Detection and replenishment of missing data in earthquake catalogs

Jiancang Zhuang
Institute of Statistical Mathematics zhuang@ism.ac.jp

Based on

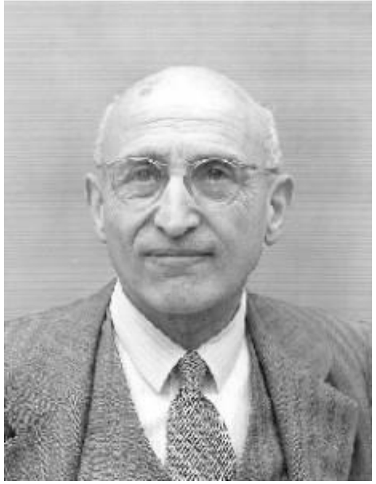
1. Zhuang, J., T. Wang, K. KiyoSugi (2019) Detection and replenishment of missing data in marked point processes. Statistica Sinica. *published online.*
2. Zhuang, J., Y. Ogata, T. Wang (2017) Earth, Planet, and Space.

报告提纲

- 模型问题
- 数据问题

地震震级模型和b值

Gutenberg-Richter magnitude-frequency relation



G-R magnitude-frequency relation

$$\log_{10} N(\geq m) = a - bm$$

$$\Pr\{M > m | M > m_0\} \approx \frac{N(\geq m)}{N(\geq m_0)}$$

$$= \frac{10^{a-bm}}{10^{a-bm_0}} = 10^{-b(m-m_0)}$$

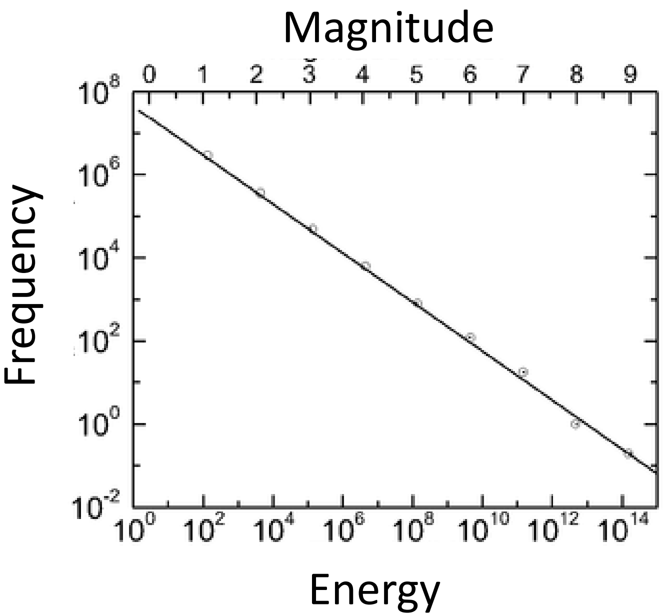
Probability density function

$$f(m) = b\, 10^{-b(m-m_0)}\, ln\, 10$$
$$= \beta e^{-\beta(m-m_0)}\,; \quad m > m_0$$

Power law distribution for energies, moments and stress drops.

$$\Pr\{E > x\} \sim C x^{-\alpha}$$

for $\alpha > 0,\ x > E_0 > 0.$



Truncated exponential distribution

$$f(m) = \begin{cases} \dfrac{\beta e^{-\beta(m-m_0)}}{1-e^{-\beta(M-m_0)}}, & m_0 \le m \le M \\ 0, & \text{otherwise} \end{cases}$$

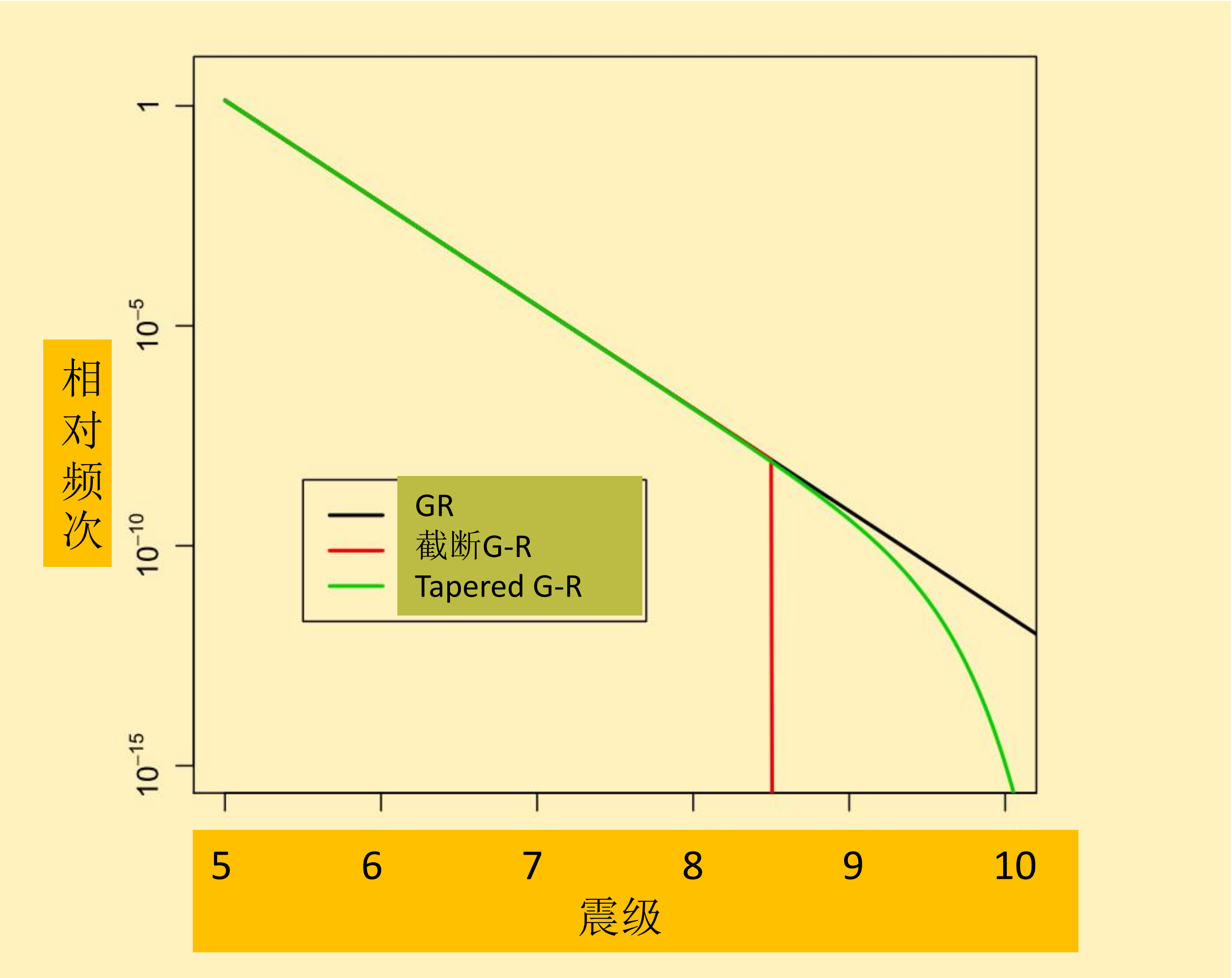
Tapered exponential distribution

$$f(m) = \left[\beta + 1.5\gamma 10^{1.5(m-m_c)} \ln 10\right] \exp\left[-\beta(m-m_0) - \gamma 10^{1.5(m-m_c)}\right]$$

m_c: corner magnitude

In moment, $\Pr\{\text{moment} < x\} = 1 - \left(\frac{x}{S_0}\right)^{-k} \exp\left[\frac{S_0 - s}{S_c}\right]$ **(Kagan) distribution**

S_c: corner moment


相对频次
1
10^{-5}
10^{-10}
10^{-15}
GR
截断G-R
Tapered G-R
5
6
7
8
9
10
震级

b值的估计和地震震级的模拟

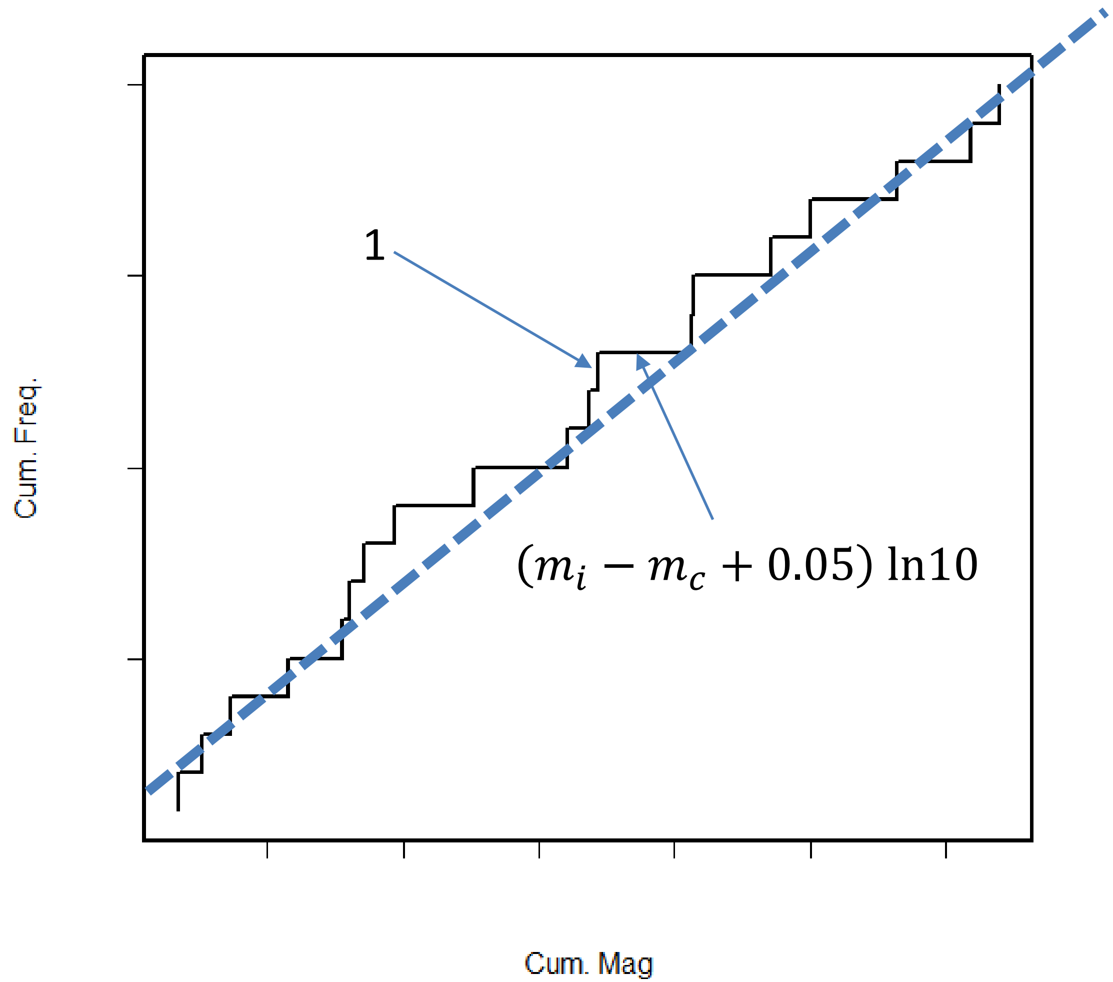
(1) b-值 最大似然估计

$$\hat{b} = \frac{1}{\left(\bar{M} - m_c + \frac{\Delta}{2}\right)\ln 10}$$

平均震级　震级下限　震级最低位

$$std.err\left(\hat{b}\right) = \frac{\hat{b}}{\sqrt{n}}$$

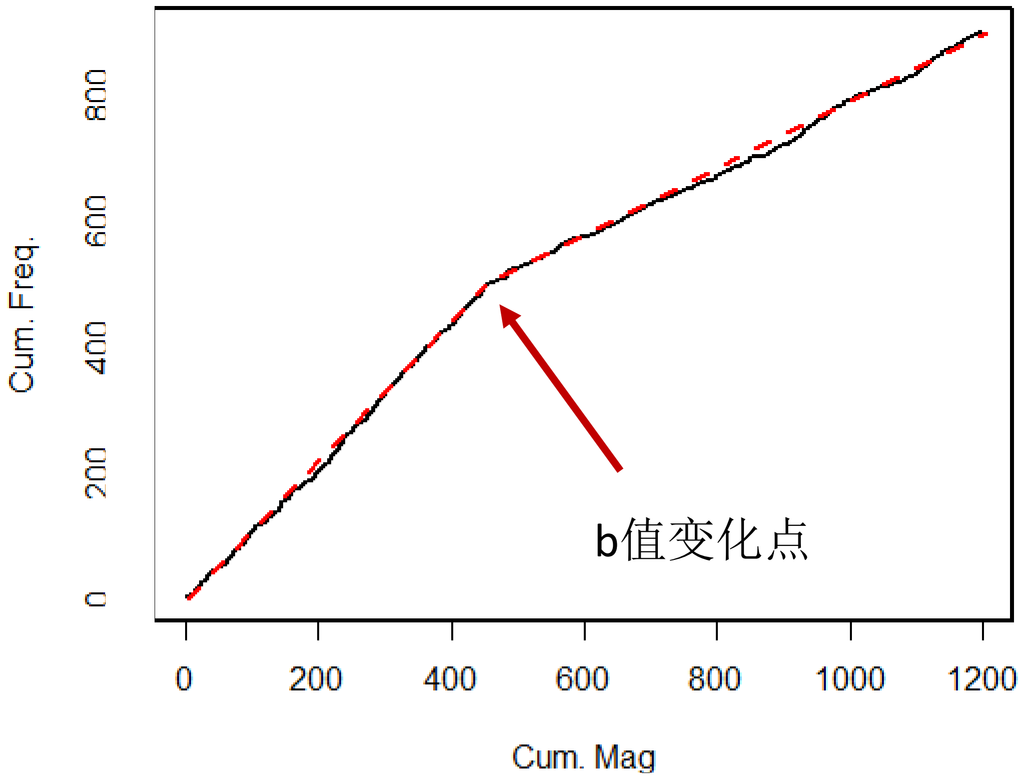
(1) b-值 最大似然估计　 绘图法



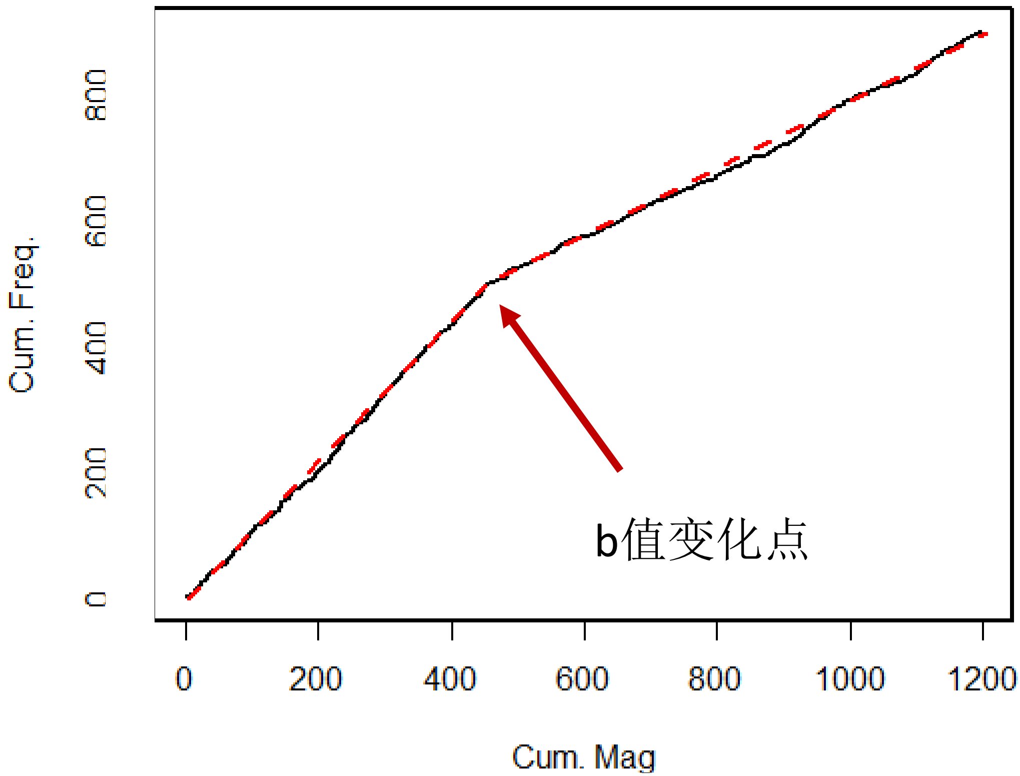
斜率 = b值

$$i \sim \sum_{k=1}^{i} m_k$$

(1) b-值 最大似然估计　　绘图法检验b值变化点



(1) b-值 最大似然估计　绘图法检验b值变化点



把地震分别按照发震时刻、纬度、经度、深度、或者沿断层投影的位置进行排序，就可以看b值在时间、纬度、经度、深度、或者断层上的变化。

短期地震活动模型
--- Omori-Utsu公式

4. Omori-Utsu formula

- Omori (1894): the rate of felt aftershocks of the 1891 $M_s 8.0$ Nobi earthquake,

$$n(t) = K(t+c)^{-1}, \tag{26}$$

t: the time from the mainshock.
K and c: constants.

