N_{SIM} = 100$ runs of the simulated annealing algorithm. The blue dot represent the estimated parameters of the best model, i.e. the run with the highest log likelohood. The red curve is a normal distribution fitted to the data for visual comparison of the results.

Parameter	Best value	Median	95% Confidence Interval
μ	$8.38 \cdot 10^{-1}$	$8.54 \cdot 10^{-1}$	$[\ 7.69 \cdot 10^{-1}\ ,\ 9.56 \cdot 10^{-1}\ ]$
k	$3.07 \cdot 10^{-2}$	$2.96 \cdot 10^{-2}$	$[\ 2.25 \cdot 10^{-2}\ ,\ 3.88 \cdot 10^{-2}\ ]$
p	$7.14 \cdot 10^{-1}$	$6.95 \cdot 10^{-1}$	$[\ 6.56 \cdot 10^{-1}\ ,\ 7.80 \cdot 10^{-1}\ ]$
c	$1.00 \cdot 10^{-5}$ *	$1.05 \cdot 10^{-5}$	$[\ 1.00 \cdot 10^{-5}\ ,\ 1.19 \cdot 10^{-5}\ ]$
α	$8.93 \cdot 10^{-2}$	$6.13 \cdot 10^{-1}$	$[\ 1.09 \cdot 10^{-2}\ ,\ 5.29 \cdot 10^{-1}\ ]$
d	$1.00 \cdot 10^{-2}$ *	$1.00 \cdot 10^{-2}$	$[\ 1.00 \cdot 10^{-2}\ ,\ 1.00 \cdot 10^{-2}\ ]$
q	3.00 *	3.00	$[\ 2.97\ ,\ 3.00\ ]$
γ	$8.18 \cdot 10^{-5}$ *	$6.57 \cdot 10^{-3}$	$[\ 8.18 \cdot 10^{-5}\ ,\ 1.50 \cdot 10^{-2}\ ]$
LogL	$-8.39 \cdot 10^{3}$	$-8.42 \cdot 10^{3}$	$[-8.56 \cdot 10^{3}\ , -8.39 \cdot 10^{3}\ ]$
N_{events}	$4.59 \cdot 10^{3}$	$4.68 \cdot 10^{3}$	$[\ 4.38 \cdot 10^{3}\ ,\ 4.88 \cdot 10^{3}\ ]$



5.4 Model testing

We will now discuss the results for the residual analysis and the N-test for the different regions.

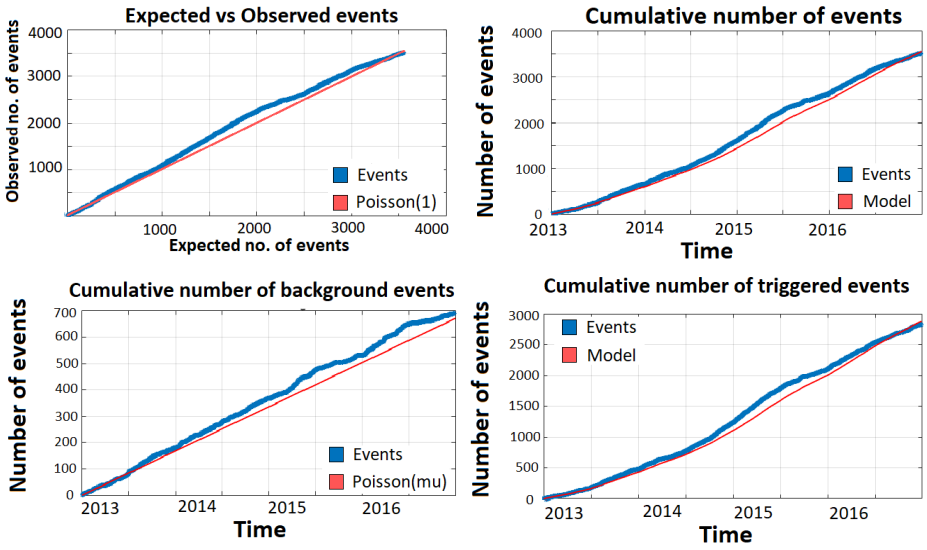
5.4.1 Residual analysis

Using the equations in section 3.3.1 we calculate the expected number of all, background and triggered events. These are compared with the observed number of events in Table 17. Note that we cannot really observe whether an event is a background or triggered event, so they are calculated as described in section 3.3.1. The numbers of expected and observed events are very close, which indicates that the model describes the data well on average. These expected and observed numbers are given over time in Figure 12, together with a plot of the transformed times for North Africa. The same results for the other three regions are very similar and are given in Figure 17, 18 and 19 in the appendix. In the top two figures, we see that the transformed times seem to follow a Poisson process with unit rate and that the number of events lies close to the number predicted by the model. In the bottom two figures, we see that this is also the case for background and triggered events. This suggests that the model is able to capture the temporal behavior of the process at first sight.

