Missing earthquake data reconstruction in the space-time-magnitude domain

Angela Stallone[1,1,1] and Giuseppe Falcone[1,1,1]

[1]Istituto Nazionale di Geofisica e Vulcanologia (INGV)

November 30, 2022

Abstract

Short term aftershock incompleteness (STAI) can strongly bias any analysis built on the assumption that seismic catalogs have a complete record of events. Despite several attempts to tackle this issue, we are far from trusting any dataset in the immediate future of a large shock occurrence. Here we introduce RESTORE (REal catalogs STOchastic REplenishment), a Python toolbox implementing a stochastic gap-filling method, which automatically detects the STAI gaps and reconstructs the missing events in the space-time-magnitude domain. The algorithm is based on empirical earthquake properties and relies on a minimal number of assumptions about the data. Through a numerical test, we show that RESTORE returns an accurate estimation of the number of missed events and correctly reconstructs their magnitude, location and occurrence time. We also conduct a real-case test, by applying the algorithm to the Mw 6.2 Amatrice aftershocks sequence. The STAI-induced gaps are filled and missed earthquakes are restored in a way which is consistent with data. RESTORE, which is made freely available, is a powerful tool to tackle the STAI issue, and will hopefully help to implement more robust analyses for advancing operational earthquake forecasting and seismic hazard assessment.

Missing earthquake data reconstruction in the space-time-magnitude domain

Angela Stallone[1]*, and Giuseppe Falcone[1]

[1]Istituto Nazionale di Geofisica e Vulcanologia (INGV)

Key Points:

- A new Python toolbox for the replenishment of incomplete seismic catalogs is developed
- The code is freely-available, data-driven and minimizes the users' inputs
- Numerical and real-case tests are provided

*Via di Vigna Murata, 605 - 00143 Rome, Italy

Corresponding author: Angela Stallone, `angela.stallone@ingv.it`

Abstract

Short term aftershock incompleteness (STAI) can strongly bias any analysis built on the assumption that seismic catalogs have a complete record of events. Despite several attempts to tackle this issue, we are far from trusting any dataset in the immediate future of a large shock occurrence. Here we introduce RESTORE (REal catalogs STOchastic REplenishment), a Python toolbox implementing a stochastic gap-filling method, which automatically detects the STAI gaps and reconstructs the missing events in the space-time-magnitude domain. The algorithm is based on empirical earthquake properties and relies on a minimal number of assumptions about the data. Through a numerical test, we show that RESTORE returns an accurate estimation of the number of missed events and correctly reconstructs their magnitude, location and occurrence time. We also conduct a real-case test, by applying the algorithm to the M_W 6.2 Amatrice aftershocks sequence. The STAI-induced gaps are filled and missed earthquakes are restored in a way which is consistent with data. RESTORE, which is made freely available, is a powerful tool to tackle the STAI issue, and will hopefully help to implement more robust analyses for advancing operational earthquake forecasting and seismic hazard assessment.

1 Introduction

It is well known that analyzing an incomplete seismic catalog could severely bias studies aimed to: 1) estimate the Gutenberg-Richter parameters, their uncertainty, together with their variation in space and/or time (e.g. Knopoff et al., 1982; Schorlemmer et al., 2003; Woessner & Wiemer, 2005; Mignan & Woessner, 2012b; Marzocchi et al., 2020); 2) estimate the Epidemic-Type Aftershock Sequence (ETAS model: Ogata, 1988, 1998) parameters by maximum-likelihood techniques (Helmstetter et al., 2005, 2006; Hainzl et al., 2013; Omi et al., 2014; Seif et al., 2017; Zhuang et al., 2017); 3) perform a statistical analysis of earthquake data (e.g. Helmstetter et al., 2006; Christophersen & Smith, 2008; Iwata, 2008; Brodsky, 2011; Felzer et al., 2015; Stallone & Marzocchi, 2019). The first two types of studies, in particular, have application in operational earthquake forecasting and seismic hazard assessment (Woessner et al., 2015), this implying that complete recording of seismic events is of primary importance in any analysis of this kind. Unfortunately, a careful estimation of the magnitude of completeness M_c is a necessary but not sufficient condition for a robust seismicity analysis. As a matter of fact, temporal changes in M_c can occur, mainly due to short term aftershock incompleteness (STAI from now on) (Ogata & Katsura, 1993; Kagan, 2004; Mignan & Woessner, 2012b; Omi et al., 2013), which arise from the under-reporting of small events after large earthquakes. These fluctuations, although transient, can severely alter the final results. For instance, (Zhuang et al., 2017) demonstrate how severe can be the influence of short-term missing aftershocks on the estimation of the ETAS parameter α (which is linked to earthquakes triggering capability). A solution to this issue would be improving the detection of early aftershocks of a large earthquake. This is possible by implementing waveform-based techniques (Peng et al., 2006, 2007; Enescu et al., 2007, 2009; Peng & Zhao, 2009). However, even in these cases, the detection capability of the missing events is far from being optimal. A quick fix could be to draw out earthquakes occurred after a large shock, for as long as the time required to the magnitude of completeness to return to the average value estimated for the whole catalog. Alternatively, one could model the magnitude of completeness as a function of time $M_c(t)$ and keep only those events whose magnitude is $\geq M_c(t)$ (e.g. Helmstetter et al., 2006; Lippiello et al., 2012). However, these approaches are not trivial, since they rely on user-defined criteria for identifying the critical events to be removed. Furthermore, a cut-and-run strategy could yield to a severe diminishment of the analyzed data, which is not always desirable. More recently, (Zhuang et al., 2017, 2019) have proposed a stochastic algorithm to replenish the portions of a seismic catalog where smaller events are missing. This approach is based on empirical distribution functions that approximately describe the time-magnitude range of data where the catalog is assumed to be complete. Furthermore, it cannot be easily extended to the spatial domain and the detection of the area where the record is incomplete is based on visual inspection. Here we present RESTORE, a Python toolbox based on a stochastic gap-filling method, which reconstructs missing events in the space-time-magnitude domain and implements an automatized recognition of the critical regions with missing events (no input required from the user). RESTORE is built on well-known empirical properties of earthquake data and relies on a fully data-driven approach, which severely minimizes the number of assumptions and approximations about the data.