- Utsu (1957): the decay of the aftershock numbers could vary. (Utusu-Omori formula)

$$n(t) = K(t+c)^{-p} \tag{27}$$

p: ranges between 0.6 and 2.5 with a median of 1.1.

- Reasenberg-Jones model

$$\lambda(t,m) = \frac{K\, s(m)}{(t+c)^p}, \tag{28}$$

$s(m)$: the magnitude probability density function.

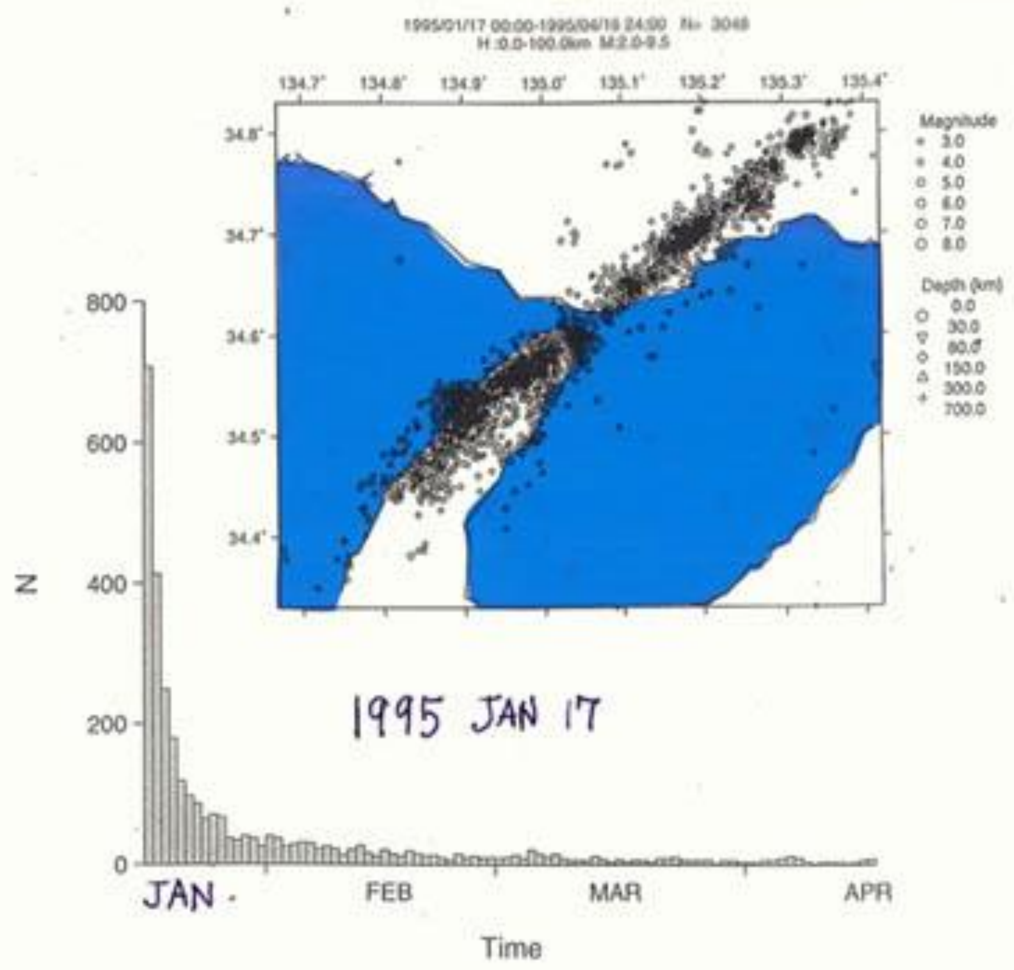
- Multiple Omori-Utsu formula: Not only mainshocks trigger aftershocks, but also large aftershocks may trigger their own aftershocks.

$$\lambda(t) = K/(t-t_0+c)^{-p} + \sum_{i=1}^{N_T} \frac{K_i H(t-t_i)}{(t-t_i+c_i)^{-p_i}}, \tag{29}$$

t_0: the occurrence time of the mainshock;
t_i, $i = 1, \ldots N_T$: the occurrence times of the triggering aftershocks;
H: Heaviside function.

Likelihood function

$$\log L = \sum_{t_i \in [0,T]} \log \lambda(t_i) - \int_0^T \lambda(t)dt$$



$$n(t) = \frac{K}{(t+c)^p}$$

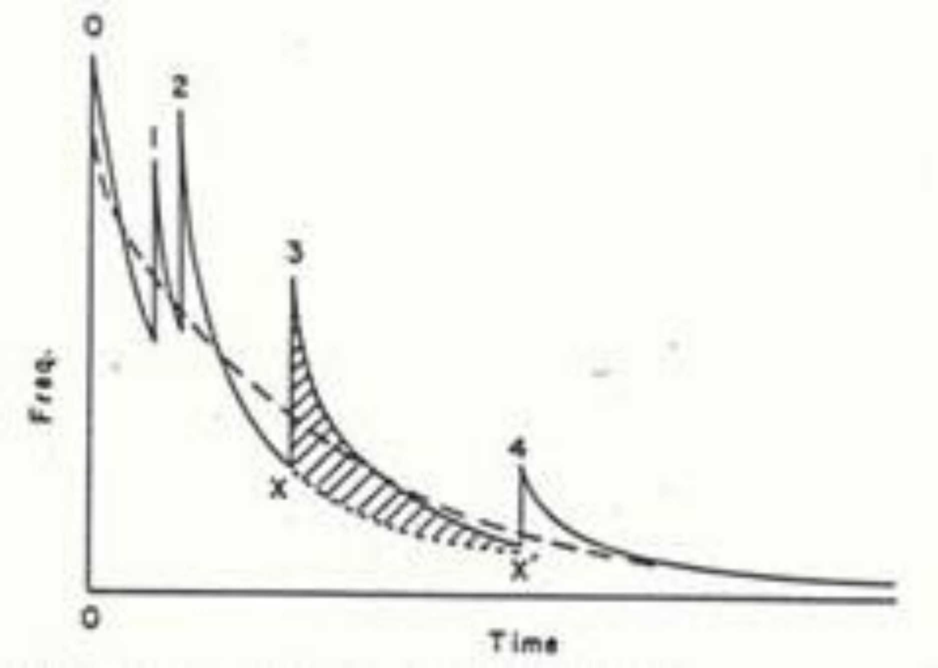
Omori-Utsu Formula (Utsu 1956)

$$n(t) = K(t+c)^{-p} + H(t-T_1)K_1(t-T_1+c_1)^{-p_1} + \ldots + H(t-T_n)K_n(t-T_n+c_n)^{-p_n}$$

(Utsu, 1970; Ogata, 1983)

$$\lambda(t) = \mu + \sum_{t_i<t} \frac{K_i}{(t-t_i+c)^p} = \mu + K_0 \sum_{t_i<t} \frac{e^{\alpha(M_i-M_0)}}{(t-t_i+c)^p}$$

(Ogata, 1988; 1989)



$\lambda(t)$: Conditional intensity, hazard function conditioning on the past history

Temporal ETAS model

- Conditional intensity

$$\lambda(t) = \mu + \sum_{i:t_i<t} \kappa(m_i) g(t-t_i)$$

1. Direct productivity: $\kappa(m) = Ae^{\alpha(m-m_C)}, \quad m \geq m_C$

2. Time p.d.f (Omori-Utsu): $g(t) = (p-1)\left(1+t/c\right)^{-p}/c, \quad t>0$

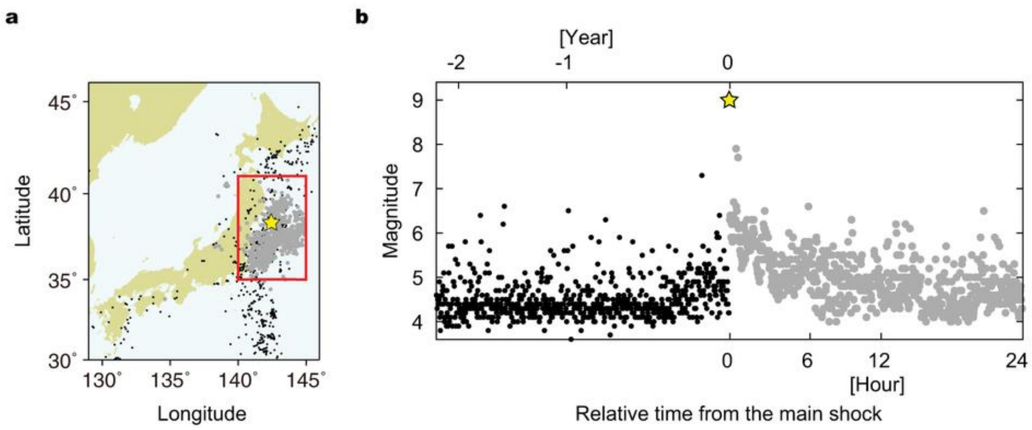
- Likelihood function $\log L = \sum_{t_i\in[0,T]} \log \lambda(t_i) - \int_0^T \lambda(t)dt$

(Ogata, 1988)

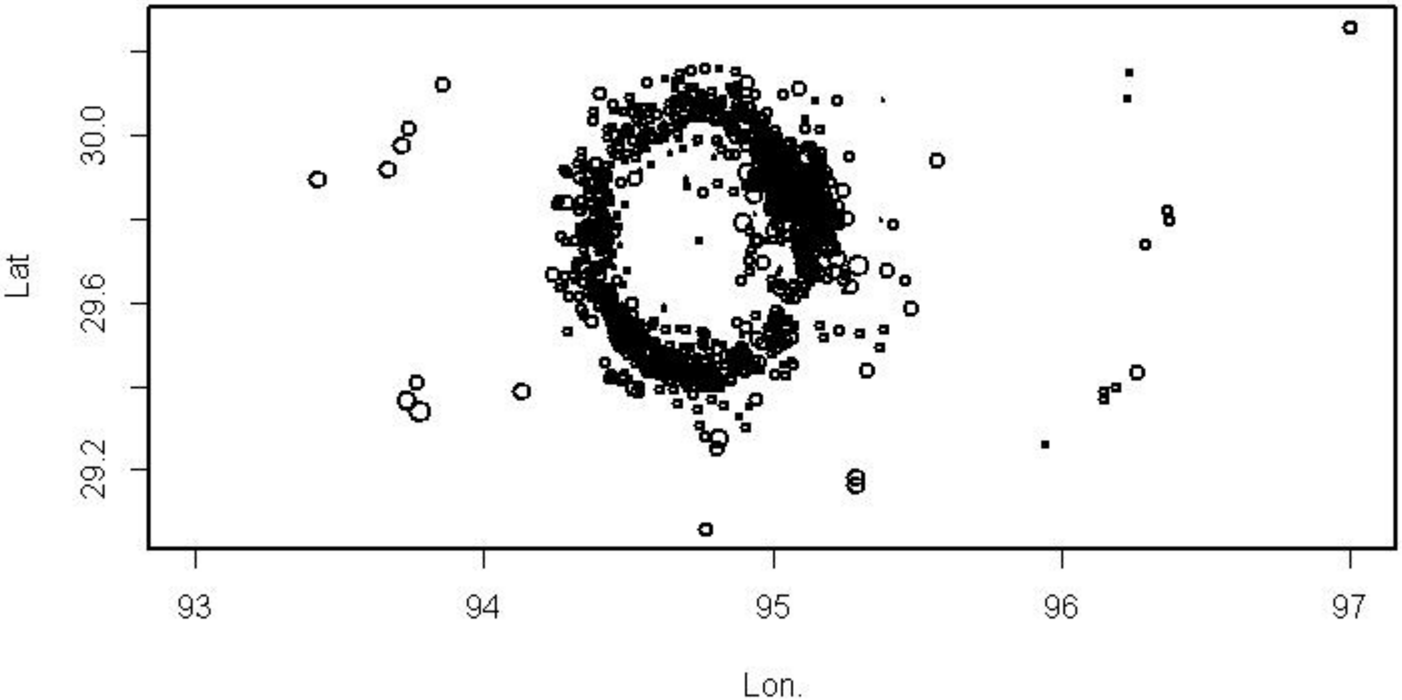
- 数据问题　　直后余震缺失问题

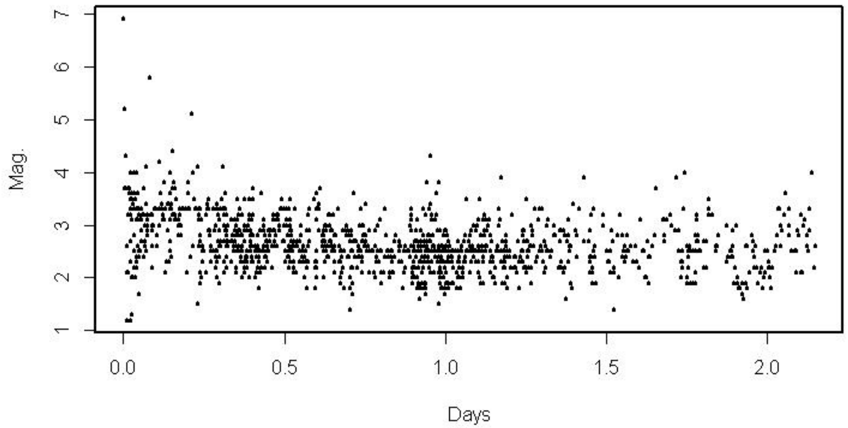
Short-term missing of aftershocks

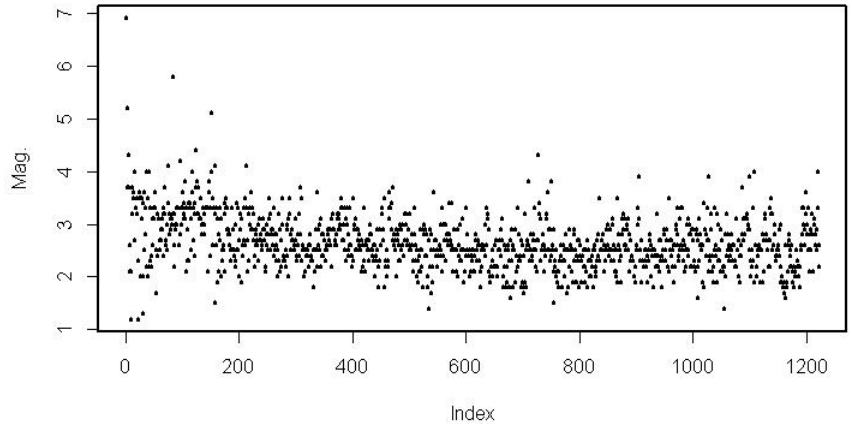
Missing events after the mainshock (from Omi etc, 2013)



林芝地震 (2017-11-18 06:34:17.30 29.875 95.057 M6.9)