Table 17: Total number of expected and observed events, background events and triggered events for each region. The expected numbers of events are calculated using the equations in section 3.3.1.

	All		Background		Triggered	
	exp	obs	exp	obs	exp	obs
North	3558	3516	671	691	2887	2824
West	2160	2163	733	705	1427	1457
Central	1889	1834	553	525	1337	1308
East	4589	4616	1223	1255	3365	3361

Figure 12: Results of the residual analysis for the North Africa catalog. In the top left figure the observed number of events is plotted against the transformed times from equation 11. In the top right figure, the observed number of events (blue) is plotted against the expected number of events by the model (red). These are calculated using Equation 12. In the bottom two figures, the observed number of background and triggered events (blue) are compared to the number of events expected by the model (red).



Whether the model captures the temporal behavior of the process can be tested more formally using the KS-test and the runs test. These results are given in Table 18. For every region, the null hypothesis that the inter-event times are exponentially distributed is rejected. This suggests that the model does not capture the temporal behavior of the process. The runs test fails to reject the null hypothesis that the inter-event times are stationary and not autocorrelated. However, since the two tests produce such extreme results, this might be an indication that the model does not fit the data very well.

Table 18: Results of the one-sided Kolmogorov-Smirnov (KS) test and the Wald-Wolfowitz RUNS test. The p-values of each test are given for each region.

Catalog	KS-test	RUNS test
North	< 0.01	1
West	< 0.01	1
Central	< 0.01	1
East	< 0.01	1

5.4.2 Number of events test

The results of the N-test are given in Table 19. For each region, the median number of events and 95% confidence interval are given for the $N_{SIM} = 100$ simulated catalogs based on the estimated ETAS model, as described in sections 3.3.2 and 4.1 . The distribution is also displayed in Figure 16 in the appendix. Furthermore, the number of events in the conflict catalog is given, together with the probability that we observe a number of events higher than the number in the conflict catalog. We see that for the regions North, Central and East, the number of events in the conflict catalog lies within the 95% confidence bounds. For the West region the observed number of events is larger than expected by the model.

Table 19: Results for the Number of Events test, or N-test, for the different regions. The estimated ETAS model for each region is used to simulate $N_{SIM} = 100$ catalogs. The median and 95% confidence interval for are calculated for the simulated catalogs, as well as the relative difference (% dif.). Furthermore, the probability (Prob.) that we observe the number of events in the event catalog or more events is given. A low probability means that we would expect more events from this model and a high probability means we would expect less events from this model.

			N-test			
Region	Median	Observed	% dif.	95% Confidence Interval		Prob
North	3101	3516	-11.8%	[2737,	3592]	0.064
West	1933	2163	-10.6%	[1774,	2107]	0.019
Central	1861	1834	+1.5%	[1652,	2076]	0.62
East	4506	4616	-2.4%	[4090,	4852]	0.34

5.5 Forecasting results

For the forecast we simulated $N_{SIM} = 100$ catalogs and calculated the median expected number of events in each cell in the year 2017, as described in section 3.4. The total forecast number of events can then be calculated as the sum over all cells. We only forecast large events, which are events with magnitude $M \geq 4$. This corresponds to events with 22 fatalities or more. In this section, we only focus on these events. The total forecast number of events in each region is given in Table 20. The forecast number of events is compared with two naive forecasts and the observed number of events. For the first naive forecast we use the average number of events in the period 2012-2016 and for the second naive forecast we use the number of large events of last year. We see that for the regions North and West, the forecasting method is much more accurate than the naive forecasts. In the Central and East region, the forecast underestimates the number of large events.

Table 20: Forecasts for the number of events with magnitude $M \geq 4$ in 2017. The average number of these events in the period 2012-2016 is given, the number in 2016, the number predicted by the model and the observed number of events in 2017.