2 The algorithm

RESTORE (REal catalogs STOchastic REplenishment) allows to generate time, location and magnitude of those earthquakes that have not been detected by the seismic network due to the overlap of earthquake signals in seismic records after the occurrence of a large earthquake. Given the transient characteristic of STAI, the replenishment of missing data only pertains to limited portions of the catalog, i.e. those being affected by the occurrence of a large event. First, the temporal variability of M_c is assessed by means of a sliding overlapping windows approach, which collects estimates of M_c at the end of each window. Since the window has a fixed number of events k and its shift δk is constant, estimates of M_c are elapsed by δk events. The algorithm implements a statistic-based approach to pinpoint those time intervals where a threshold value for the magnitude of completeness M_c^* is significantly exceeded ("STAI gaps" from now on). For each interval, fluctuations in the completeness magnitude, represented by the δk-shifted moving-window estimates of M_c, are accounted for to reconstruct the missing earthquakes: the higher the estimated M_c, the higher the number of earthquakes to be replenished. It follows that the moving-window approach is functional for both the identification of STAI gaps and for their discretization. The latter is essential for a high-resolution temporal reconstruction of M_c inside the STAI gaps. The algorithm evaluates, for each magnitude bin in each step, the difference between the observed counts and the counts predicted by the Gutenberg-Richter relationship. This approach returns the simultaneous estimation of both the number and magnitudes of missing events at the bin level: the first is derived from the difference between observed and estimated counts, whereas the second is derived from the magnitude value in the bin. Occurrence time and location of the simulated events are reconstructed implementing Monte Carlo sampling techniques (inverse method, (Devroye, 1986)). More specifically, occurrence times are simulated from an uniform distribution whose support are the time limits of the δk-step. The latter is based on the assumption that earthquake detection rate can be assumed constant within intervals including few events, i.e within very short time intervals. In other words, the probability of missing events within a δk-step can be considered time-independent if the step width is much shorter than the whole STAI gap width. As regards the spatial information, latitude and longitude of missing events are assigned with a probability that increases as the average rate of earthquake increases, the latter being derived from a Gaussian smoothing kernel. In the following, we examine the algorithm steps in more detail.

2.1 Query user inputs

The user is required to load the catalog as a csv file, in ZMAP format (i.e., Longitude, Latitude, Year, Month, Day, Magnitude, Depth, Hour, Minute, Second). Alternatively, he/she can download it from web services based on FDSN specification, by providing the parameters listed in Table 1, left column. There are two main requirements for the correct implementation of RESTORE. First, the magnitude type in the seismic catalog must be the moment magnitude Mw (a bin size of 0.1 is assumed by default). This is required since magnitude scales other than the moment magnitude are inappropriate for rigorous statistical analyses (Kagan, 2013). Second, the catalog should include a period of seismic quiescence before the onset of one or more relatively strong seismic sequences. This is necessary for the estimation of the reference value for the magnitude

Table 1. RESTORE input parameters[a].

CATALOG PARAMETERS (optional)	INPUT PARAMETERS
Minimum magnitude	Moving-window size
Minimum longitude	Moving-window step
Maximum longitude	Spatial map domain limits
Minimum latitude	t_{seq}
Maximum latitude	b-value
Maximum depth	α (Lilliefors test)
t_{start}	
t_{end}	

[a]Left: catalog parameters (to be provided only when downloading the catalog from web services based on FDSN specification) - t_{start}: string representing the start time of the catalog in a recognizably valid format; t_{end}: string representing the end time of the catalog in a recognizably valid format.
Right: RESTORE parameters - t_{seq}: starting time of the seismic sequence (i.e., end of the seismic quiescent period).

of completeness (M_c^*), which must not be affected by STAI. The parameters that need to be set for running RESTORE are reported in Table 1, right column. They will be explained in more detail in the subsequent sections.

2.2 Reference value for the magnitude of completeness

The reference value M_c^* must be evaluated for the seismically quiescent period preceding the onset of one or more relatively strong seismic sequences. By default, it is estimated as the first magnitude value such that the hypothesis of exponentially-distributed data cannot be rejected at a significance level α (Lilliefors test, Lilliefors, 1969; Clauset et al., 2009). Alternatively, the user could input his/her own value for M_c^*, based on a priori information. RESTORE relies on *Mc-Lilliefors*, a Python routine which returns a robust and rigorous estimation of the magnitude of completeness by the Lilliefors test (Herrmann & Marzocchi, 2020b, 2020a). From now on, we always mean that the magnitude of completeness estimation has been performed by the *Mc-Lilliefors* routine.

2.3 Temporal variations in M_c

RESTORE implements a moving window approach to analyze the variation of the magnitude of completeness as a function of time. By default, the window size is $k = 1000$ events (following Mignan & Woessner, 2012a), but it could be increased or decreased, depending on both the catalog size and the resolution the user needs to achieve. Intuitively, a small window highlights short-term variation in M_c, but it could return a biased estimate of M_c if the sample size is too small (due to the decreased power of the Lilliefors test); on the contrary, a larger window returns a faster and more robust estimate of M_c, but it is less sensitive to its transient fluctuations. The window is shifted by a step of δk events. By default, $\delta k = 250$ (following Mignan & Woessner, 2012a). The same considerations made for a larger/smaller window apply for a larger/smaller step. M_c is estimated and its values are collected at the end of each window. Since the window has a fixed number of events k and its shift δk is constant, estimates of M_c are elapsed by δk events.

2.4 Automatic detection of STAI gaps

STAI gaps are identified as those where $M_c \geq M_c^* + 2\sigma$, i.e. where M_c is significantly larger than the reference value. The bootstrap method (Efron, 1992) is implemented to estimate the uncertainty σ about the estimate of M_c^* returned by the Lilliefors test. Specifically, σ is obtained from 200 bootstrap samples, as suggested in (Woessner & Wiemer, 2005). The onset time of each gap is set equal to the time of the largest earthquake in the first step. Intuitively, it is the one responsible for the raise of the magnitude of completeness among the δk events. The end time of each gap is coincident with the occurrence time of the last event in the last step. In order to account for statistical fluctuations of the magnitude of completeness, small gaps - defined as those with a number of events $< 2 * \delta k$ - are removed.