跳开最震后短时间内的地震进行拟合，结合使用较大震级下限

- 汶川地震 M4.0以上

The following is the earthquake probability forecast
for the 31th (2008-6-12 14:28 to 2008-6-13 14:28) day after the mainshock

	Expected #	Prob.	Expected.WT	50%	95%	99% waiting time
M>=4.0	1.14	0.68	0.87	0.60	2.59	3.97
M>=4.5	0.54	0.41	1.83	1.26	5.50	8.48
M>=5.0	0.18	0.16	5.41	3.79	16.15	24.57
M>=5.5	0.05	0.05	16.86	12.48	46.94	72.50
M>=6.0	0.03	0.03	25.21	18.63	71.77	117.31

2008-07-15 17:26 15 Jul 2008 09:26 Mianzhu, Sichuan 31.57 103.98 5.0 Ms

2008-07-24 03:54 23 Jul 2008 19:54 Ningqiang, Shaanxi 32.8 105.5 5.6 Ms

2008-07-24 15:09 24 Jul 2008 07:09 Qingchuan, Sichuan 32.82 105.47 6.0 Ms

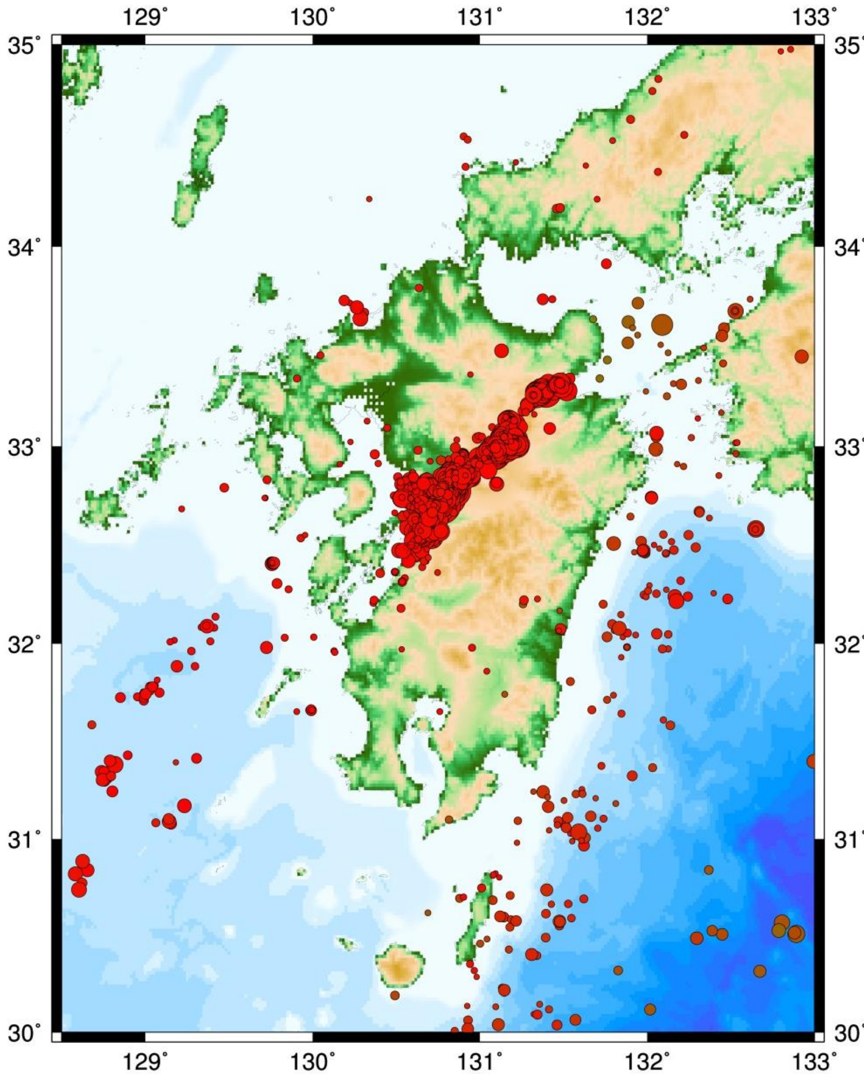
2008-08-01 16:32 01 Aug 2008 08:32 Pingwu and Beichuan, Sichuan 32.1 104.7 6.1 Ms

The Kumamoto aftershock sequence data

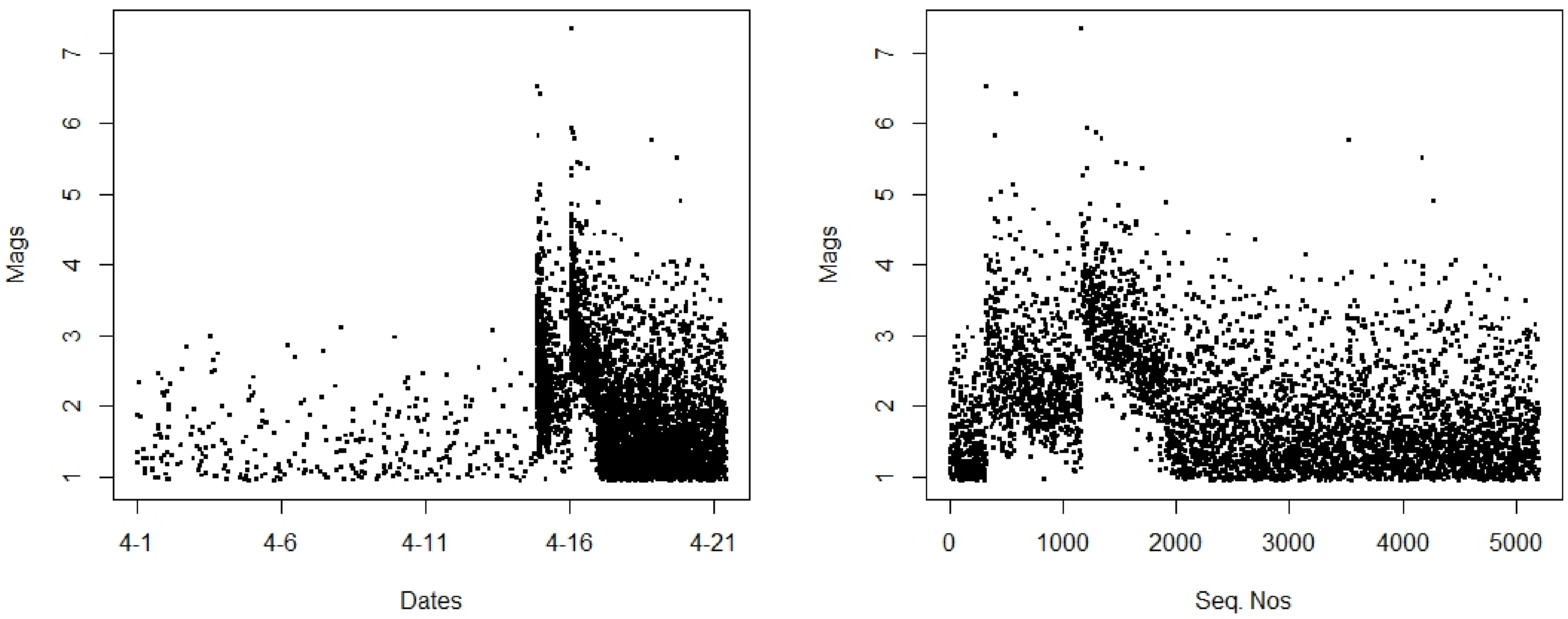
Data Selection

Time: 2016/4/1~2016/4/21
Mag.: 1.0+
Depth: < 100 km
Space: 128° -- 133°
30° -- 35°

2016-04-14 21:26 (130.81 32.74) M6.5
2016-04-15 00:03 (130.78 32.70) M6.4
2016-04-16 01:25 (130.76 32.75) M7.3



Application to the recent Kumamoto aftershock sequence data



Previous studies for fixing the problems of short-term missing aftershocks

Observational approaches

- waveform-based earthquake detection methods (e.g., Enescu et al., 2007, 2009; Peng et al., 2007; Marsan and Enescu, 2012; Hainzl, 2016).
- Energy based description (Sawazaki and Enescu, 2014)

Statistical approaches

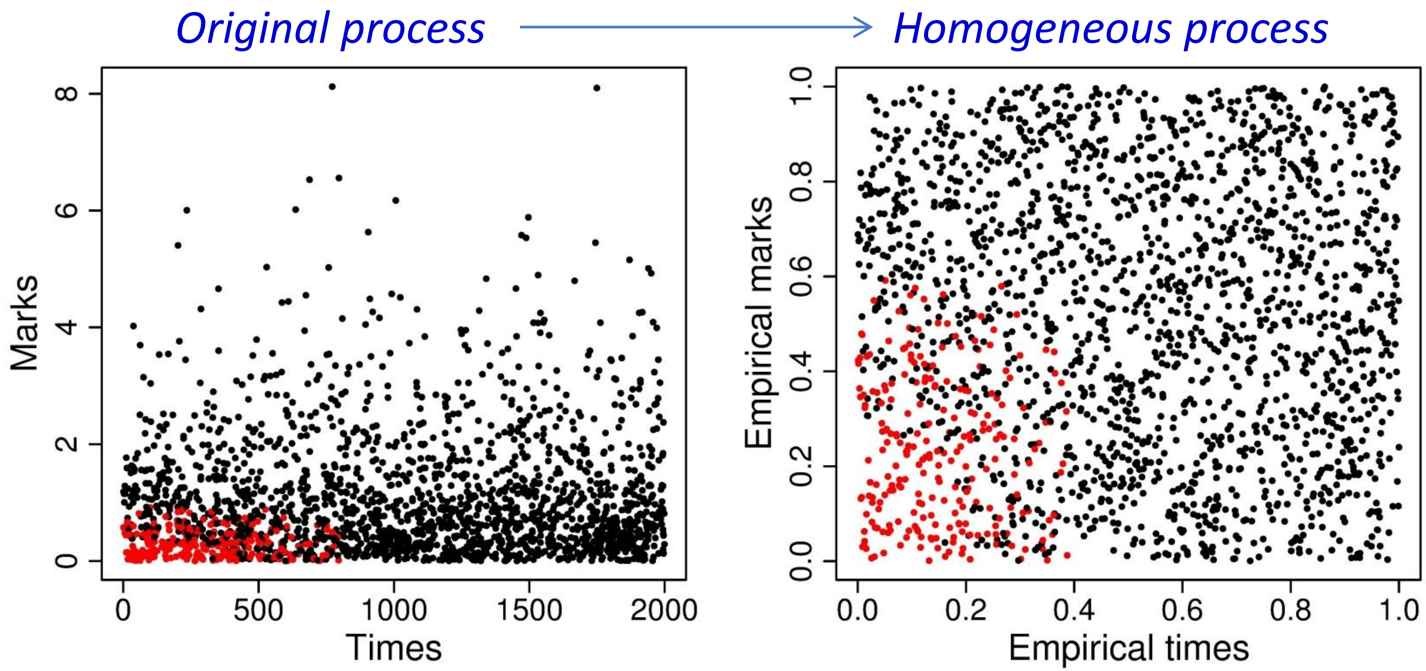
- (Ogata, Omi, Iwata) Bayesian, assuming GR relation for whole range
- (Marsan and Enescu, 2012) Assuming Omori-Utsu formula or ETAS model
- This study: Independence between magnitudes and occurrence times

When data is complete ……

Biscale empirical transformation

$$t_i \to \quad \tau_i = \frac{i}{N} = \frac{\#\ of\ times < t_i}{N}$$

$$m_i \to \quad s_i = \frac{\sum_k \mathbf{1}(m_k < m_i)}{N} = \frac{\#\ of\ marks < m_i}{N}$$



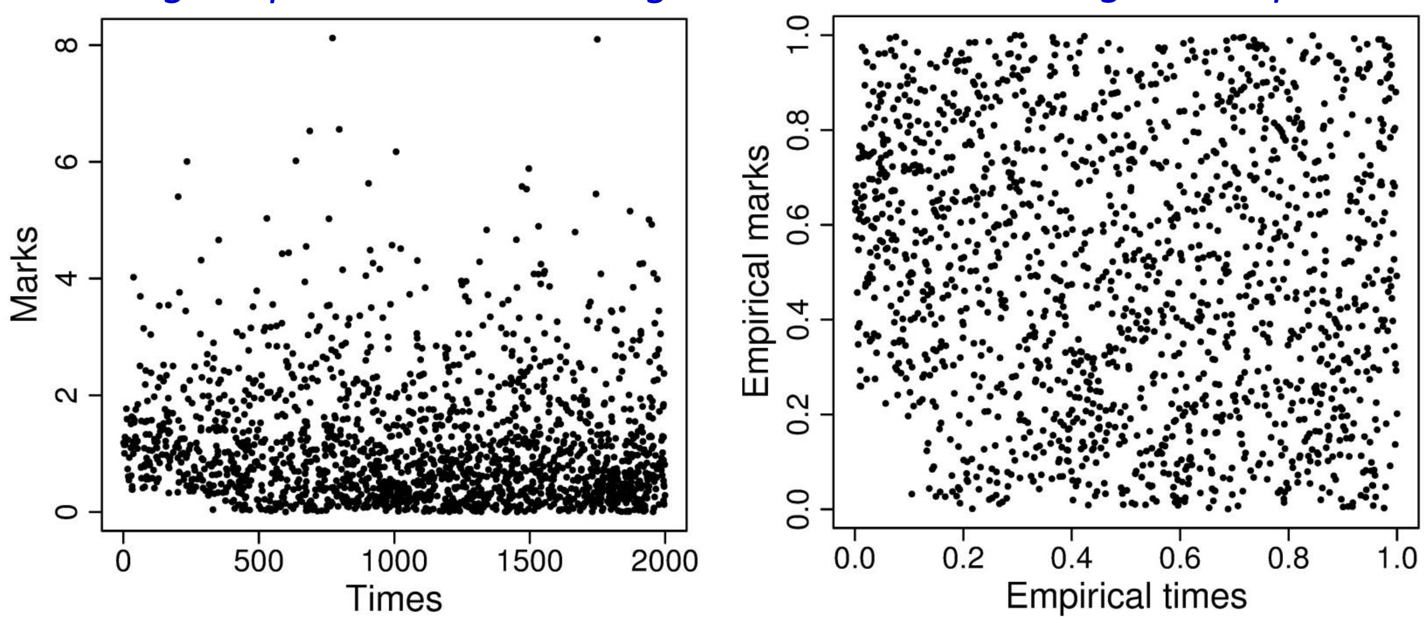
When data missing exists……

Biscale empirical transformation

$$t_i \to \quad \tau_i = \frac{i}{N} = \frac{\#\ of\ times < t_i}{N}$$

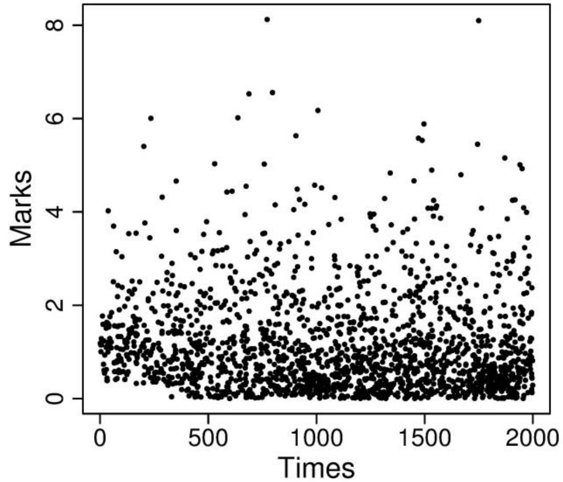
$$m_i \to \quad s_i = \frac{\sum_k \mathbf{1}(m_k < m_i)}{N} = \frac{\#\ of\ marks < m_i}{N}$$