Region	Average	Last year	Model	Observed
North	54.2	58	16.7	17
West	61.8	39	18.3	22
Central	23.2	13	13.8	39
East	58.4	50	30.5	69

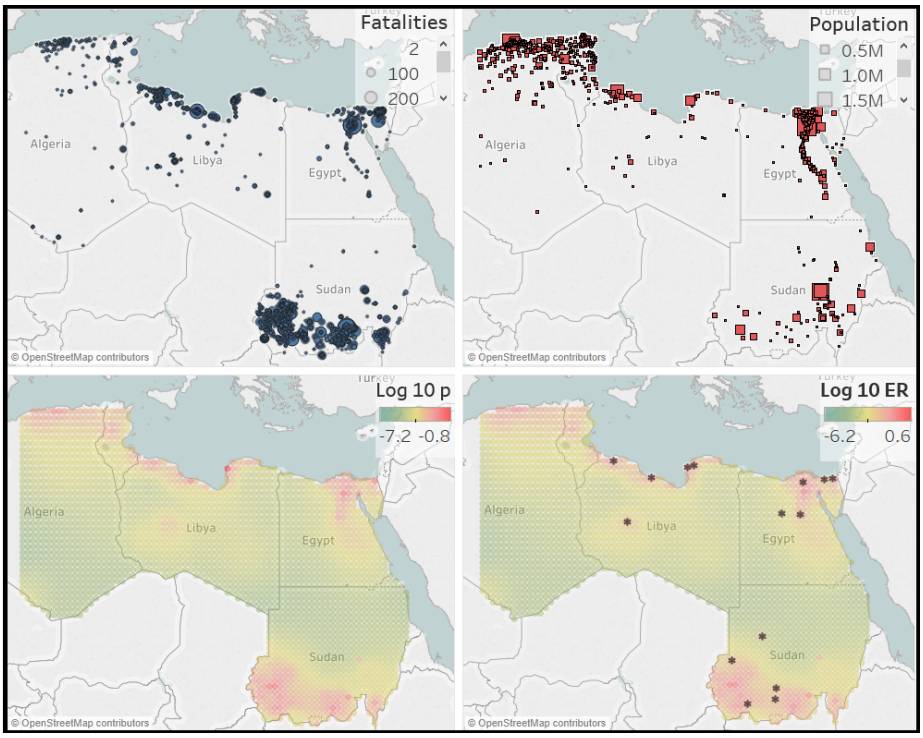
The results for the forecast locations of events is given in Figure 13 for the North and West catalog and in Figure 14 for the Central and East catalog. For the forecasting of locations we use three other predictions for comparison. Two of them are naive predictions: we predict that future events happen at places were past events have happened and we predict that future events happen in highly populated areas or around country borders. For this, a map with the events is displayed, together with a map that shows every city with a population greater than 1000. The country borders can also be seen on this map. The third other prediction is based on the model and is the estimated spatial background distribution. This predicts that future events will happen at places where there is a high probability of background events occurring. These three maps are displayed together with the forecast map. The forecast map also includes the actual locations of the large events, indicated by a '*'.

The first thing we can notice is that there is little connection between population density and conflict areas. This can for example be seen in the West catalog where most conflicts take place in the north-east side of Nigeria or in the East catalog where most conflicts take place in South Sudan. The locations of previous conflicts seem to be a good predictor for future events. Since many events overlap on the maps, these locations can be better seen in the figure with

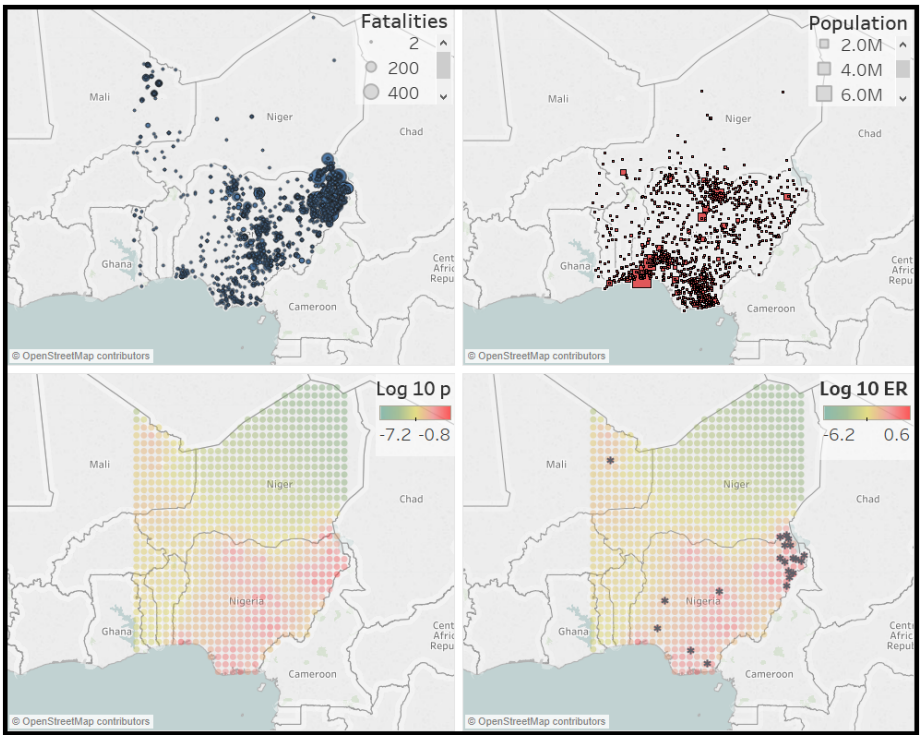
the spatial background distribution.