2.5 Simulation of missing earthquakes

RESTORE implements a multi-scale approach for addressing the inherent problem of multidimensionality of the seismic process:

- Small scale: magnitude bin-level estimation of the number and magnitudes of missing events (Section 2.5.1);
- Medium scale: step-level estimation of the occurrence times of missing events (Section 2.5.2);
- Coarse scale: STAI gap-level simulation of missing events epicenters (Section 2.5.3).

2.5.1 Simulation of number of missing events and magnitudes

For a given STAI gap, the algorithm stores as many M_c estimates as the number of δk-steps in the gap. This information is used to evaluate the number of expected events at the magnitude bin level by means of the following equation, which relies on the Gutenberg-Richter frequency-magnitude relationship (we refer to Appendix 1 for its derivation):

$$N(M \geq M_{LB}) = N(M \geq M_{UB}) \cdot 10^{b \cdot mbin}, \tag{1}$$

where: 1) M_{LB} (M_{UB}) is the lower (upper) bound of the magnitude bin $mbin$, with $mbin = 0.1$ by default; 2) $N(M \geq M_{LB})$ is the expected number of events with magnitude $M \geq M_{LB}$; 3) $N(M \geq M_{UB})$ is the observed number of events with magnitude $M \geq M_{UB}$. Equation 1 allows to extrapolate the expected number of events with magnitude $M \geq M_{LB}$, given the complete recording of events at magnitudes $M \geq M_{UB}$. It is then straightforward to retrieve, for each bin, the expected number of events with magnitudes $M = M_{LB}$ ($M_{LB} \leq M < M_{UB}$), given the complete recording of events at magnitudes $M \geq M_{UB}$:

$$\begin{aligned} N(M = M_{LB}) &= N(M \geq M_{LB}) - N(M \geq M_{UB}) \\ &= N(M \geq M_{UB}) \cdot 10^{b \cdot mbin} - N(M \geq M_{UB}) \end{aligned} \tag{2}$$

Finally, the number of missing events in the bin is derived from the difference between the expected number of events in the bin, $N(M = M_{LB})$, and the observed number of events in the bin. RESTORE recursively implements Equation 1 and Equation 2 in order to estimate the number and magnitudes of missing events for all the bins between M_c^* - the reference value for the magnitude of completeness - and M_c^S, the magnitude of completeness estimated within the step. The algorithm starts at the bin whose upper bound is M_c^S: since magnitudes in the step are complete above M_c^S, Equation 1 and Equation 2 are implemented for the estimation of the number of missing events in the preceding bin. Then, variables are updated and the algorithm proceeds with the next preceding bin, following the recursive approach explained in Algorithm 1.

Algorithm 1: Magnitude simulation

for *each STAI gap* **do**
  for *each step in the gap* **do**
    $M_{UB} = M_c^*$;
    $M_{LB} = M_{UB} - mbin$;
    $N(M \geq M_{UB}) \longrightarrow$ Counts of $M \geq M_{UB}$ in the step
    for *each bin in the step* **do**
      $N(M \geq M_{LB}) = N(M \geq M_{UB}) \cdot 10^{b \cdot mbin}$
      $N(M = M_{LB}) = N(M \geq M_{LB}) - N(M \geq M_{UB}) \longrightarrow$ Expected number of magnitudes in the bin
      $n(M = M_{LB}) \longrightarrow$ Observed number of magnitudes in the bin
      $N_{ghost} = N(M = M_{LB}) - n(M = M_{LB}) \longrightarrow$ Number of missing events in the bin
      $M = [M_{LB}] * N_{ghost} \longrightarrow$ Vector of missing magnitudes in the bin
      Update variables:
      $M_{UB} = M_{LB}$;
      $M_{LB} = M_{UB} - mbin$;
      $N(M \geq M_{UB}) = N(M \geq M_{LB})$
    end
  end
end

2.5.2 Simulation of occurrence times

Occurrence times are simulated from an uniform distribution whose support are the time limits of the δk-step. As already discussed, earthquake detection rate can be assumed constant within intervals including few events, i.e within very short time intervals. The main steps are summarized in Algorithm 2.

Algorithm 2: Occurrence times simulation

for *each STAI gap* **do**
  for *each step in the gap* **do**
    while *count* $\leq$ *Number of earthquakes missing in the step* **do**
      t_{i-1} = start time of the step;
      t_i = end time of the step;
      U = RAND(0,1);
      $T = t_{i-1} + U \cdot (t_i - t_{i-1})$;
      count $+= 1$;
    end
  end
end

2.5.3 Simulation of epicenter coordinates

Latitude and longitude of missing events are assigned with a probability that increases as the rate of earthquakes increases, i.e. as the distance from the large event diminishes. The rationale is based on kernel smoothing techniques, commonly implemented to forecast the density of future seismicity given the spatial distribution of past events (e.g. Frankel, 1995; Helmstetter et al., 2006; Zechar & Jordan, 2010). Specifically, a Gaussian kernel (Zechar & Jordan, 2010) is used, which is a function of the smoothing distance σ only. For each STAI gap, RESTORE extracts the pertaining subset from the catalog, that is all the events meeting the following two criteria: 1) their occurrence times range between the start and end time of the STAI gap; 2) their epicenter coordinates fall within a rectangular grid representing the large shock "influence area". As a proxy for the latter, the algorithm uses the estimation of the subsurface rupture length through the relation proposed by (Mai & Beroza, 2000):

$$M_o = 10^{\frac{3}{2}(M_w+10.7)} \cdot 10^{-7}; \tag{3}$$

$$R_l = 10^{-5.20+0.35 \cdot log(M_o)} \tag{4}$$

The grid is discretized in cells, whose width depends on the bin in the latitude and longitude direction *sbin* (*sbin* = 0.01 deg in both the directions, by default). The smoothing kernel is defined as follow:

$$K_\sigma = \frac{1}{2\pi\sigma^2} exp\left(\frac{-R^2}{2\sigma^2}\right), \tag{5}$$

where σ is the smoothing distance (set to 1 by default) and R is the distance of a given earthquake from a given grid node. The kernel smoothing technique offers an intuitive representation of seismicity clustering in space: as a matter of fact, events that are close in space will mainly contribute to the same (few) nodes in the grid. The events count at each grid node is estimated by summing up the contributions from all the events in the grid to that specific point. Normalizing the smoothed rate by the total rate yields the expected earthquake density over all the grid nodes. The latter is used as the basis for assigning epicenter locations to a given grid point, i.e. with a probability that is proportional to the expected earthquake rate at that location. This is achieved by simply applying the discrete version of the inverse method to the cumulative distribution of the normalized smoothed rate. Once an epicenter has been linked to a specific grid point XY, its latitude (longitude) is simulated from an uniform distribution whose support is $([lat(XY) - sbin, lat(XY) + sbin]\ [lon(XY) - sbin, lon(XY) + sbin])$. Main steps are summarized in Algorithm 3:

Algorithm 3: Epicenters latitude and longitudes simulation

```
input: CUMSUM: Cumulative sum of the (sorted) smoothed rate
for each STAI gap do
    while count ≤ Number of earthquakes missing in the STAI gap do
        U = RAND(0,1);
        for each grid point XY do
            if CUMSUM(XY - 1) ≤ U < CUMSUM(XY) then
                U2 = RAND(0,1);
                LON = LON(XY − 1) + U2 · sbin;
                LAT = LAT(XY − 1) + U2 · sbin;
            end
        end
        count + = 1;
    end
end
```

2.6 Output

RESTORE replenishes the original catalog with the reconstructed events, by properly taking into account the occurrence time of the latter. The resulting catalog is saved in ZMAP format and differs from the original one only for two aspects: 1) the depth column is now a zeros vector, as this information has not been taken into account for the spatial simulation of missing earthquakes; 2) there is an additional column which flags events to 0 or 1, depending on whether they belong to the original catalog or they have been simulated. Additionally, several graphical outputs are returned:

- Time evolution of magnitude of completeness, with highlighted all detected STAI gaps (the plot neglects the seismic quiescent period);
- Magnitudes versus sequential numbers for the original and replenished catalogs: this is a great, tough qualitative, tool to highlight STAI issues which could possible affect earthquake magnitudes through time;
- Magnitude versus time for 1) the original catalog and the reconstructed events; 2) the original catalog only;
- Spatial map of the original events with overlapping reconstructed events;
- Magnitude-frequency distribution (MFD) for both the original and the replenished catalogs.

Finally, the magnitude of completeness is estimated for both the original and the replenished catalogs. This provides an additional test for validating the outputs by RESTORE: intuitively, we expect the M_c estimated for the replenished catalog to be very close to the pre-sequence value M_c^*. As for all the previous cases, this is done by means of the Lilliefors test. However, the user should keep in mind that the statistical power of the Lillieforst test (and, more in general, of the Kolmogorov-Smirnov test) greatly increases with the sample size (Stallone, 2018; Marzocchi et al., 2020). It follows that for a large number of events, which can be the case for the replenished catalog, the Lilliefors test becomes very sensitive to even slight deviations from an exponential distribution. This is not necessary ideal, since the detected departures could actually arise from magnitude errors. We therefore strongly recommend to inspect the magnitude of completeness of the replenished catalog by alternative methods as well, as those implemented in the ZMAP software (Wiemer, 2001).

3 Synthetic test

As a validation test, we implement numerical modeling, which enables us to control the number of missing events and their collocation in the magnitude-time-space domain. The goal is to check whether the algorithm is capable of reconstructing this information with an acceptable degree of accuracy. First, we simulate a seismic catalog by implementing the stochastic program described in (Felzer et al., 2002), which simulates the ETAS model (Ogata, 1988) as a branching process. In the original code, earthquakes with a magnitude larger than 6.5 are modeled as planar sources. We change that by modeling all the events as point sources. We use the program to simulate a 2-years-long catalog in Southern California, with magnitudes ranging from 2 to 6.9. We leave unchanged the remaining parameters needed for the simulation as indicated in the code. The b-value is set equal to 1. Since our next step is to simulate incompleteness of aftershocks following the largest earthquake in the catalog, we select a subset of the simulated dataset, which ranges from 1 year before to 3 months after the occurrence of the largest earthquake ($M = 6.9$). After this step, the original catalog includes 11,169 events. We simulate the STAI issue for the largest event by following the approach described in (Ogata & Katsura, 2006). Specifically, earthquakes are filtered out at a magnitude-dependent rate, according to the cumulative normal distribution:

$$F(M|\mu,\sigma) = \int_{-\infty}^{M} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \tag{6}$$

where μ and σ are constant: the first is the magnitude with a detection rate of 50%; the latter is the standard deviation of the normal distribution. $F(M|\mu,\sigma)$ is the probability of detection at magnitude M. See (Stallone, 2018) for more details. For our simulations, we set $\mu = 3$ and $\sigma = 0.2$; we assume that the magnitude of completeness is restored to the reference value 3 days after the occurrence of the large event. The catalog after STAI modeling includes 7,744 events. Figure 1 shows the frequency-magnitude distribution for the original (white circles) and incomplete (yellow circles) catalog, for which the STAI issue has been modeled. Figure 2 plots the magnitude of events as a func-

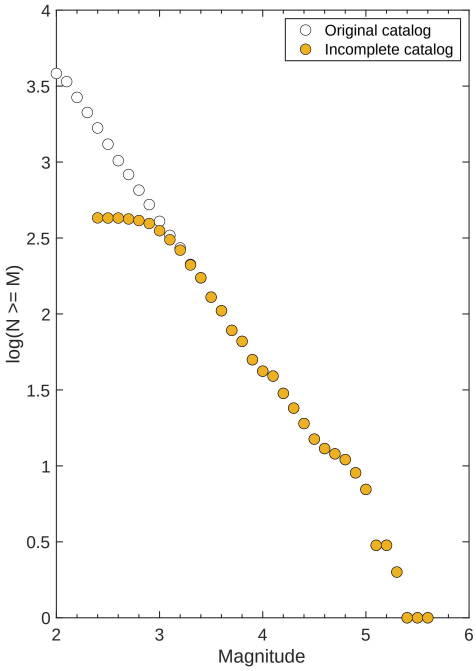

Figure 1. Frequency-magnitude distribution for the original synthetic catalog before STAI modeling (white circles) and after STAI modeling (yellow circles).

tion of time (over a period of 0-3 days from the mainshock) for the original (white circles) and incomplete (yellow circles) catalog.