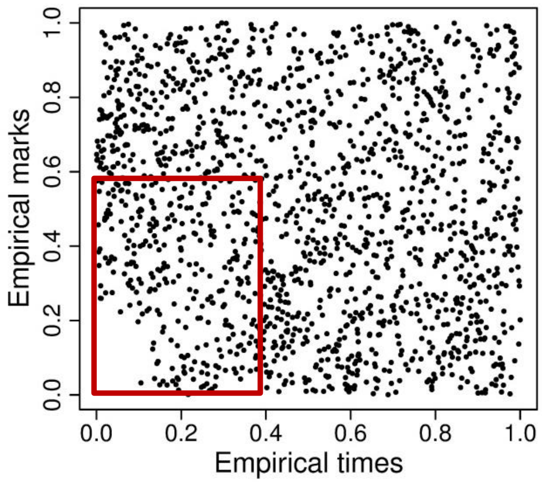
Original process with missing → *Non-homogeneous process*



Key points for Replenishing

Non-complete dataset

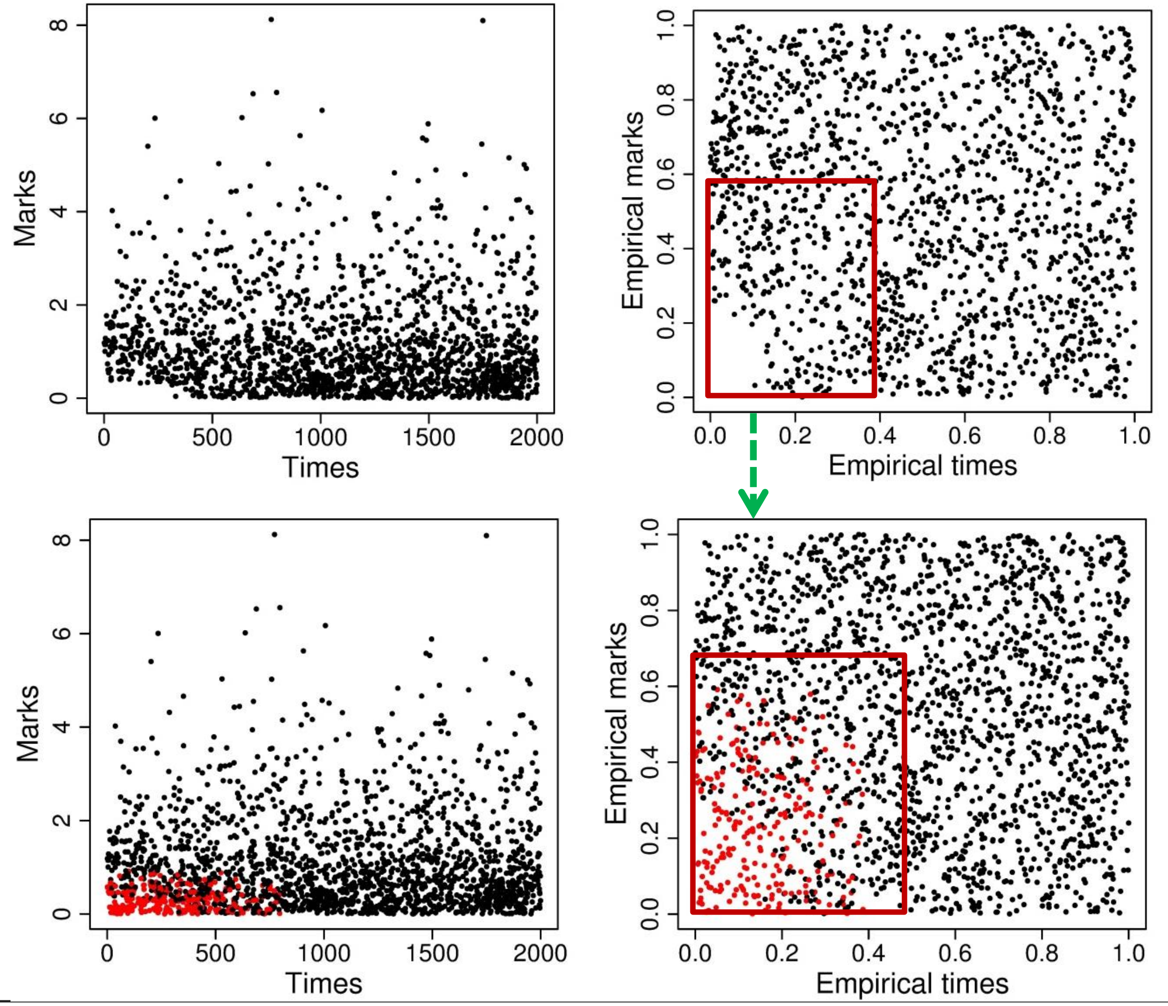




Key points for Replenishing

Non-complete dataset

Complete dataset



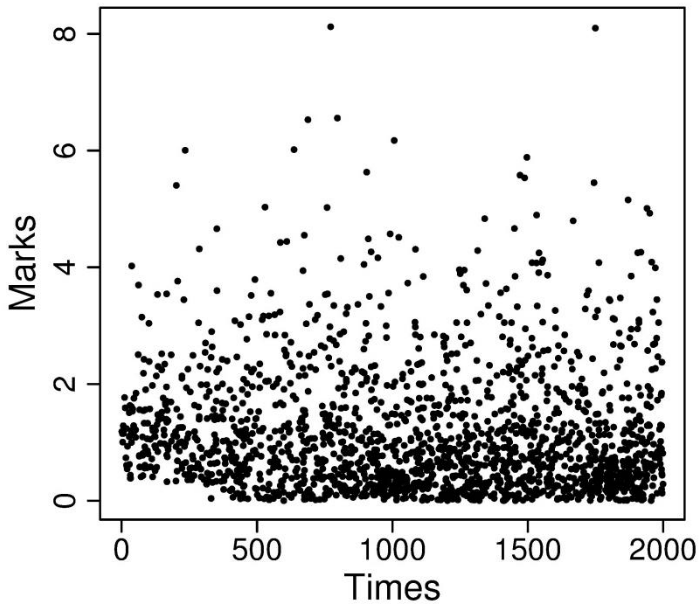
Restore *missing area* without knowing missing data (red dots)?

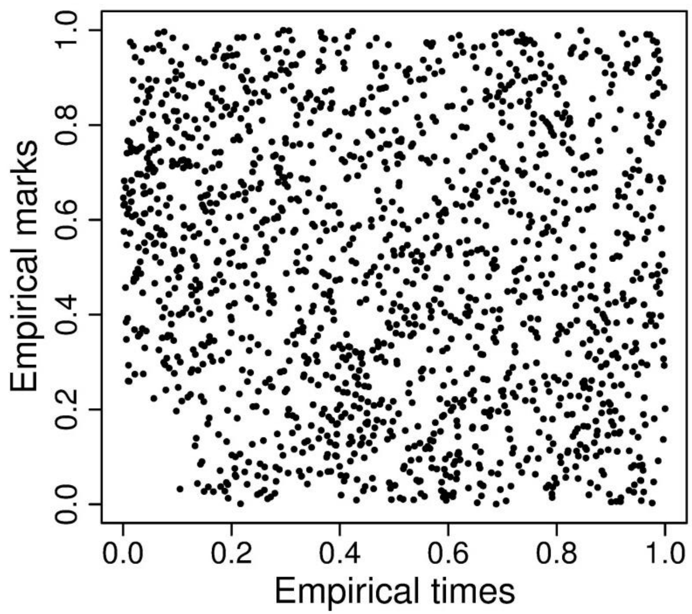
Replenishing algorithm

Step 1. Transform the process using the biscale empirical transformation

$$t_i \rightarrow \quad \tau_i = \frac{i}{N} = \frac{\#\ of\ times < t_i}{N}$$

$$m_i \rightarrow \quad s_i = \frac{\sum_k \mathbf{1}(m_k < m_i)}{N} = \frac{\#\ of\ marks < m_i}{N}$$



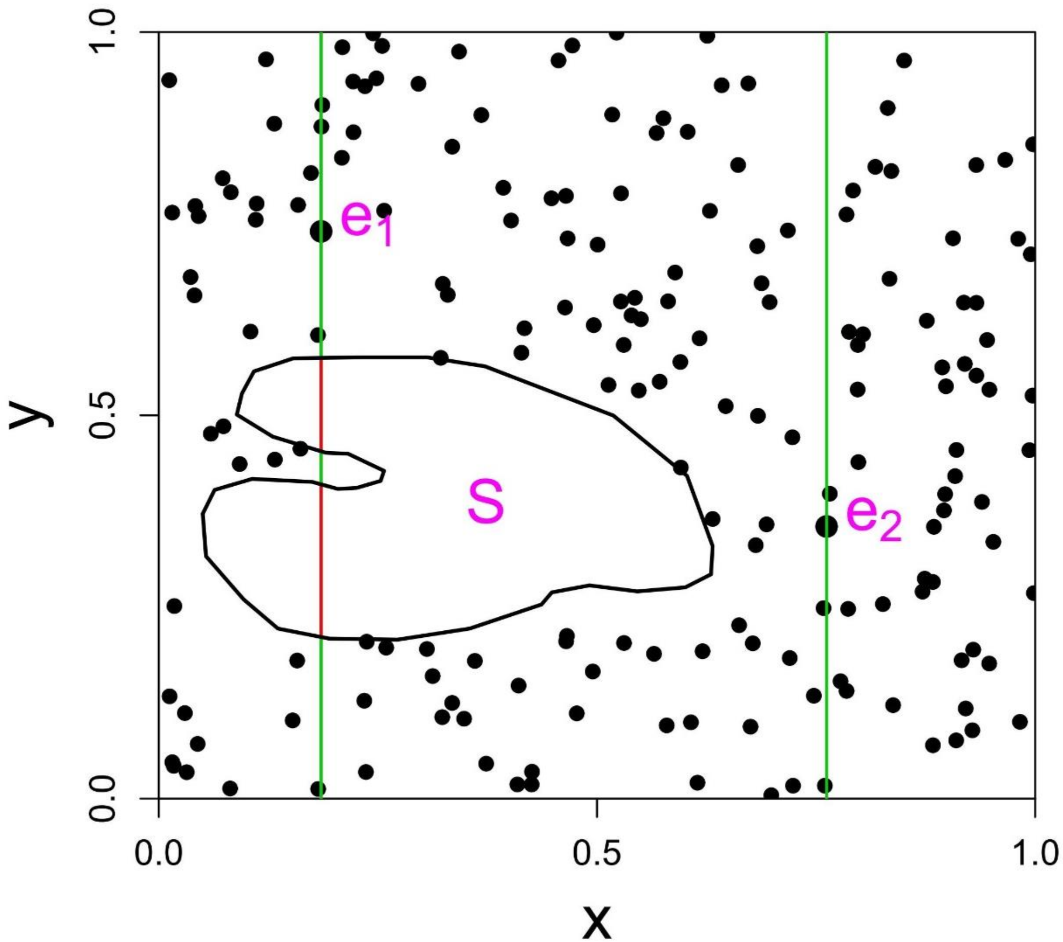


Heuristic illustration

Heuristic illustration for estimating empirical probability distribution function when missing happens in homogeneous process

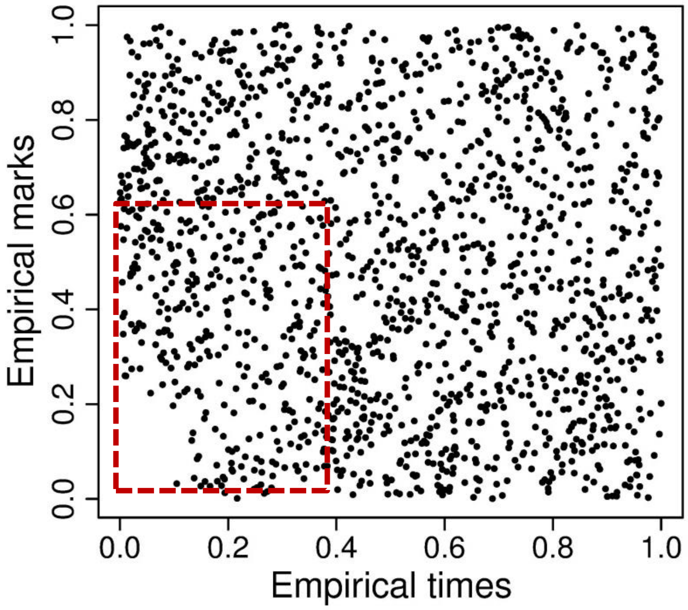
$$\widehat{F_X}(x) = \frac{\sum_i w(e_i) I(x_i < x)}{\sum_i w(e_i)}$$

$$w(e_i) = \frac{1}{1 - \int_0^1 I\big((x_i, y) \in S\big)\, dy}$$



Replenishing algorithm

Step 2. Specify area *S* that contains the missing data



The missing area *S* satisfies

$$\int_M \mathbf{1}\left((t,m) \notin S\right) dF_2(m) > 0$$

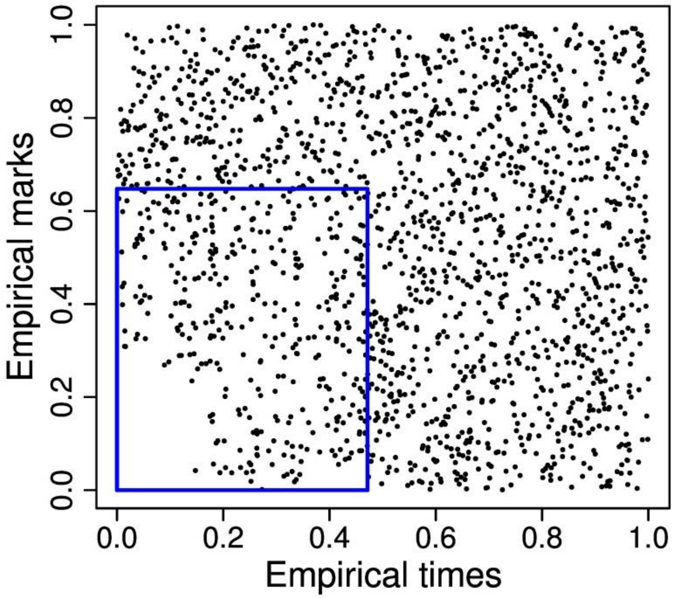
for all $t \in [0,T]$ and

$$\int_0^T \mathbf{1}\left((t,m) \notin S\right) \mu_g(t)\, dt > 0$$

for all $m \in M$.