For the forecasting map, the similarity with the spatial background distribution is very noticeable. Further investigation showed that each cell in the forecasting map is very highly correlated with its equivalent cell in the background distribution. One possible explanation could be that because the parameter $p \leq 1$ or that the branching ratio $\phi > 1$, the process becomes explosive and the simulations for the forecasting period are not accurate. Nonetheless, the forecasting maps are still good predictors of the actual locations of future large events. On average, the forecasts coincide with the locations in the event catalogs. However, there are also some areas in which it performs better. These are regions with relatively few or small events in the past, but still show a high expected number in the forecast. For example, in the North catalog, we would not expect a future event in the center of Libya, since the previous events at that place were relatively small. Other such regions can be found in the West catalog, where two events in the west of Nigeria occurred. Finally, one of the clearest examples is in the Central catalog. In the forecast map, there is a cluster of large conflicts in an area of high intensity in the south of the Democratic Republic of the Congo. This was not expected from the event map, where only a few small events can be seen. Hence, even though the forecasting map is highly correlated with the spatial background distribution, it still produces useful predictions for the locations of future large events.

Figure 13: Four maps for the regions North- (a) and West Africa (b). Top left: event catalog 2012-2016. Top right: City map with cities with 1000+ population. Bottom left: spatial background distribution estimated from the data from 2012-2016. Bottom right: forecast for the expected number of events with $M \geq 4$ in each cell in the year 2017, together with the locations of the actual conflicts that have taken place with $M \geq 4$.

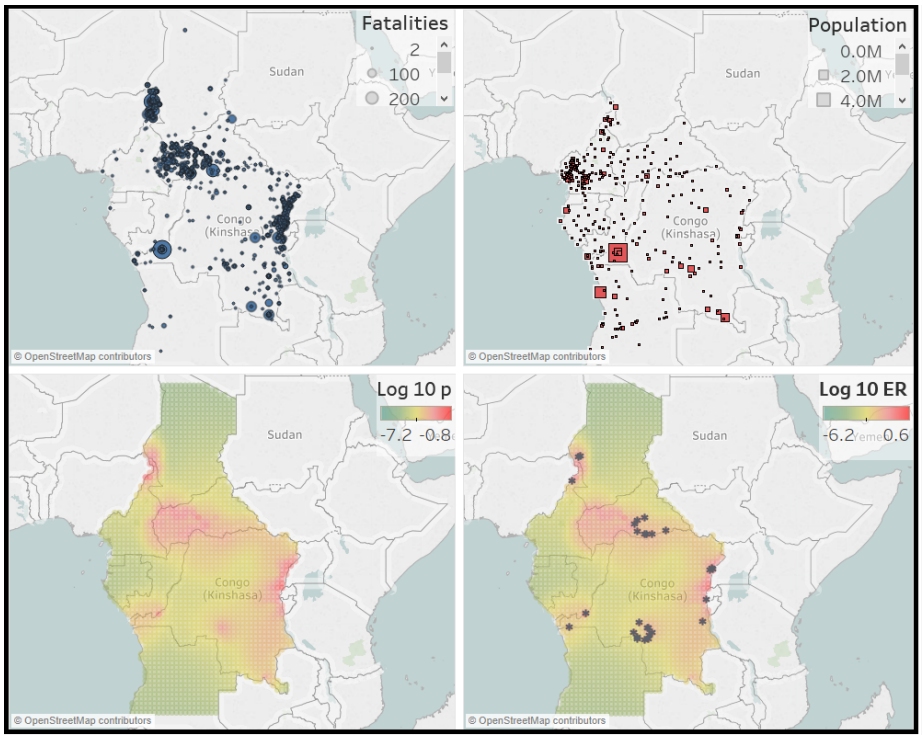


(a) Forecast for the region North Africa.