As a next step, we implement RESTORE for reconstructing the missing events in the magnitude-time-space domain. We leave the default values for the window size (1000 events) and the step (250). The reference value for the magnitude of completeness equals the minimum magnitude in the synthetic catalog, i.e. 2.0. We set the b-value for the Gutenberg-Richter law to 1. Figure 3 shows some of the graphical outputs returned by the algorithm. We observe that occurrence times, magnitude range and locations of missing events have been correctly reconstructed. The replenished catalog includes $11,199$ events, i.e. 30 events more than the original synthetic catalog. The magnitude of completeness estimated by the Lilliefors test is 2.8 and 2.1 for the incomplete and replenished catalog, respectively. In order to further inspect the algorithm performance, we compare the frequency-magnitude distribution for 1) the original synthetic catalog before STAI modeling; 2) the original synthetic catalog after STAI modeling; 3) the replenished catalog. Results are shown in Figure 4. This comparison further proves the good performance of the algorithm when reconstructing missing events in the magnitude-time-space domain.

4 Real-case test (Amatrice earthquake)

We apply RESTORE to the 24 August 2016 Mw 6.2 Amatrice earthquake. The downloaded catalog covers the period from 1^{st} January 2016 to 30 September 2016 and includes 18,623 events. We leave the default values for the window size (1000 events) and the step (250). The seismically quiescent period ranges from 1^{st} January 2016 to 24 August 2016 and includes 2,351 events. We estimate the reference value for the magnitude of completeness M_c^* with the Lilliefors test provided by the algorithm, which returns $M_c^* =$

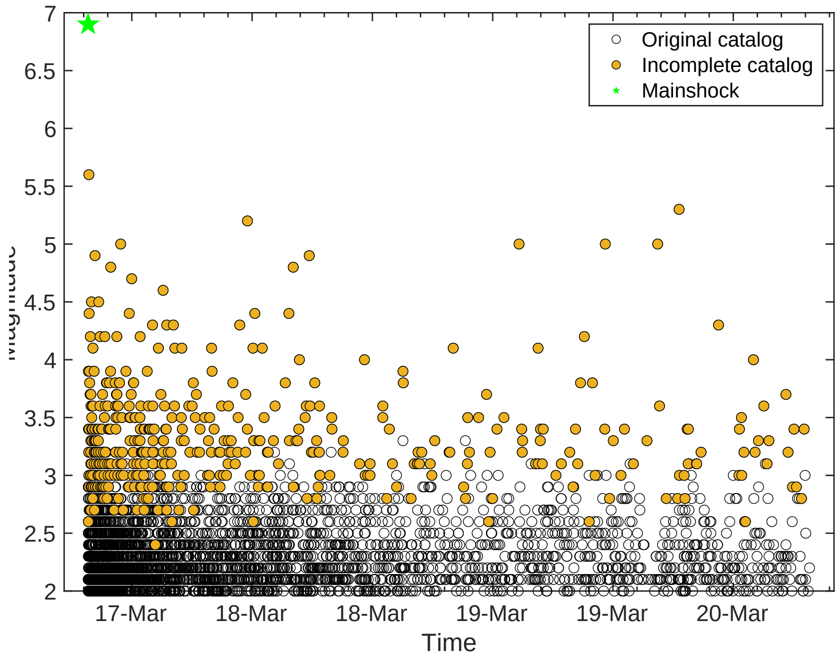

Figure 2. Magnitude-time plot for events occurred within 3 days from the large shock. White circles: before STAI modeling. Yellow circles: after STAI modeling.

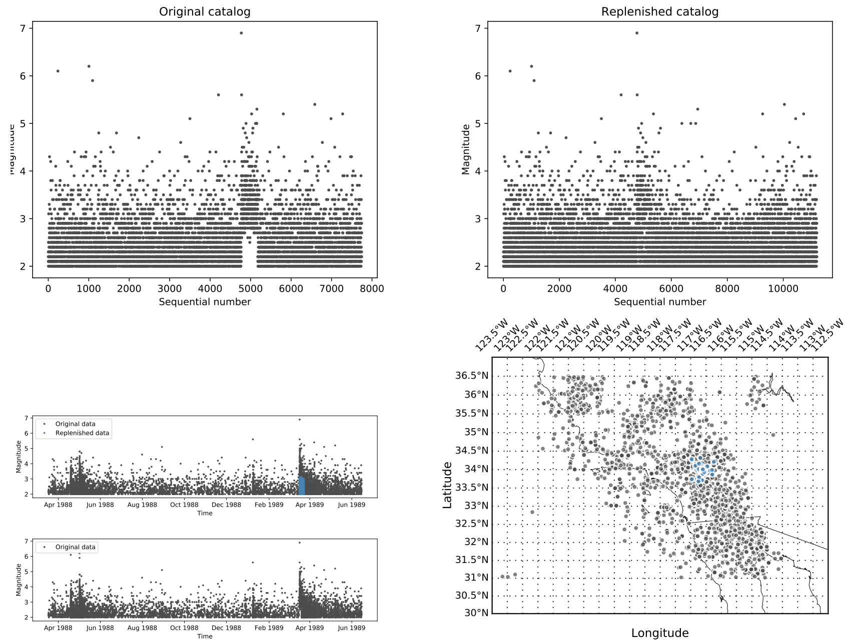

Figure 3. Main graphical outputs of the algorithm. Top Left: Magnitudes versus sequential numbers for the original (synthetic) catalog; Top Right: Magnitudes versus sequential numbers for the replenished catalog; Bottom Left: Magnitude versus time for 1) the original catalog and the reconstructed events 2) the original catalog only; Bottom Right: Spatial map of the original events with overlapping reconstructed events.