Replenishing algorithm

Step 3. Calculate the missing area in the biscale transformation domain based on complete data



$$F_1^*(t) = \frac{\sum_{j=1}^{n} w_1(t_j, m_j, S)\mathbf{1}(t_j < t)}{\sum_{j=1}^{n} w_1(t_j, m_j, S)}$$

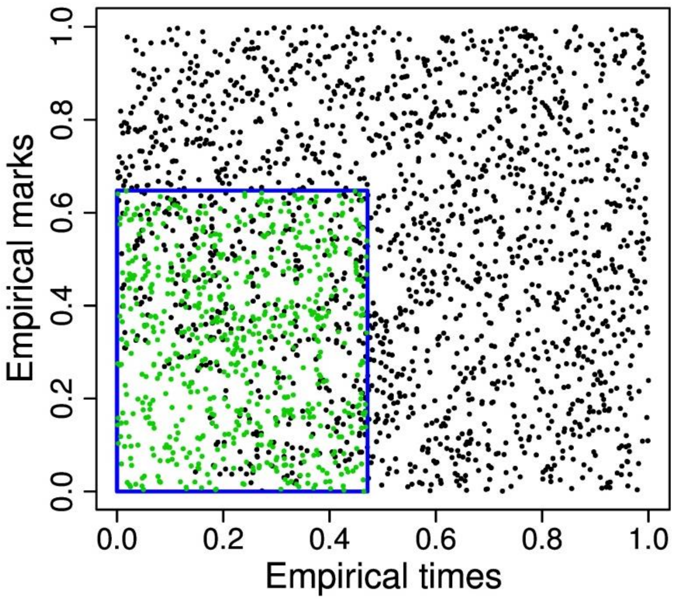
$$F_2^*(m) = \frac{\sum_{j=1}^{n} w_2(t_j, m_j, S)\mathbf{1}(m_j < m)}{\sum_{j=1}^{n} w_2(t_j, m_j, S)}$$

$$w_1(t, m, S) = \frac{\mathbf{1}((t, m) \notin S)}{\int_M \mathbf{1}((t, s) \notin S) dF_2^*(s)}$$

$$w_2(t, m, S) = \frac{\mathbf{1}((t, m) \notin S)}{\int_M \mathbf{1}((\tau, m) \notin S) dF_1^*(\tau)}$$

Replenishing algorithm

Step 4. Generate data point in the missing area



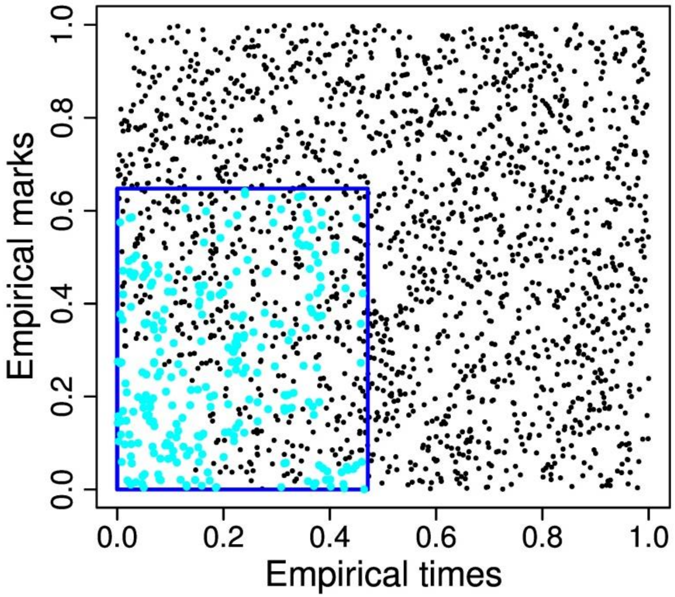
Generate events uniformly distributed in the missing region $S*$ (image of S)

#events ~ $NB(k, 1-|S^*|)$

k: # of observed events outside of S^*

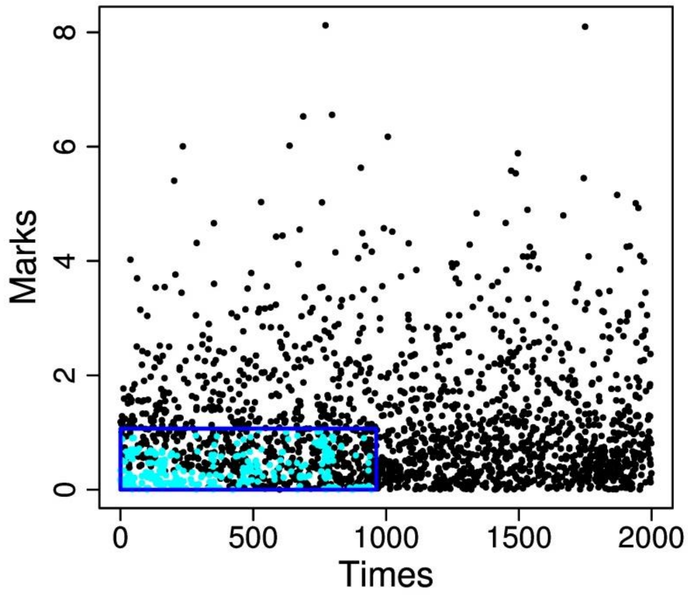
Replenishing algorithm

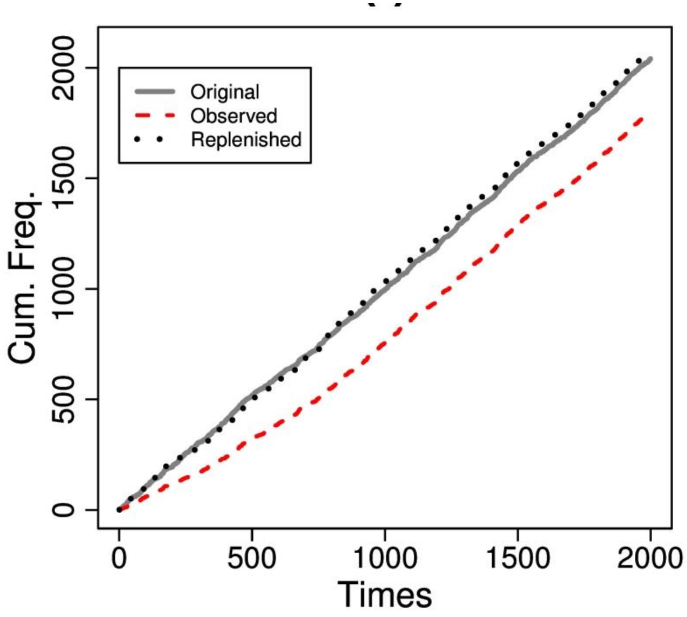
Step 4. Remove sequentially a simulated data point for each existing point in the missing area



Replenishing algorithm

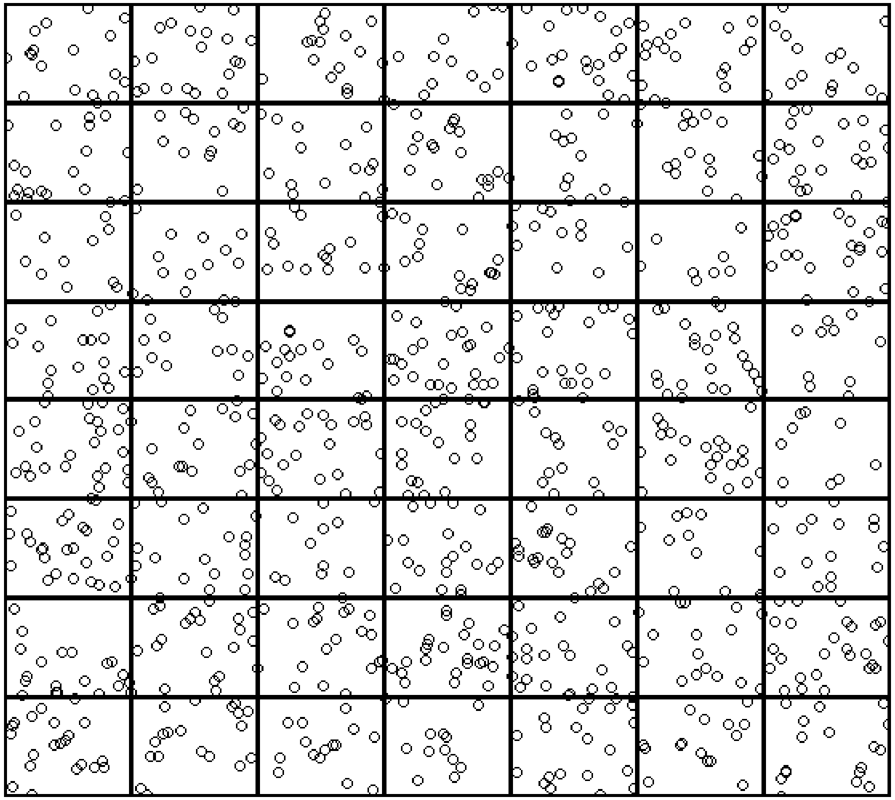
Step 5. Transform back all the events into the original domain





How to test existence of missing: Testing method

L_2 columns



L_1 rows

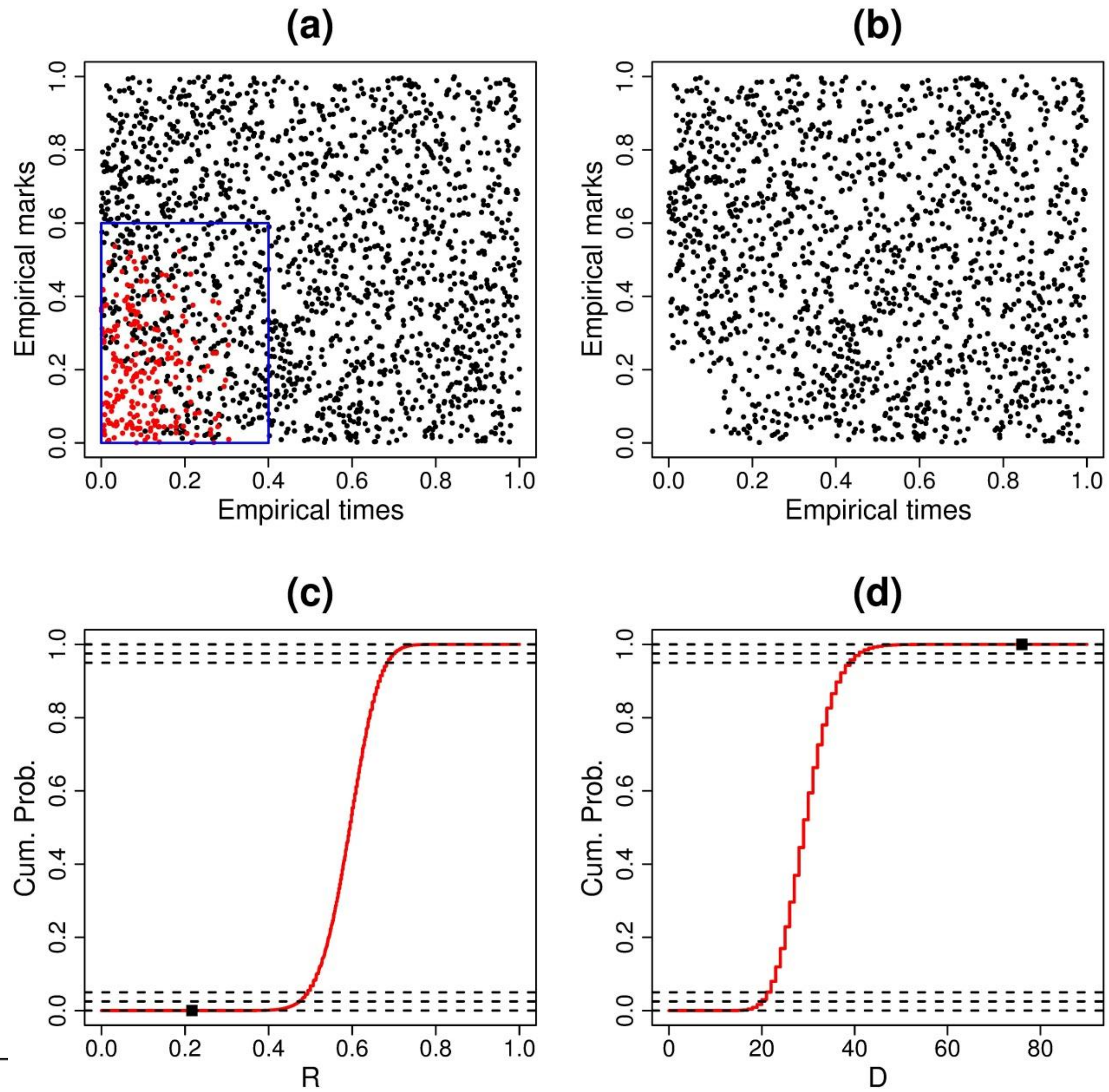
$$L = L_1 \times L_2$$

$$R = \frac{\min\{C_1, C_2, \cdots, C_L\}}{\max\{C_1, C_2, \cdots, C_L\}}$$

$$D = \max\{C_1, C_2, \cdots, C_L\} - \min\{C_1, C_2, \cdots, C_L\}$$

C_i: #events in cell i

Testing method

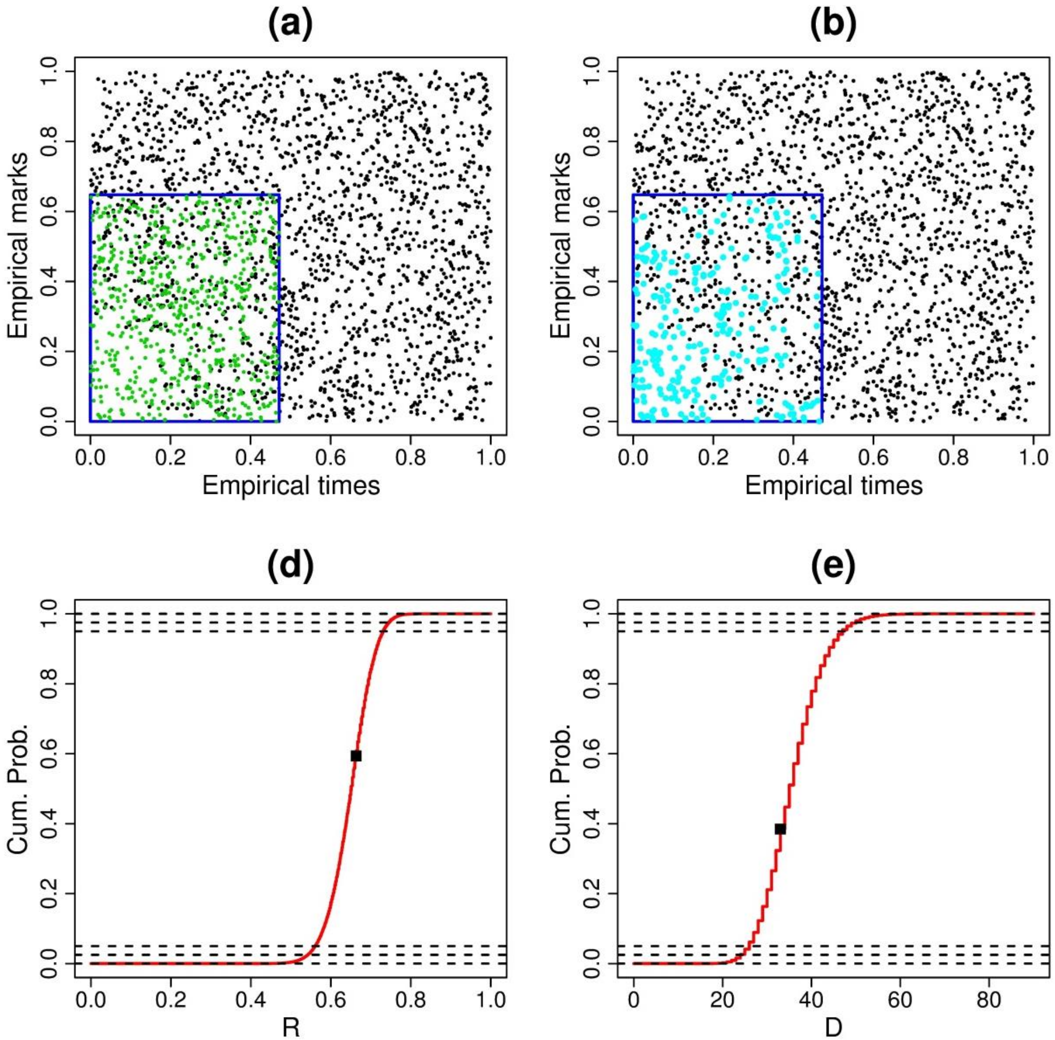


Observed data

$$L_1 = L_2 = 5$$
$$L = 25$$

Distributions of *R* and *D* when the same number of data points are completely observed

Testing method



Replenished data

$$L_1 = L_2 = 5$$
$$L = 25$$

Distributions of *R* and *D* when the same number of data points are completely observed