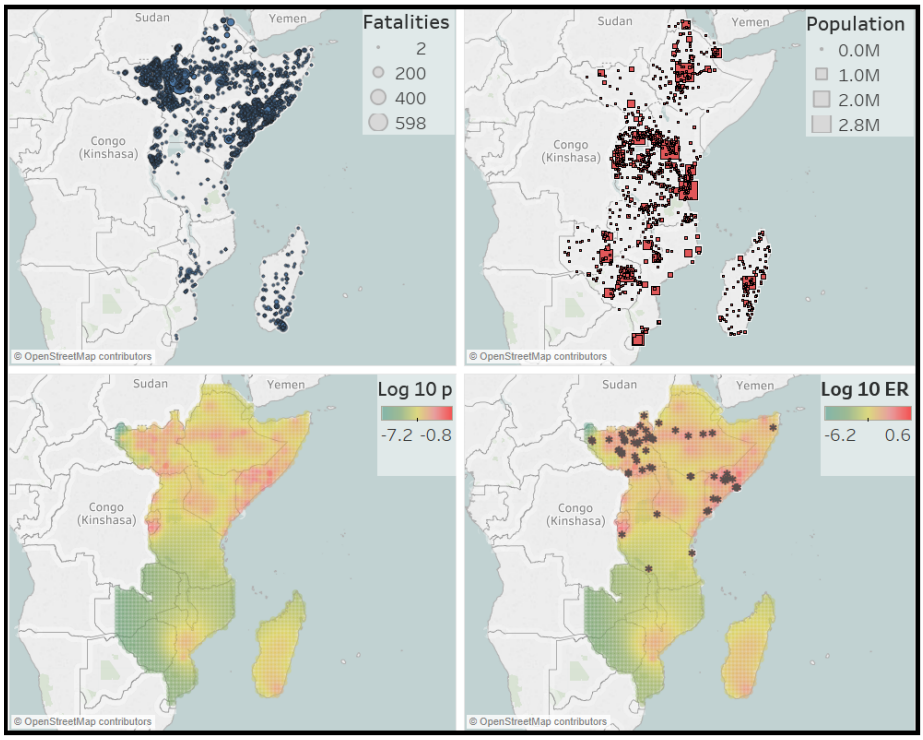


(b) Forecast for the region West Africa.

Figure 14: Four maps for the regions Central- (a) and East Africa (b). Top left: event catalog 2012-2016. Top right: City map with cities with 1000+ population. Bottom left: spatial background distribution estimated from the data from 2012-2016. Bottom right: forecast for the expected number of events with $M \geq 4$ in each cell in the year 2017, together with the locations of the actual conflicts that have taken place with $M \geq 4$.



(a) Forecast for the region Central Africa.



(b) Forecast for the region East Africa.

6 Conclusion & discussion

Conclusion

We have successfully estimated the spatial-temporal ETAS model on social conflict data in Africa. Using this model we could make a forecast for the number of large future events and their locations. The estimated spatial background distribution showed that the probability distribution of a background events was in line with the locations in the conflict catalogs: high probability areas were associated with high event density areas. With the parameter estimation, certain parameters went to their constraint values in each catalog, which could suggest that not every parameter in the ETAS model is identified. However, with these parameters at their constraint values, we could still make a comparison for the other parameters across the four major regions. These parameter values were in line with what was expected from the catalogs. The N-test showed that for three out of four regions, the number of events in the conflict catalog were close to the number of events expected by the model. Only for the West region the number of events was slightly underestimated. The KS-test showed that the inter-event times were not exponentially distributed, indicating that the model does not capture the temporal behavior of the data very well. The runs test suggested that the inter-event times are stationary and not autocorrelated. However, due to the extreme results of the KS-test and the runs test, it might be the case that the model does not fit the data very well. A visual comparison of the residuals with the observed events indicated that the model captures the temporal behavior of the data fairly well.

Finally, we were able to use the estimated model to make predictions for the number and locations of future large events. The combination of estimated parameter values gave a branching ratio greater than one, which possibly caused the forecasts to become highly correlated with the spatial background distribution. Nonetheless, the forecast locations were very much in line with the locations of actual events taking place.

In conclusion, even though social conflicts and earthquakes do not look very similar at first sight, we were still able to estimate the spatial-temporal ETAS model on the conflict data and produce useful forecasts for future events.

Suggestions for future research

The ETAS model and estimation procedures have been optimized specifically for estimating earthquake catalogs. For example, the constants in the cooling procedure in the simulated annealing algorithm were found to provide an efficient algorithm for estimating earthquake catalogs. This, however, does not mean that these are the optimal constants for estimating conflict catalogs. Moreover, the parameter constraints in estimating the model are obtained from the literature about estimating earthquake occurrences and might be too strict for modeling conflicts. However, relaxing these constraints still gave the same esti-

mation problems, i.e. some parameters became arbitrarily small or large. This problem still persisted after many trial and error with different constraint values. Hence, further research in the estimation procedure and parameter constraints regarding social conflicts catalogs might produce better results.