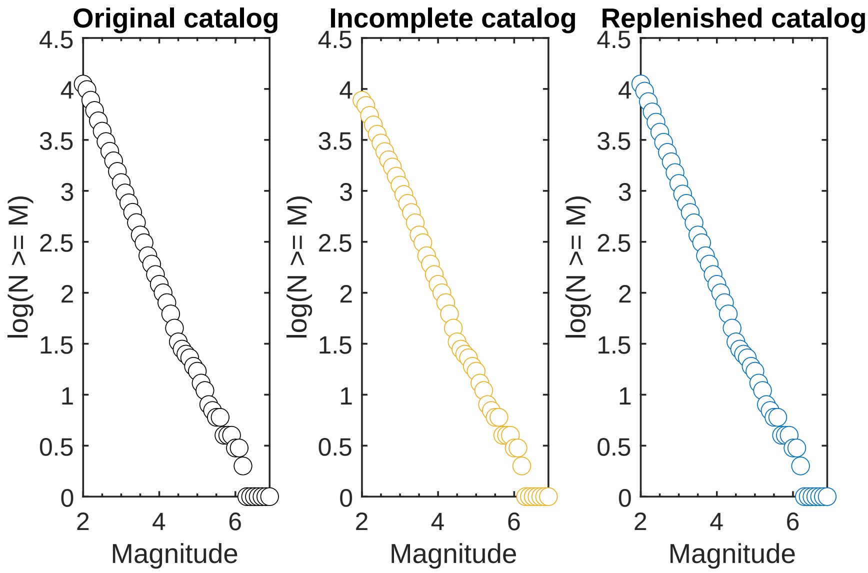

Figure 4. Frequency-magnitude distribution. From left to right: original synthetic catalog before STAI modeling, original synthetic catalog after STAI modeling, replenished catalog.

1.3. This leaves 11,429 earthquakes with $M \geq M_c^*$. Finally, we set the b-value for the Gutenberg-Richter law equal to 1. The replenished catalog includes 17,428 events. Figure 5 plots the magnitude of completeness as a function of time, with highlighted the detected STAI gaps (four in this case). The magnitude of completeness is recovered to the reference value M_c^* after about 1 month. Figure 6 shows the other graphical outputs returned by the algorithm. While the ground truth is not known in the real-case test, we observe that the missing events are correctly reconstructed in a way which is consistent with data.

5 Conclusions

We have presented RESTORE, a new Python toolbox for the reconstruction of magnitude, time and location of events missed in the coda of large shocks. It relies on very few assumptions - e.g. the detection rate of events can be assumed to be constant within periods of time that are much shorter than the STAI extent. It also relies on a data-driven approach, which is built on well-known empirical properties of earthquake data, such as the Gutenberg-Richter law for the frequency-magnitude distribution and the aftershocks clustering in space. The critical subsets of the catalog that are affected by STAI are automatically detected through a moving-window approach, which identifies statistically significant departures of the magnitude of completeness with respect to a reference value. We demonstrate the robustness of the algorithm by means of a numerical and a real-case test. In the first case, the ground truth is accurately recovered: not only the number of missing earthquakes is correctly retrieved, but their space-time-magnitude stochastic distribution is correctly resolved as well. The real-test case, which applies to the Mw 6.2 Amatrice earthquake, further proves the good performance of the algorithm, which reconstructs the missed events in a way that is consistent with the data. The main advantage of RESTORE lies in its fully data-driven approach. However, this could also represent a drawback if the following aspects are not carefully taken into consideration:

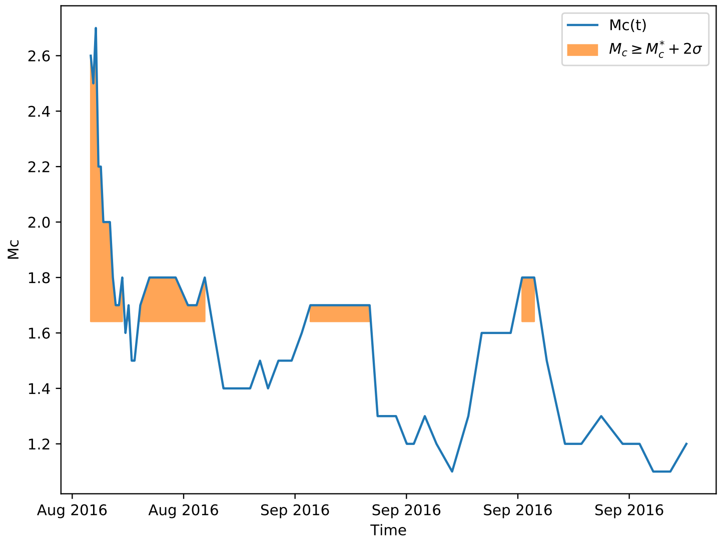

Figure 5. Temporal evolution of the magnitude of completeness, with highlighted the detected STAI gaps. The moving-window includes 1000 events and is shifted by 250 events.

- the quality of the seismic catalog: strong uncertainties about the earthquake parameters (epicenter coordinates, magnitude, occurrence time) will affect the properties of the simulated events;
- the seismic quiescent period: it must be carefully selected for an accurate estimation of the reference value of the magnitude of completeness. Furthermore, it must be long enough to include a number of events which must be substantially higher than the chosen window size; for an unbiased estimation of M_c^*, the user is required to select the quiescent period so to include a number of events N which is a multiple of the selected window size k (we recommend at least $N \simeq 4 * k$);
- spatial map domain: the same reasoning for the seismic quiescent period applies here as well; the selected area should obviously include the large shock/s and, at the same time, enough events in the seismic quiescent period;
- the window size and step: the output provided by RESTORE will be affected by the values provided for these parameters; we recommend to test several alternatives and opt for those assuring the best replenishment. As detailed in the text, too small values for the size and step will likely bias the M_c estimate, whereas too large values will shadow short-term fluctuations of M_c.

RESTORE is made freely available and can be downloaded at the link provided in the Acknowledgments. It promises to become a valuable research tool to tackle the STAI issue, which can severely bias any study based on the analysis of real seismic catalogs. Hopefully, it will help reducing these sources of bias, thus leading to better operational earthquake forecasting and seismic hazard assessment.