More Simulations

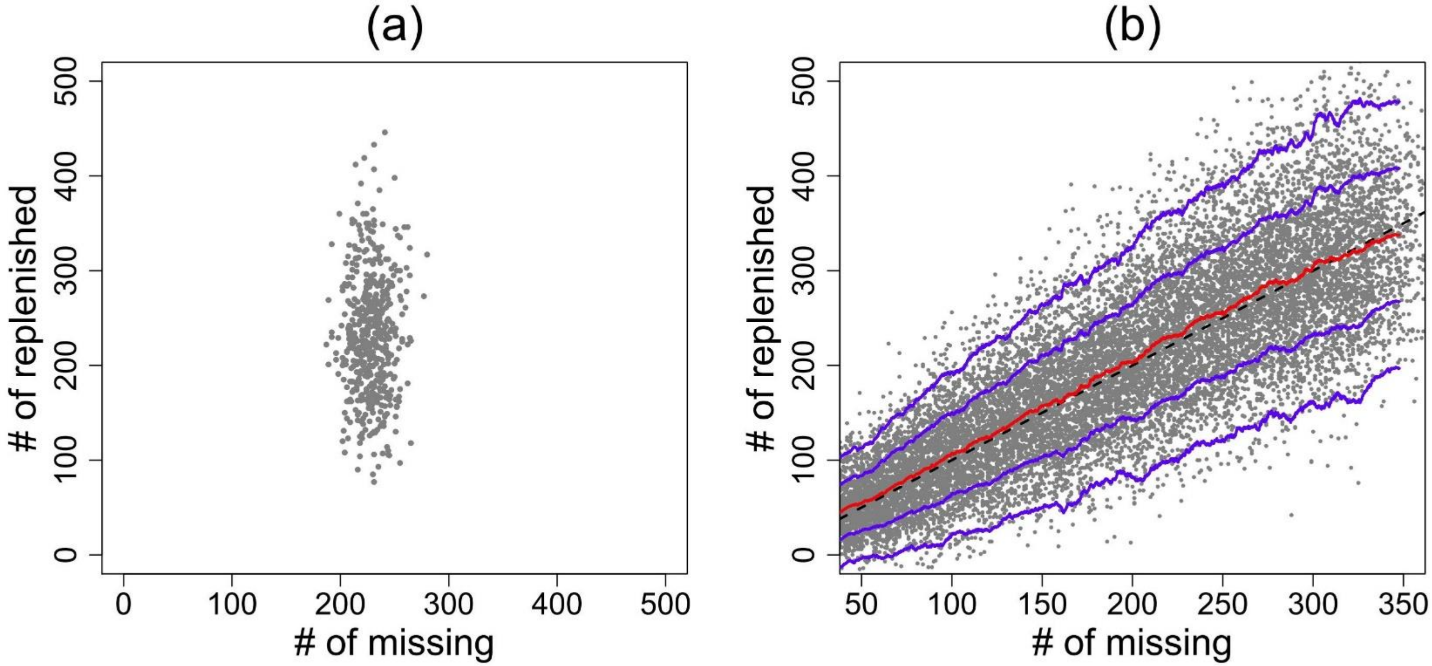
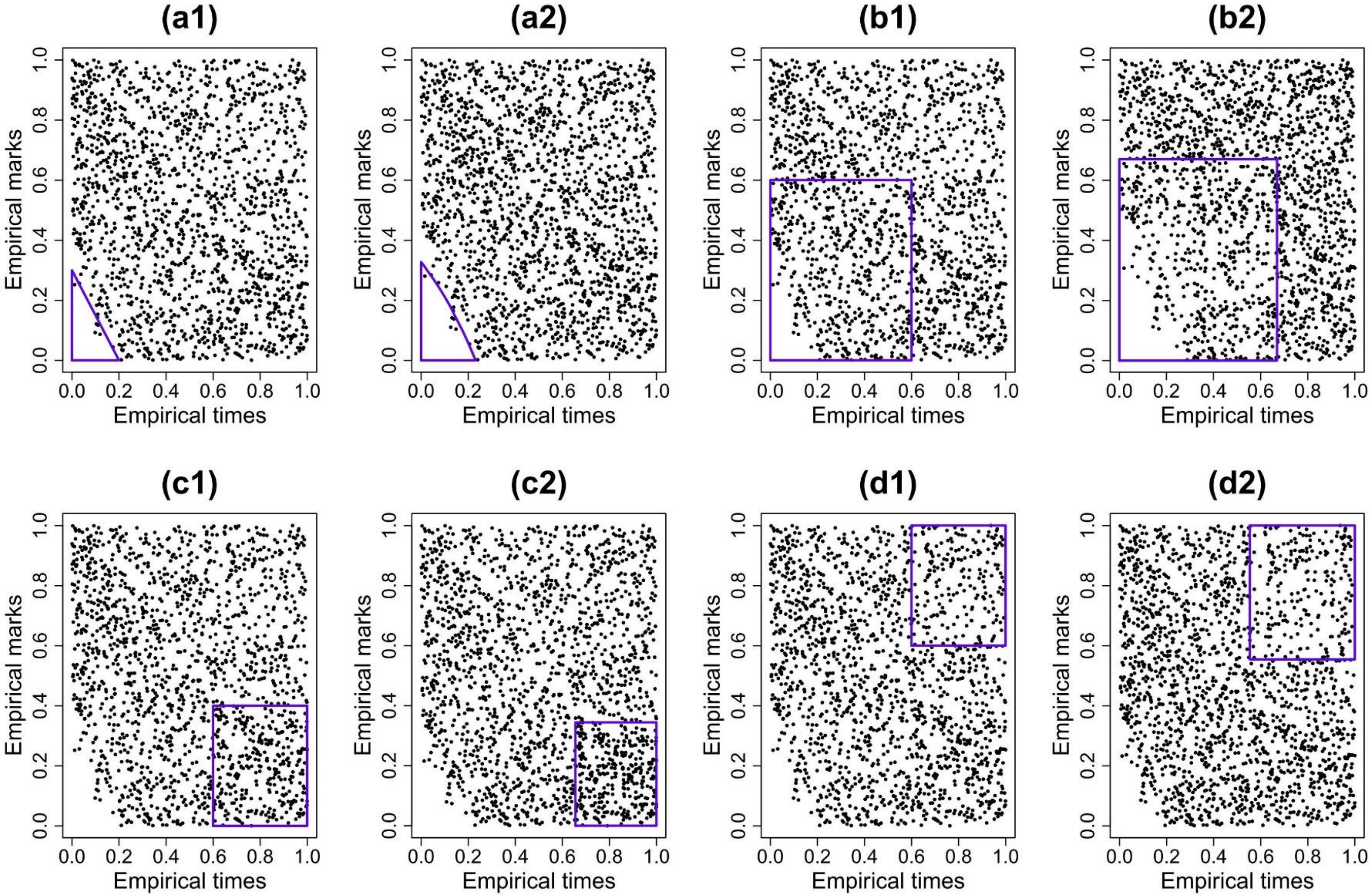
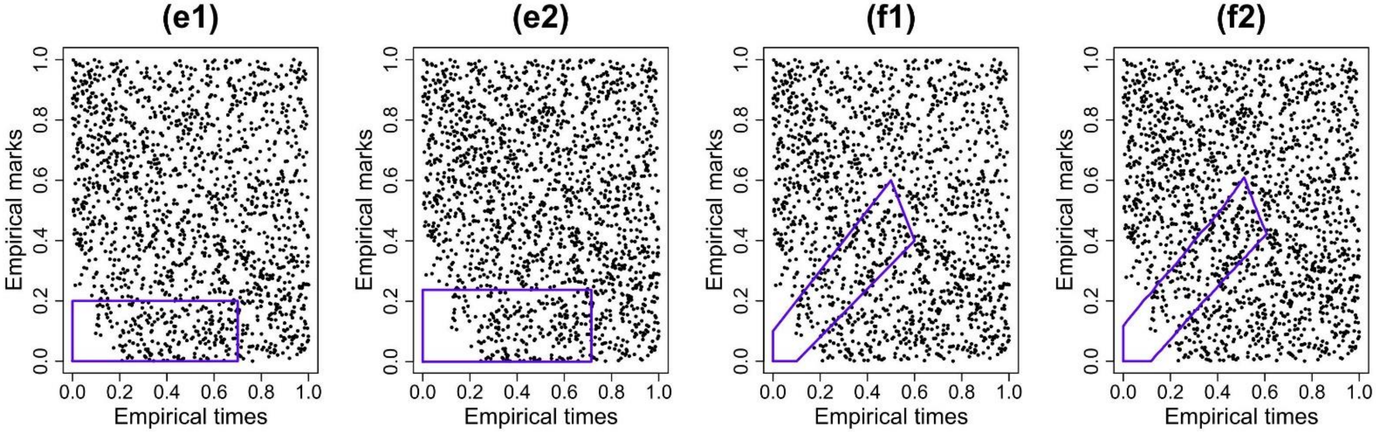


Figure 5: Comparison between the number of true missing events and the number of replenished events. (a) $\lambda = 2,000$ fixed. (b) λ is drawn from a uniform distribution between 100 and 3000. The dashed lines represent the case where the numbers of missing events and replenished events are equal. The blue and red curves represent the running mean and the corresponding single and double standard deviation bands.

If wrong selection of missing area (1)

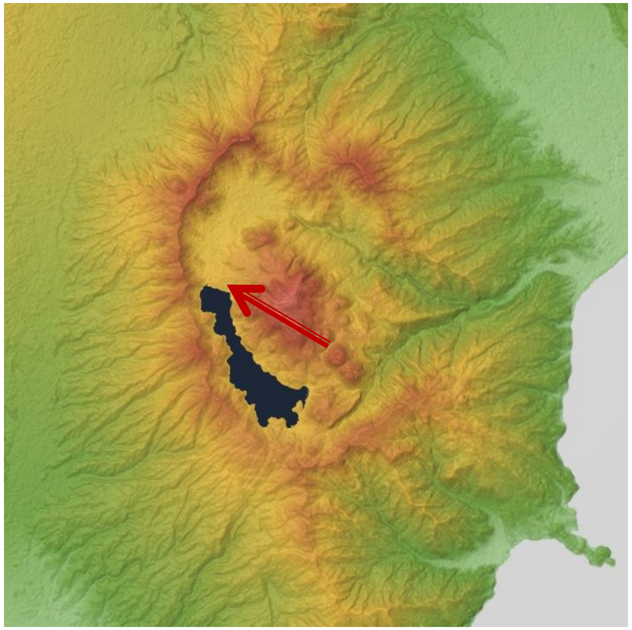


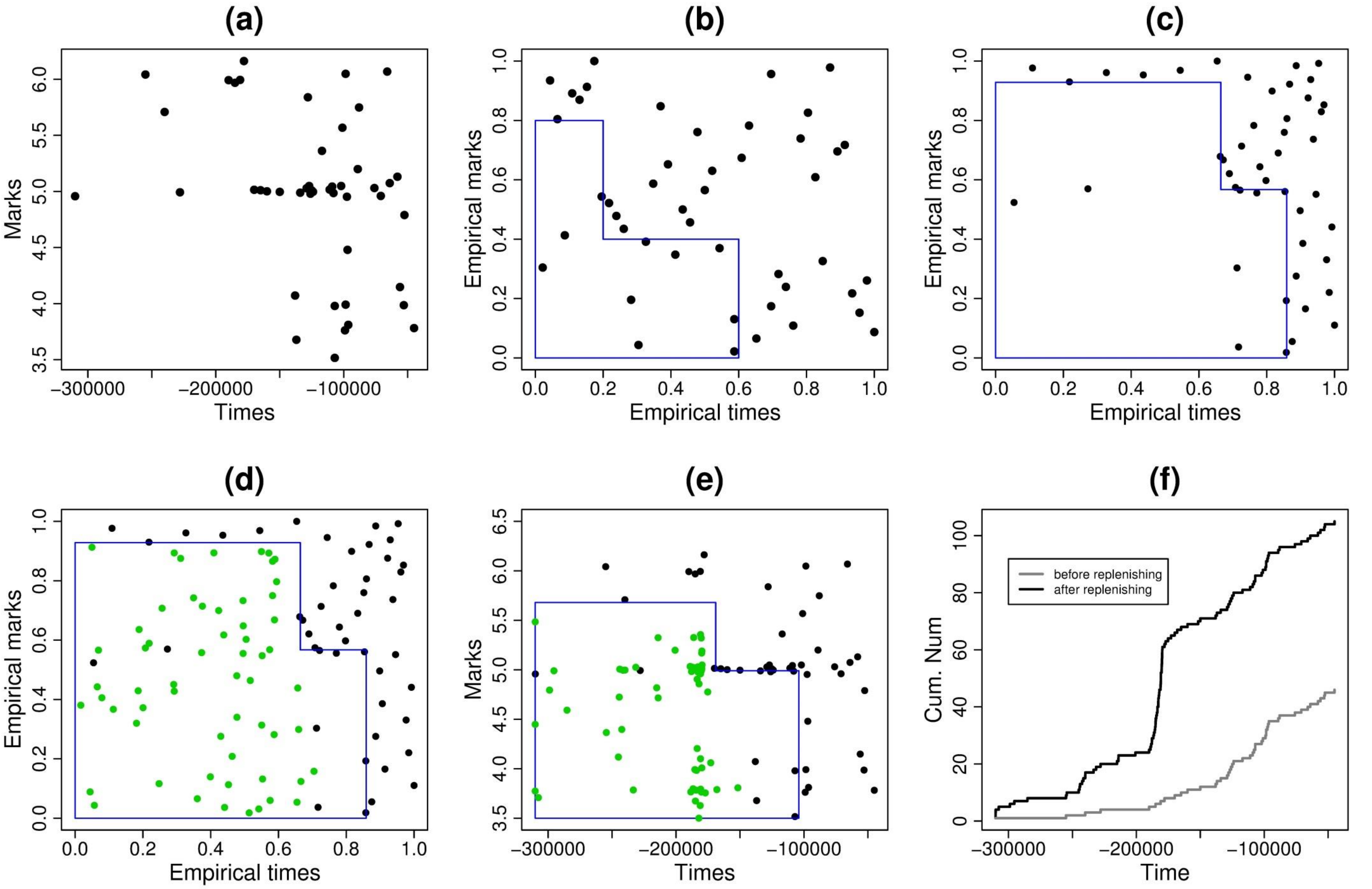
If wrong selection of missing area (1)



Application 1: Volcanic eruption record

- Data: eruptions at Hakone volcano
- Data source:
  - Smithsonian's Global Volcanism Program database
  - Large Magnitude Explosive Volcanic Eruptions database (LaMEVE database)
  - additional Japanese databases





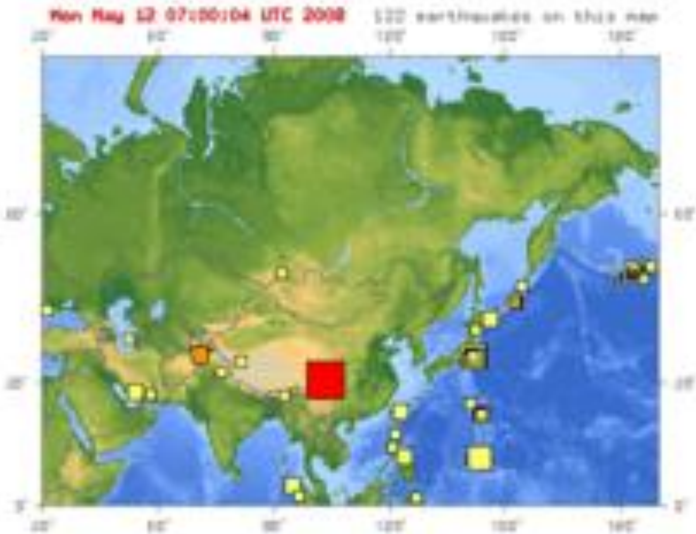
(a)
(b)
(c)
(d)
(e)
(f)
Marks
Times
Empirical marks
Empirical times
Cum. Num
Time
before replenishing
after replenishing