The model estimated in this study is the stationary ETAS model, which assumes that the background rate does not change over time. This is not always the case for conflict data. For example, the rise of the Arab Spring in North Africa caused a large influx of events. This study can be extended using a non-stationary ETAS model as in Kumazawa et al. (2014) ,where the background rate varies over time. Furthermore, the models for earthquake occurrences work with a continuous time, while the conflicts are only reported with a daily precision. Another possible extension might therefore be a discrete time implementation. Finally, the model used in this study assumed the Gutenberg-Richter law for the magnitude distribution, since it closely resembles the magnitude distribution of earthquakes. One could use a different distribution that is more natural to the size distribution of conflicts.

Regarding the data itself, it would be interesting to apply the same procedures to a comparison of smaller regions. Regions of interest could be the north-east of Nigeria, where the Boko Haram insurgency is active, the conflicts around lake Chad or around the Sudan South-Sudan border. This study focused more on the larger scale of conflict dynamics and it would be interesting to see if these findings still hold at a smaller scale.

References

Brandt, P. T., Freeman, J. R., and Schrodt, P. A. (2011). Real time, time series forecasting of inter-and intra-state political conflict. *Conflict Management and Peace Science*, 28(1):41–64.

Cilliers, J. (2015). Future (im) perfect? Mapping conflict, violence and extremism in Africa. *Institute for Security Studies Papers*, 2015(287):24.

Ingber, L. (2000). Adaptive simulated annealing (ASA): Lessons learned. *arXiv preprint cs/0001018*.

Johnson, N., Hitchman, A., Phan, D., and Smith, L. (2017). Self-exciting point process models for political conflict forecasting. *European Journal of Applied Mathematics*, 28(6):1–23.

Kumazawa, T., Ogata, Y., and Others (2014). Nonstationary ETAS models for nonstandard earthquakes. *The Annals of Applied Statistics*, 8(3):1825–1852.

Lin, F.-T., Kao, C.-Y., and Hsu, C.-C. (1993). Applying the genetic approach to simulated annealing in solving some NP-hard problems. *IEEE Transactions on systems, man, and cybernetics*, 23(6):1752–1767.

Lombardi, A. M. (2015). Estimation of the parameters of ETAS models by Simulated Annealing. *Nature Publishing Group*, 5(February):8417.

Lombardi, A. M. (2017). SEDA : A software package for the Statistical Earthquake Data Analysis. *Nature Publishing Group*, 7(February):1–16.

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

Ogata, Y. (1998). Space-Time Point-Process Models for Earthquake Occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402.

Sen, M. K. and Stoffa, P. L. (2013). *Global optimization methods in geophysical inversion*. Cambridge University Press.

Vere-Jones, D. (1970). Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(1):1–62.

Zammit-Mangion, A., Dewar, M., Kadirkamanathan, V., and Sanguinetti, G. (2012). Point process modelling of the Afghan War Diary. *Proceedings of the National Academy of Sciences*, 109(31):12414–12419.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380.

7 Appendix

7.1 Magnitude distribution of social conflicts

The following is the description of assigning magnitudes to a number of fatalities in the social conflict catalogs.

Let M_{min} and M_{max} be the minimum and maximum of our magnitude range. For a bin size of 0.1, i.e. magnitudes are reported with one decimal accuracy as in our data, we have a total of

$$N_{steps} = 10 \cdot (M_{max} - M_{min}) + 1$$

magnitude steps. One extra step is included since M_{min} is included in the range. Given that a percentage x of events has magnitude M_{min}, we want a percentage $x/10$ to have magnitude $M_{min} + 1$, a percentage $x/100$ to have magnitude $M_{min}/100$, etc. In smaller magnitude steps, an increase of 0.1 corresponds to a frequency decrease of $\omega = 10^{1/10} \approx 1.259$. Then, given that a percentage x of events has magnitude M_{min}, we want a percentage x/ω to have magnitude $(M_{min} + 0.1)$, a percentage x/ω^2 to have magnitude $(M_{min} + 0.2)$, etc.

To determine the percentage of events with a magnitude of M_{min}, we solve the following equation

$$\sum_{n=0}^{N_{\text{steps}}} \frac{x}{\omega^n} = 100\%,$$

which gives us $x \simeq 20\%$ for $10 < N_{steps} < 100$. Next, we have to estimate the number of magnitude steps. The largest event is one event with 600 fatalities in a dataset of about 20,000 events in total. Hence, the largest event, which will be given the largest magnitude M_{max}, has a occurrence frequency of 1 in 20,000 or 0.005%. If the percentage of events with magnitude M_{min} is roughly 20%, then the percentage of events with magnitude $(M_{min}+4)$ will be $20\%/10^4 = 0.002\%$, which is about the frequency of the largest event. Hence, if we use 4 full magnitude steps, or 40 smaller steps, the frequency of the smallest and largest events follow the Gutenberg-Richter law. This means we will use a magnitude range of $[M_{min}, M_{max} = M_{min} + 4]$. Lastly, we have to determine M_{min}, which can be chosen rather arbitrarily. In earthquake studies, the minimum magnitude is often set at $M_{min} = 3$, since this is when an earthquake can be considered of reasonable size. To be in line with the earthquake models, we will also set $M_{min} = 3$ which gives a magnitude range of $[M_{min}, M_{max}] = [3, 7]$. For this

magnitude range, $x \simeq 20.56\%$ and the set $\{x/\omega^n\}_{n=0}^{n=40}$ gives the ideal percentages of events for each magnitude $\{3.0,\ 3.1\ , ...,\ 7.0\}$.