6 Data Availability Statement

The algorithm RESTORE is available at the following Zenodo repository: `https://doi.org/10.5281/zenodo.3952182`, and can also be downloaded from GitHub at this

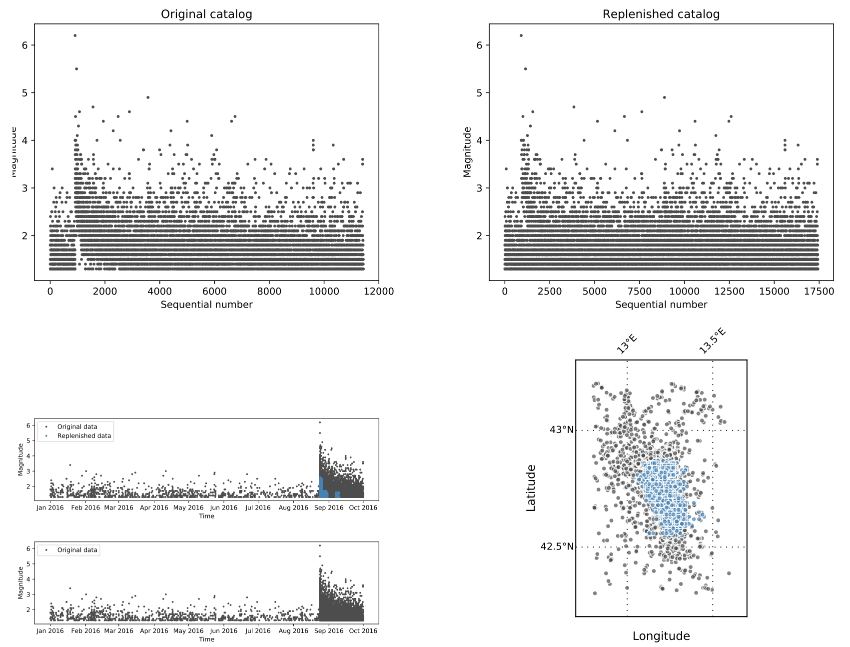

Figure 6. Main graphical outputs of the algorithm. Top Left: Magnitudes versus sequential numbers for the original catalog; Top Right: Magnitudes versus sequential numbers for the replenished catalog; Bottom Left: Magnitude versus time for 1) the original catalog and the reconstructed events 2) the original catalog only; Bottom Right: Spatial map of the original events with overlapping reconstructed events.

link: `https://github.com/angystallone/RESTORE`. The repository includes the dataset used for the synthetic test as well. The seismic catalog used for the real-case test (Amatrice earthquake) is the HOmogenized instRUmental Seismic catalog (HORUS) of Italy (Lolli et al., 2020) and it can be downloaded at this link: `https://horus.bo.ingv.it/`. The routine *Mc-Lilliefors* implemented in RESTORE for the magnitude of completeness estimation is available at the following Zenodo repository: `https://doi.org/10.5281/zenodo.4162496`.

Acknowledgments

This project has been founded by the Seismic Hazard Center (Centro di Pericolosità Sismica, CPS, at the Istituto Nazionale di Geofisica e Vulcanologia, INGV).

Appendix A Calculation of number of missing events

Here we derive Equation 1 in the text. The frequency-magnitude distribution of earthquakes is typically described by the Gutenberg-Richter (G-R) exponential law (Gutenberg & Richter, 1944):

$$N(M) = 10^{a-bM}, \tag{A1}$$

where $N(M)$ is the number of events with magnitude above M ($M >= Mmin$, i.e. the minimum magnitude in the earthquake catalog), a is a constant related to the total seismic rate and b is the b-value, controlling the relative number of large earthquakes in the catalog. Let us consider the case where $M_2 \geq M_1$. We have:

$$N(M \geq M_1) = 10^{a-bM_1}$$
$$N(M \geq M_2) = 10^{a-bM_2}$$

We start by expressing $N(M \geq M_1)$ as a function of $N(M \geq M_2)$ and b only, by calculating the ratio:

$$\frac{N(M \geq M_1)}{N(M \geq M_2) = 10^{-b(M_1-M_2)}} \tag{A2}$$

This simple trick enables us to rescale the problem, i.e. to get rid of the term 10^a, which is related to the total seismic rate:

$$N(M \geq M_1) = N(M \geq M_2) \cdot 10^{-b(M_1-M_2)} \tag{A3}$$

We observe that $M_2 = M_1 + n \cdot mbin$, where $mbin$ is the magnitude binning (usually equal to 0.1). It follows that:

$$N(M \geq M_1) = N(M \geq M_2) \cdot 10^{b \cdot n \cdot mbin} \tag{A4}$$

For one bin only (i.e., n = 1):

$$N(M \geq M_1) = N(M \geq M_2) \cdot 10^{b \cdot mbin} \tag{A5}$$

This equation allows to retrieve the number of expected events with magnitude $M \geq M_1$ as a function of the number of events with magnitude $M \geq M_2$. In other words, we can extrapolate the frequency of earthquakes above a given magnitude to any lower magnitude cutoff. Note that we implicitly assume the b-value is constant for any subset of the whole catalog.

References

Brodsky, E. (2011). The spatial density of foreshocks. *Geophysical Research Letters*, *38*(10).

Christophersen, A., & Smith, E. G. (2008). Foreshock rates from aftershock abundance. *Bulletin of the Seismological Society of America*, *98*(5), 2133–2148.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, *51*(4), 661–703.

Devroye, L. (1986). Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on winter simulation* (pp. 260–265).

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics* (pp. 569–593). Springer.

Enescu, B., Mori, J., & Miyazawa, M. (2007). Quantifying early aftershock activity of the 2004 mid-niigata prefecture earthquake (mw6. 6). *Journal of Geophysical Research: Solid Earth*, *112*(B4).

Enescu, B., Mori, J., Miyazawa, M., & Kano, Y. (2009). Omori-utsu law c-values associated with recent moderate earthquakes in japan. *Bulletin of the Seismological Society of America*, *99*(2A), 884–891.

Felzer, K. R., Becker, T. W., Abercrombie, R. E., Ekström, G., & Rice, J. R. (2002). Triggering of the 1999 mw 7.1 hector mine earthquake by aftershocks of the 1992 mw 7.3 landers earthquake. *Journal of Geophysical Research: Solid Earth*, *107*(B9), ESE–6.

Felzer, K. R., Page, M. T., & Michael, A. J. (2015). Artificial seismic acceleration. *Nature Geoscience*, *8*(2), 82–83.

Frankel, A. (1995). Mapping seismic hazard in the central and eastern united states. *Seismological Research Letters*, *66*(4), 8–21.

Gutenberg, B., & Richter, C. F. (1944). Frequency of earthquakes in california. *Bulletin of the Seismological Society of America*, *34*(4), 185–188.

Hainzl, S., Zakharova, O., & Marsan, D. (2013). Impact of aseismic transients on the estimation of aftershock productivity parameters. *Bulletin of the Seismological Society of America*, *103*(3), 1723–1732.