Next, we have to assign each number of fatalities to a magnitude, in such a way that the resulting magnitude distribution matches the above distribution as close as possible. For this, we first calculate the cumulative percentages $\{\sum_{k=0}^{n} x/\omega^k\}_{n=0}^{n=40}$ for each magnitude. Then we use these cumulative percentages as fatality quantiles. If a number of fatalities corresponds to multiple quantiles, it will be assigned to the magnitude of the lowest quantile. For each magnitude, we have the number of fatalities, expected and observed percentage of events, and the expected and observed number of events.

Table 21: Full table of the magnitude distribution of the social conflict data. For each magnitude (Magn.), we have the corresponding number of fatalities (Fatal.), the expected percentage of events (% exp.), , the observed percentage of events (% obs.), the expected number of events (N exp.) and the observed number of events (N obs.)

Magn.	Fatl.	% exp.	% obs.	N exp	N obs
3	2	20.57	26.94	3565	4669
3.1	3	16.34	14.49	2832	2511
3.2	4	12.98	17.29	2250	2997
3.3	6	10.31	8.64	1787	1498
3.4	8	8.19	4.76	1419	825
3.5	10	6.50	10.00	1127	1733
3.6	10	5.17	0.00	896	0
3.7	12	4.10	3.91	711	677
3.8	15	3.26	3.64	565	631
3.9	19	2.59	2.02	449	351
4	22	2.06	1.77	357	306
4.1	27	1.63	1.29	283	224
4.2	31	1.30	1.07	225	186
4.3	38	1.03	0.93	179	162
4.4	45	0.82	0.33	142	58
4.5	50	0.65	0.85	113	147
4.6	58	0.52	0.44	90	76
4.7	67	0.41	0.34	71	59
4.8	78	0.33	0.27	57	47
4.9	97	0.26	0.05	45	9
5	100	0.21	0.33	36	57
5.1	107	0.16	0.13	28	22
5.2	129	0.13	0.07	22	13
5.3	150	0.10	0.00	18	0
5.4	183	0.08	0.07	14	12
5.5	204	0.07	0.05	11	8
5.6	248	0.05	0.05	9	8
5.7	310	0.04	0.03	7	6
5.8	349	0.03	0.00	6	0
5.9	400	0.03	0.05	4	8
6	400	0.02	0.00	4	0
6.1	409	0.02	0.02	3	3
6.2	458	0.01	0.01	2	2
6.3	590	0.01	0.00	2	0
6.4	597	0.01	0.02	1	4
6.5	597	0.01	0.00	1	0
6.6	597	0.01	0.00	1	0
6.7	598	0.00	0.01	1	1
6.8	598	0.00	0.00	1	0
6.9	599	0.00	0.00	0	0
7	600	0.00	0.01	0	1

7.2 Model tests

Here, the distributions of the N-tests are given for the simulated catalog and the four regions. Furthermore, the plots from the residual analysis are given for the regions West, Central and East. The analysis for the North catalog is given in the main text.

Figure 15: Distribution of the Number of Events test for the simulated catalog

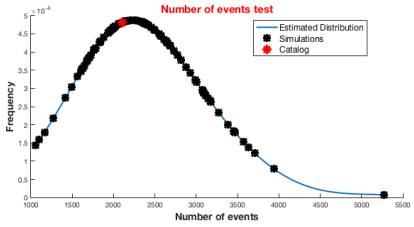
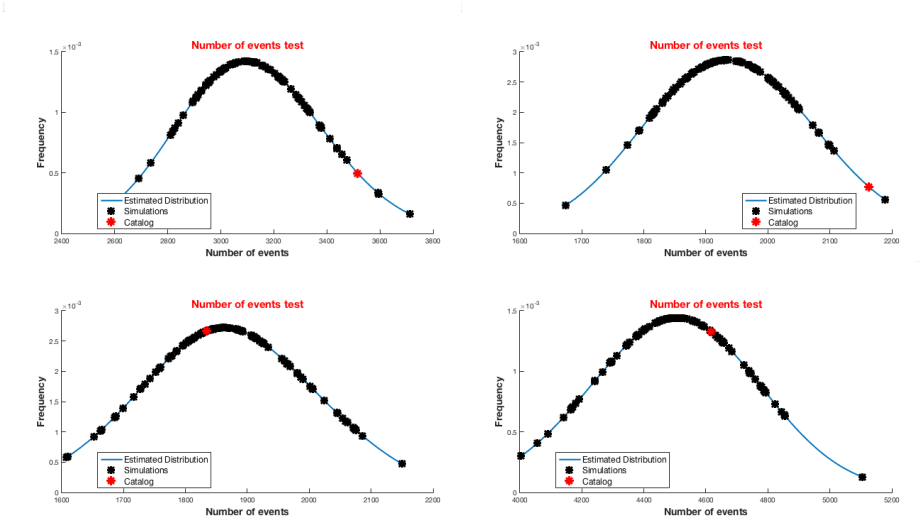


Figure 16: Distribution of the number of events from the Number of Events test for each region.



7.3 Residual analysis

Figure 17: Residual analysis for the West Africa catalog.

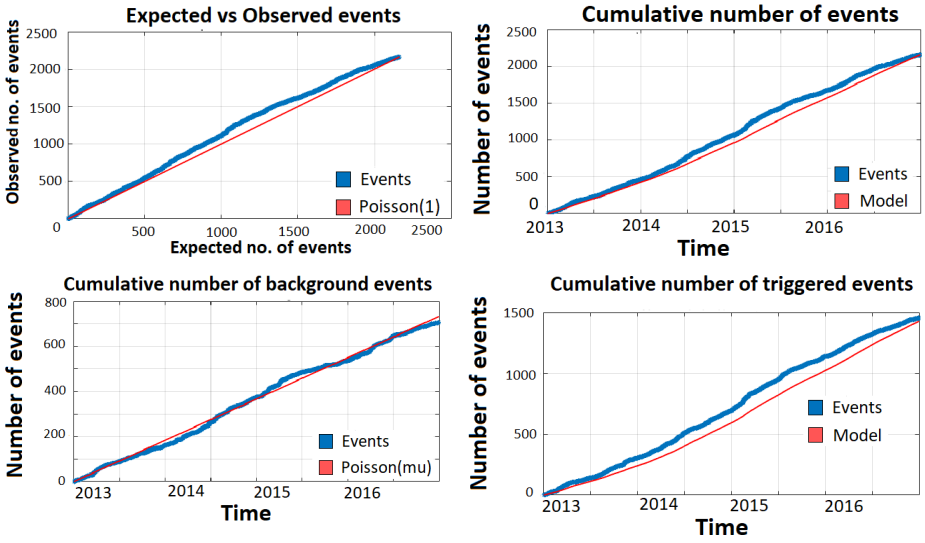


Figure 18: Residual analysis for the Central Africa catalog.

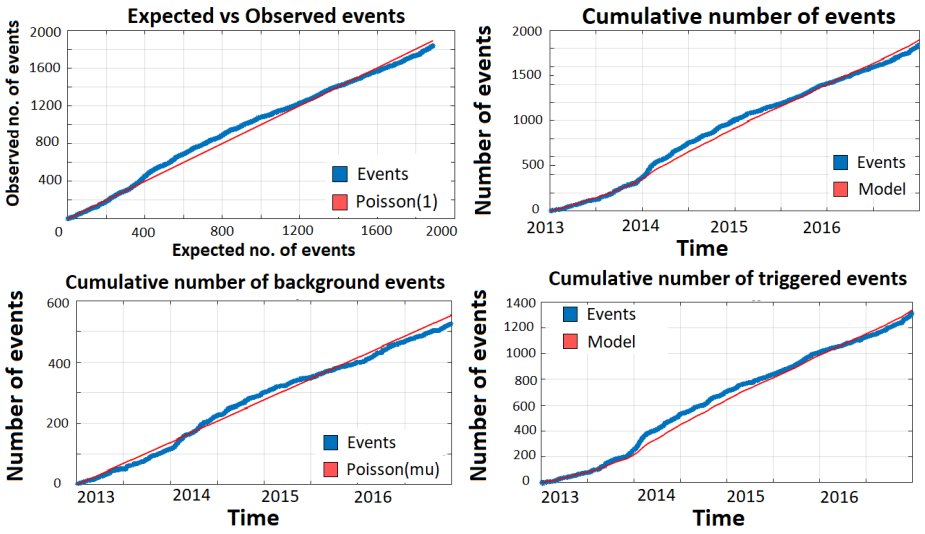


Figure 19: Residual analysis for the East Africa catalog.