Helmstetter, A., Kagan, Y. Y., & Jackson, D. D. (2005). Importance of small earthquakes for stress transfers and earthquake triggering. *Journal of Geophysical Research: Solid Earth*, *110*(B5).

Helmstetter, A., Kagan, Y. Y., & Jackson, D. D. (2006). Comparison of short-term and time-independent earthquake forecast models for southern california. *Bulletin of the Seismological Society of America*, *96*(1), 90–106.

Herrmann, M., & Marzocchi, W. (2020a). Inconsistencies and lurking pitfalls in the magnitude–frequency distribution of high-resolution earthquake catalogs. *Seismological Research Letters*.

Herrmann, M., & Marzocchi, W. (2020b, November). *Mc-Lilliefors: a completeness magnitude that complies with the exponential-like Gutenberg–Richter relation.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.4162497` doi: 10.5281/zenodo.4162497

Iwata, T. (2008). Low detection capability of global earthquakes after the occurrence of large earthquakes: Investigation of the harvard cmt catalogue. *Geophysical Journal International*, *174*(3), 849–856.

Kagan, Y. Y. (2004). Short-term properties of earthquake catalogs and models of earthquake source. *Bulletin of the Seismological Society of America*, *94*(4), 1207–1228.

Kagan, Y. Y. (2013). *Earthquakes: models, statistics, testable forecasts.* John Wiley & Sons.

Knopoff, L., Kagan, Y. Y., & Knopoff, R. (1982). b values for foreshocks and aftershocks in real and simulated earthquake sequences. *Bulletin of the Seismological Society of America*, *72*(5), 1663–1676.

Lilliefors, H. W. (1969). On the kolmogorov-smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, *64*(325), 387–389.

Lippiello, E., Godano, C., & de Arcangelis, L. (2012). The earthquake magnitude is influenced by previous seismicity. *Geophysical Research Letters*, *39*(5).

Lolli, B., Randazzo, D., Vannucci, G., & Gasperini, P. (2020). The homogenized instrumental seismic catalog (horus) of italy from 1960 to present. *Seismological Society of America*, *91*(6), 3208–3222.

Mai, P. M., & Beroza, G. C. (2000). Source scaling properties from finite-fault-rupture models. *Bulletin of the Seismological Society of America*, *90*(3), 604–615.

Marzocchi, W., Spassiani, I., Stallone, A., & Taroni, M. (2020). How to be fooled searching for significant variations of the b-value. *Geophysical Journal International*, *220*(3), 1845–1856.

Mignan, A., & Woessner, J. (2012a). Estimating the magnitude of completeness for earthquake catalogs. *Community Online Resource for Statistical Seismicity Analysis*, 1–45.

Mignan, A., & Woessner, J. (2012b). *Theme iv—understanding seismicity catalogs and their problems* (Tech. Rep.). Technical Report doi: https://doi.org/10.5078/corssa-00180805, Community ....

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.

Ogata, Y., & Katsura, K. (1993). Analysis of temporal and spatial heterogeneity of magnitude frequency distribution inferred from earthquake catalogues. *Geophysical Journal International*, *113*(3), 727–738.

Ogata, Y., & Katsura, K. (2006). Immediate and updated forecasting of aftershock hazard. *Geophysical research letters*, *33*(10).

Omi, T., Ogata, Y., Hirata, Y., & Aihara, K. (2013). Forecasting large aftershocks within one day after the main shock. *Scientific reports*, *3*, 2218.

Omi, T., Ogata, Y., Hirata, Y., & Aihara, K. (2014). Estimating the etas model from an early aftershock sequence. *Geophysical Research Letters*, *41*(3), 850–857.

Peng, Z., Vidale, J. E., & Houston, H. (2006). Anomalous early aftershock decay rate of the 2004 mw6. 0 parkfield, california, earthquake. *Geophysical Research Letters*, *33*(17).

Peng, Z., Vidale, J. E., Ishii, M., & Helmstetter, A. (2007). Seismicity rate immediately before and after main shock rupture from high-frequency waveforms in japan. *Journal of Geophysical Research: Solid Earth*, *112*(B3).

Peng, Z., & Zhao, P. (2009). Migration of early aftershocks following the 2004 parkfield earthquake. *Nature Geoscience*, *2*(12), 877–881.

Schorlemmer, D., Neri, G., Wiemer, S., & Mostaccio, A. (2003). Stability and significance tests for b-value anomalies: Example from the tyrrhenian sea. *Geophysical research letters*, *30*(16).

Seif, S., Mignan, A., Zechar, J. D., Werner, M. J., & Wiemer, S. (2017). Estimating etas: The effects of truncation, missing data, and model assumptions. *Journal of Geophysical Research: Solid Earth*, *122*(1), 449–469.

Stallone, A. (2018). *Statistical analysis of earthquake occurrences and implications for earthquake forecasting and seismic hazard assessment* (Doctoral dissertation, University of Roma TRE). doi: 10.13140/RG.2.2.35672.65282/1

Stallone, A., & Marzocchi, W. (2019). Empirical evaluation of the magnitude-independence assumption. *Geophysical Journal International*, *216*(2), 820–839.

Wiemer, S. (2001). A software package to analyze seismicity: Zmap. *Seismological Research Letters*, *72*(3), 373–382.

Woessner, J., Laurentiu, D., Giardini, D., Crowley, H., Cotton, F., Grünthal, G., ...

others (2015). The 2013 european seismic hazard model: key components and results. *Bulletin of Earthquake Engineering*, *13*(12), 3553–3596.

Woessner, J., & Wiemer, S. (2005). Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty. *Bulletin of the Seismological Society of America*, *95*(2), 684–698.

Zechar, J. D., & Jordan, T. H. (2010). Simple smoothed seismicity earthquake forecasts for italy. *Annals of Geophysics*, *53*(3), 99–105.

Zhuang, J., Ogata, Y., & Wang, T. (2017). Data completeness of the kumamoto earthquake sequence in the jma catalog and its influence on the estimation of the etas parameters. *Earth, Planets and Space*, *69*(1), 36.

Zhuang, J., Wang, T., & Koji, K. (2019). Detection and replenishment of missing data in marked point processes.. Retrieved from `http://bemlar.ism.ac.jp/zhuang/pubs/zhuang2019statsini.pdf`