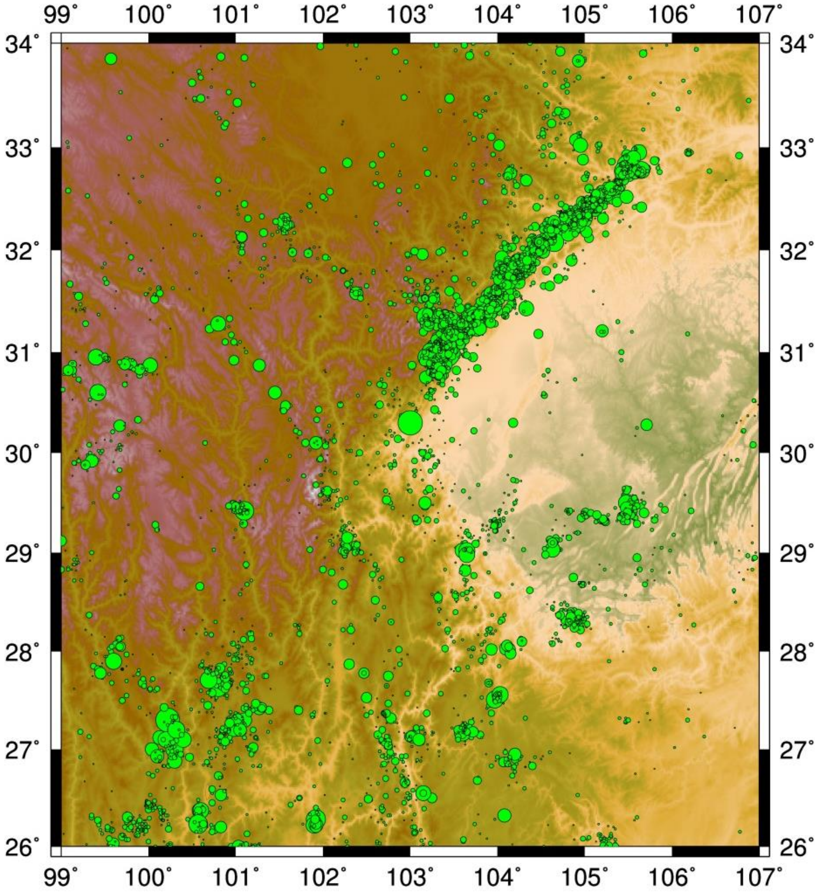
Application to the Wenchuan aftershock sequence data

Wenchuan EQ: Mw7.9 (Ms8. 2) 2008/5/12

Data selection:

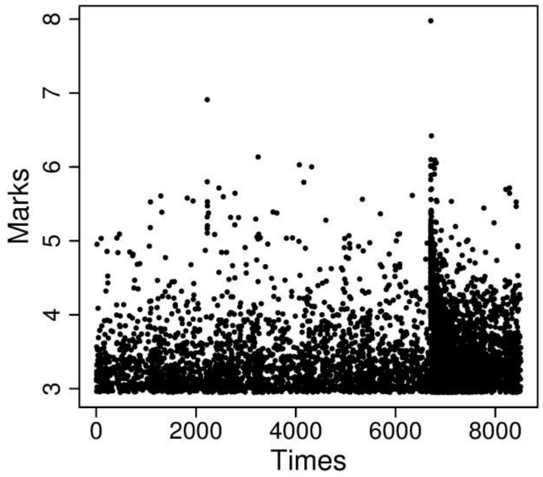
Time: 1990/1/1~2013/4/20
Magnitude: 3.0+



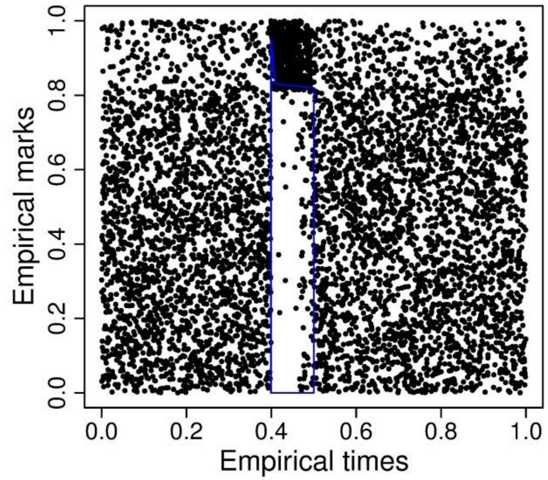


Application to the Wenchuan aftershock sequence data

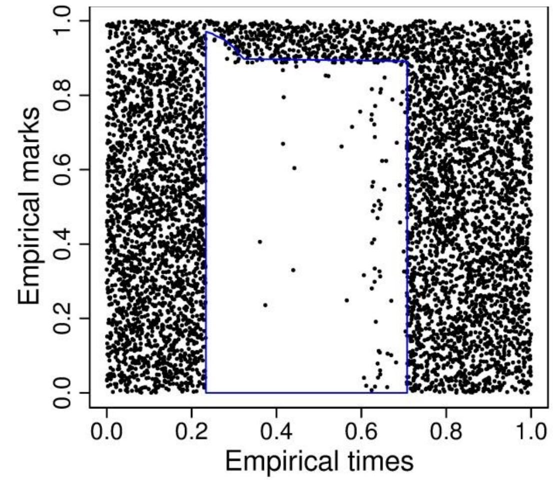
Observed data



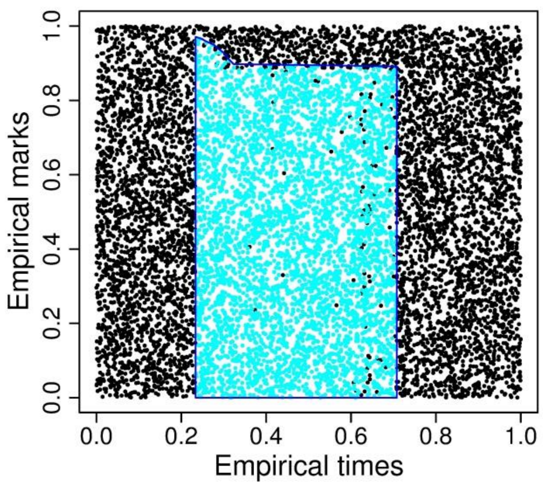
Biscale transformed



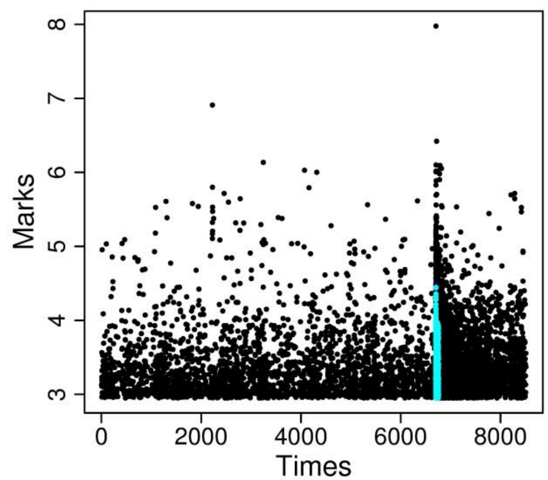
Estimate transformation under complete data



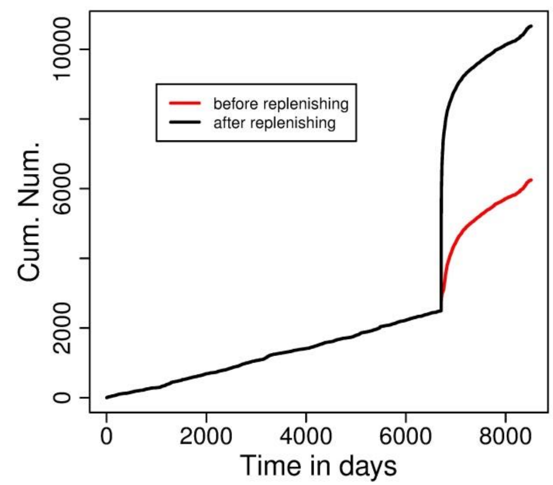
Transformed replenished data



Original replenished data



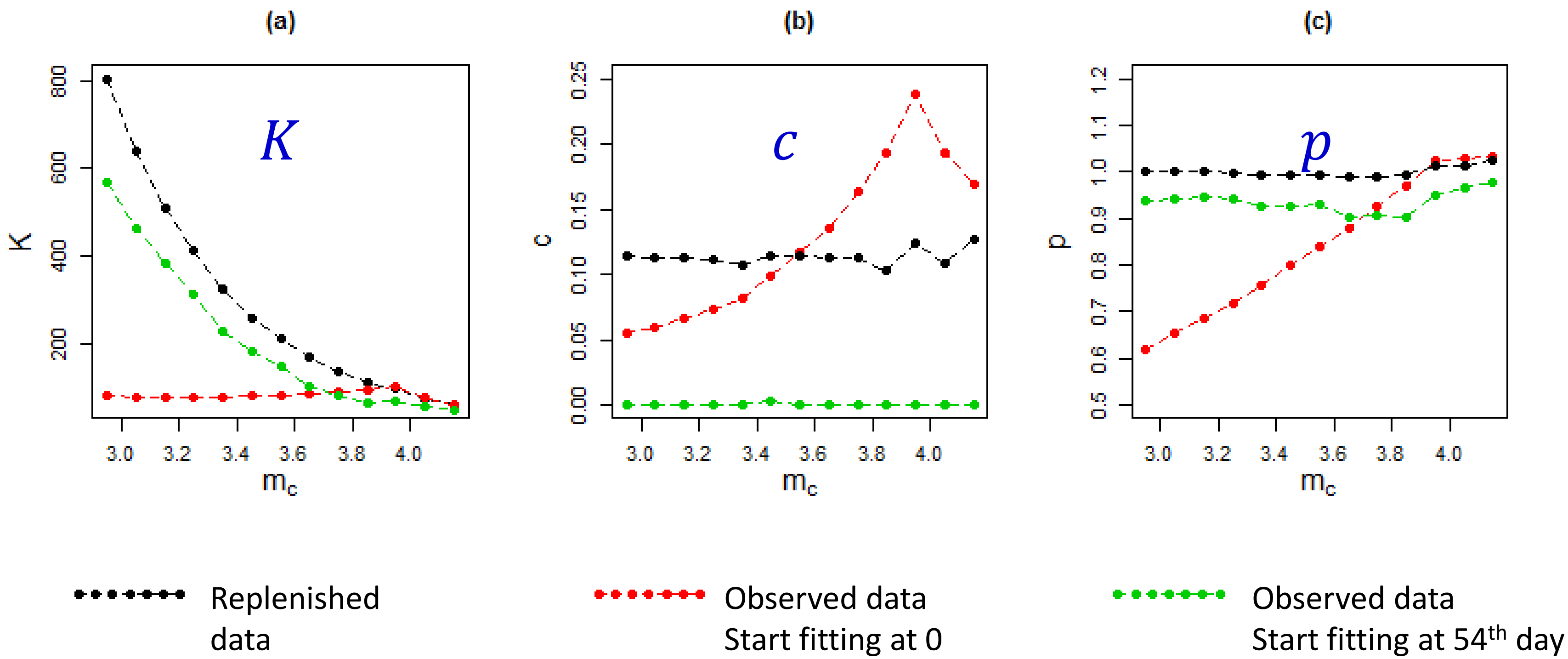
Cumulative freq. vs time



Influence of short-term aftershock missing on estimating the Omori formula

Omori-Utsu formula:

$$\lambda(t) = \frac{K(p-1)}{c}\left(1+\frac{t}{c}\right)^{-p}, \qquad t > 0$$

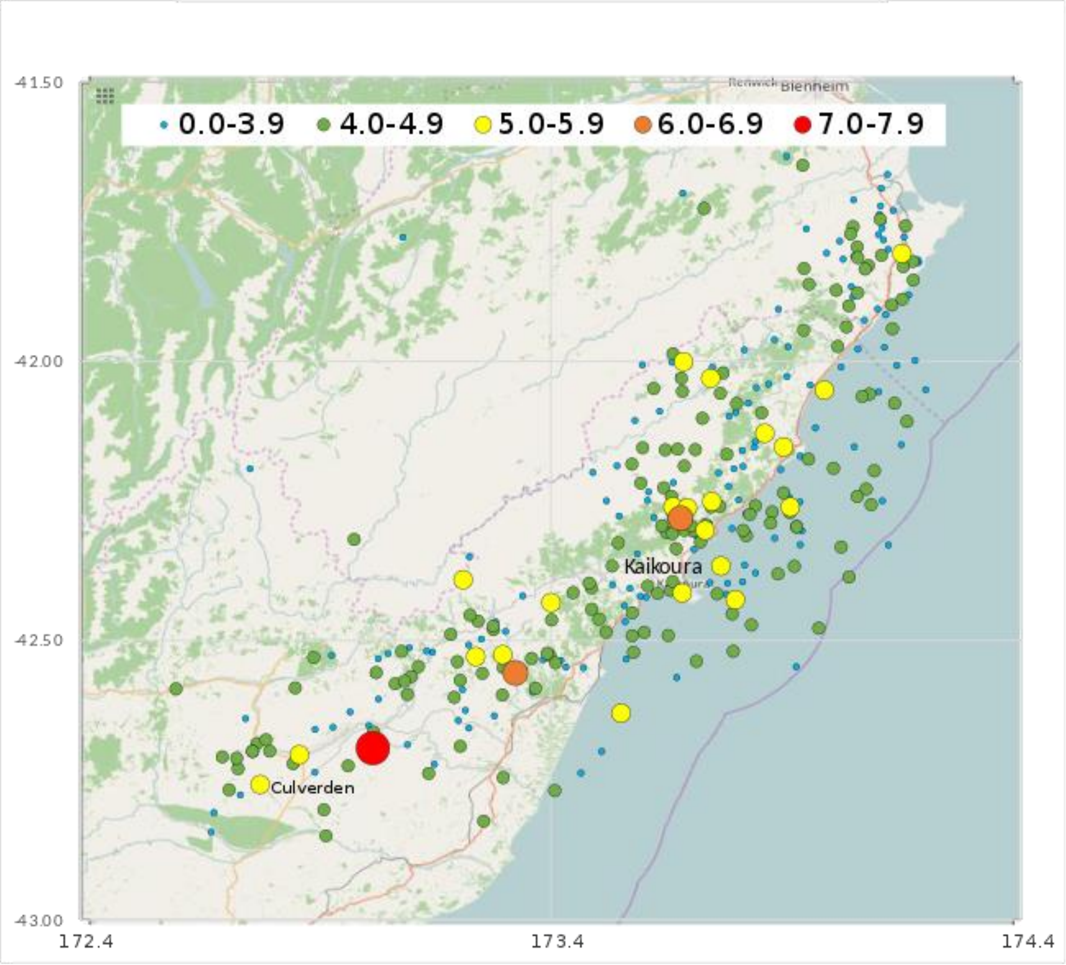


Application 3: the Kaikoura aftershock sequence data

Kaikoura EQ: Mw7.8 2016/11/14

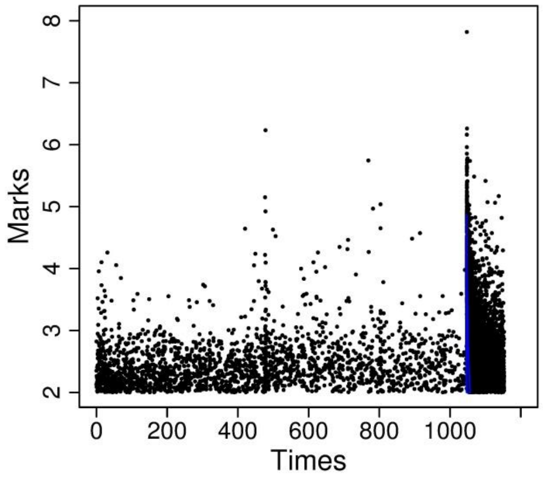
Data selection:

Time: 2014/1/1~2017/2/17
Magnitude: 2.0+

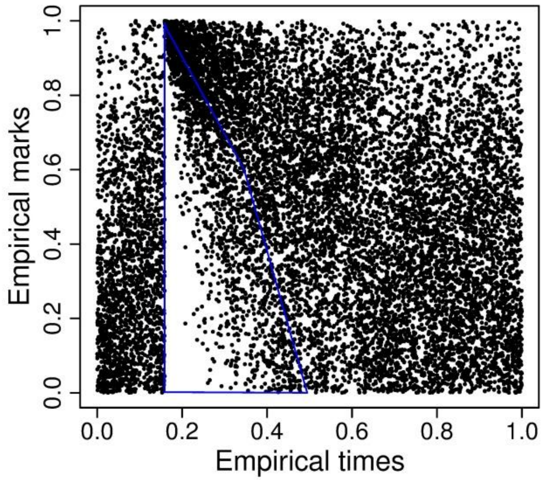


Application 3: the Kaikoura aftershock sequence data

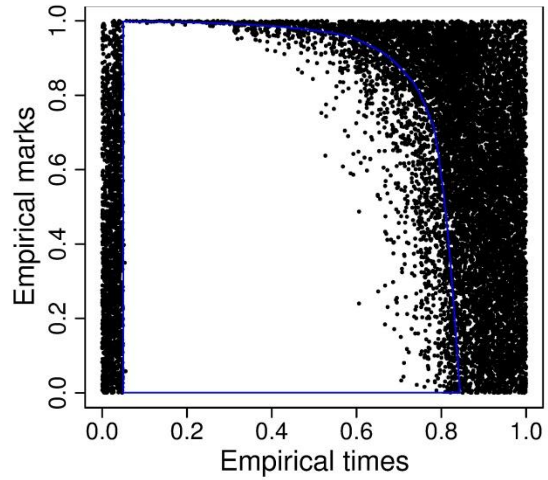
Observed data



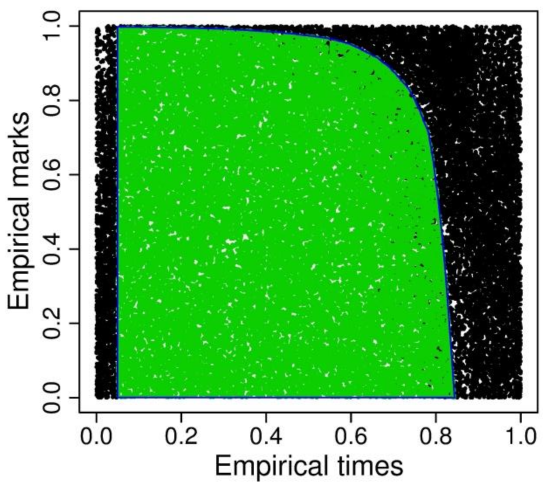
Biscale transformed



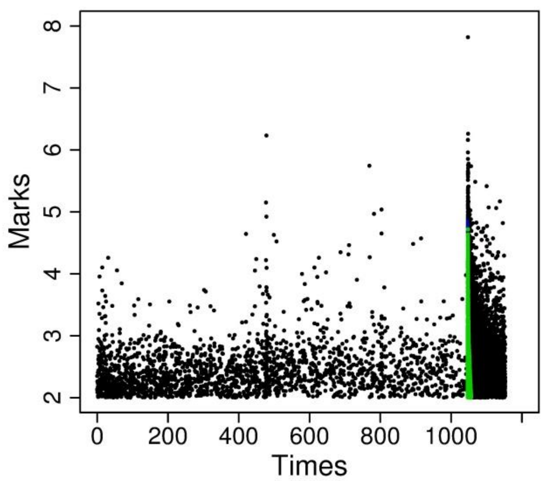
Estimate transformation under complete data



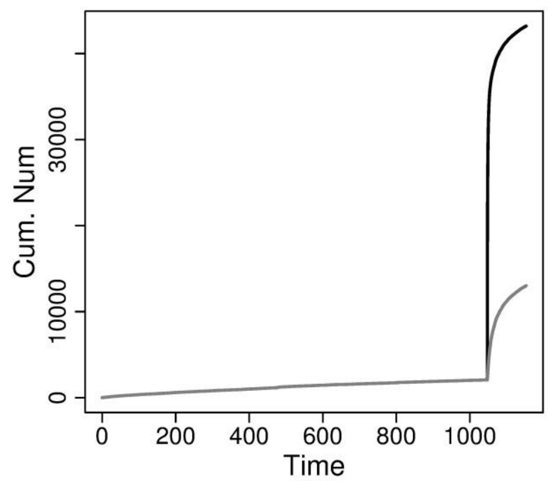
Transformed replenished data



Original replenished data



Cumulative freq. vs time



On the biased estimate of earthquake clustering parameters caused by short-term aftershock missing

Errors in estimated parameters propagate in forecasting.

The ETAS model

- **Conditional intensity** (Ogata, 1998)

$$\lambda(t,x,y) = E[N(dtdxdy)|\mathrm{H}_t] / |dtdxdy|$$
$$= \nu\mu(x,y) + \sum_{i:t_i<t} \kappa(m_i) g(t-t_i) f(x-x_i, y-y_i, m_i)$$

1. Direct productivity: $\kappa(m) = Ae^{\alpha(m-m_C)}, \quad m \geq m_C$
2. Time p.d.f (Omori-Utsu): $g(t) = (p-1)(1+t/c)^{-p}/c, \quad t>0$
3. Location p.d.f: $f(x,y|m) = \dfrac{q-1}{\pi De^{\gamma(m-m_c)}}\left(1+\dfrac{x^2+y^2}{De^{\gamma(m-m_c)}}\right)^{-q}$

Temporal version (Ogata 1988, JASA)

$$\lambda(t) = \mu + K \sum_{i:t_i<t} \frac{e^{\alpha(m_i-m_c)}}{(t-t_i+c)^p}$$

$$A = \frac{Kc^{1-p}}{p-1} = \int_0^\infty \frac{K}{(t+c)^p} dt$$

expected # of direct offspring from m_c

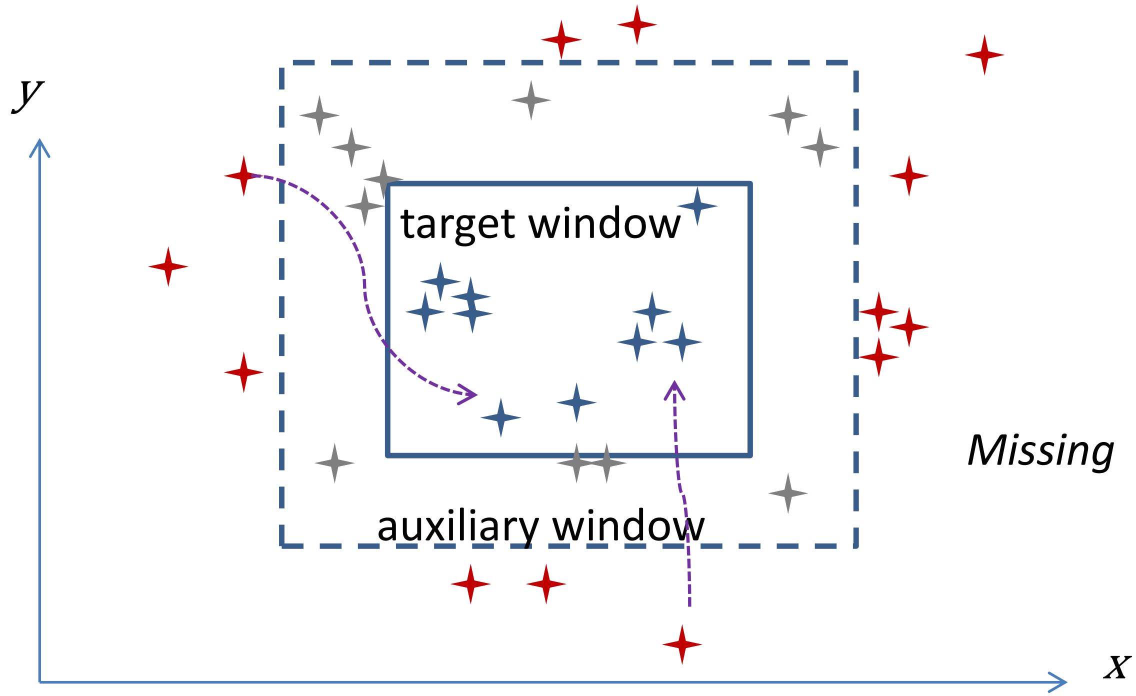
What influence the estimate of the ETAS parameters

What influence the estimate of the ETAS parameters

Missing links!

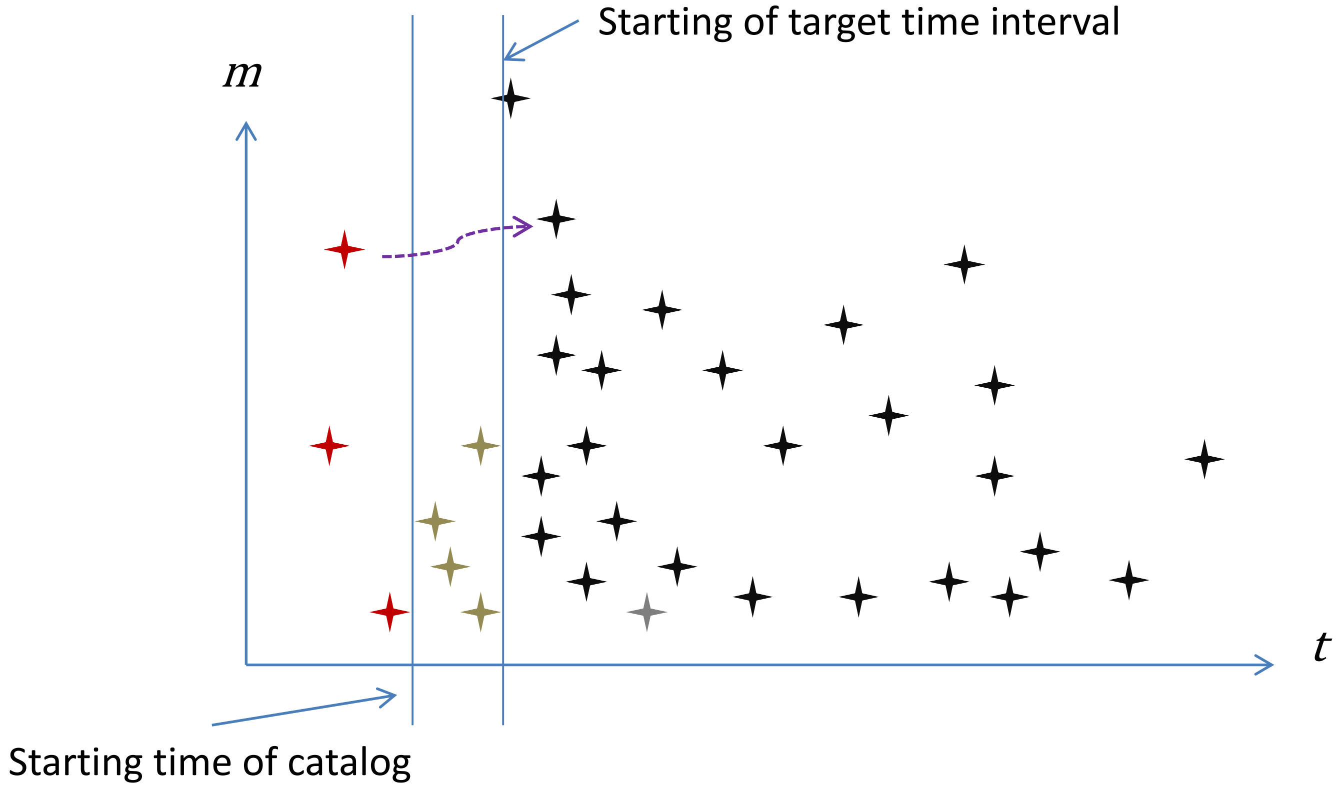
What influence the estimate of the ETAS parameters: *Missing links!*

1. Missing links in space



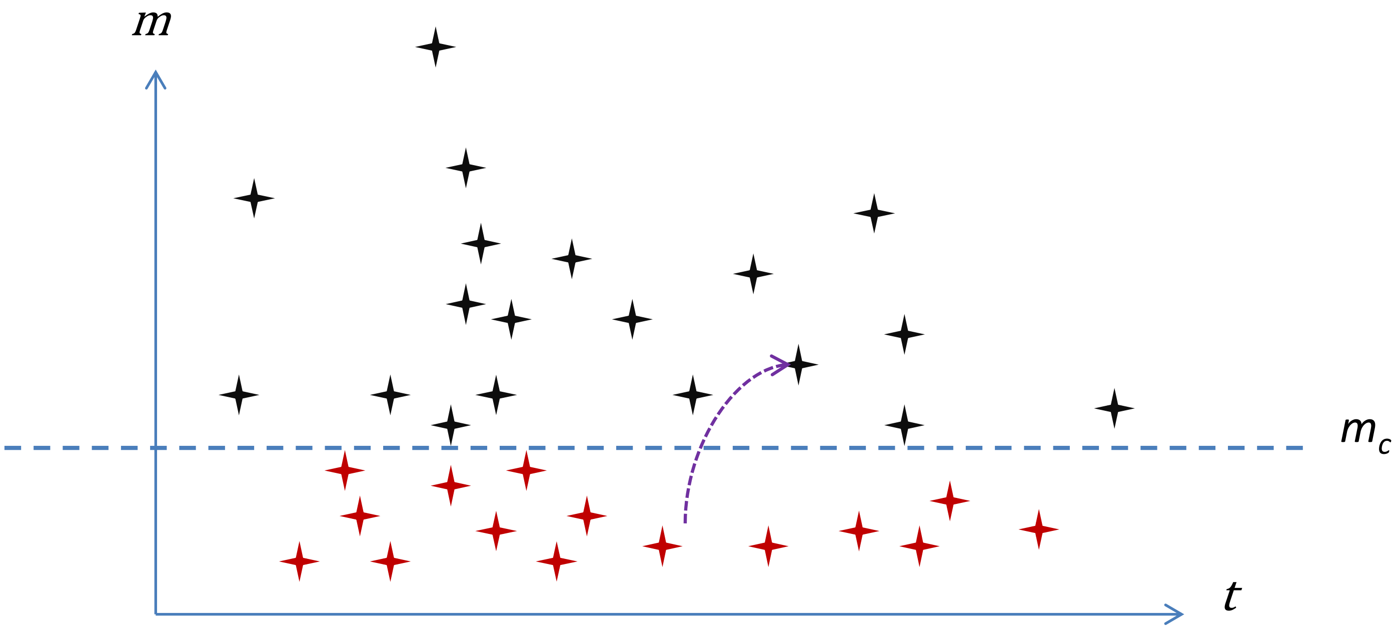
What influence the estimate of the ETAS parameters: *Missing links!*

2. Missing links in time



What influence the estimate of the ETAS parameters: *Missing links!*

3. *Missing links in magnitude*



The influence of the above missing links has been studied in

- Sornette & Werner (2005) **JGR**, 110, B09303.
- Sornette & Werner (2005) **JGR**, 110, B08304.
- Wang, et al. (2010) **BSSA**, 100 , 1989 – 2001.
- Wang, Jackson & Zhuang. (2013) **GRL**., 37 , L21307.
- etc..

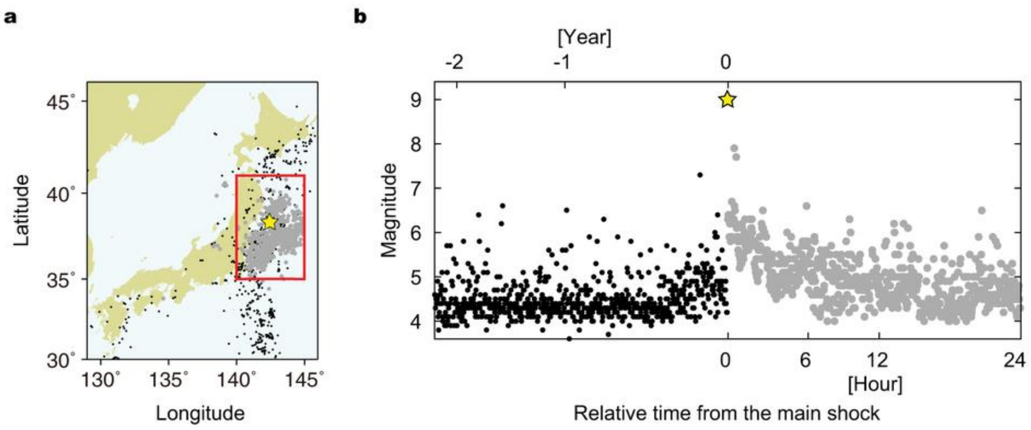
What influence the estimate of the ETAS parameters

4. Missing links caused by short-term missing of aftershocks.

What influence the estimate of the ETAS parameters

4. Missing links caused by short-term missing of aftershocks.

Missing events after the mainshock (from Omi etc, 2013)



Previous studies for fixing the problems of short-term missing aftershocks

Observational approaches

- waveform-based earthquake detection methods (e.g., Enescu et al., 2007, 2009; Peng et al., 2007; Marsan and Enescu, 2012; Hainzl, 2016).
- Energy based description (Sawazaki and Enescu, 2014)

Statistical approaches

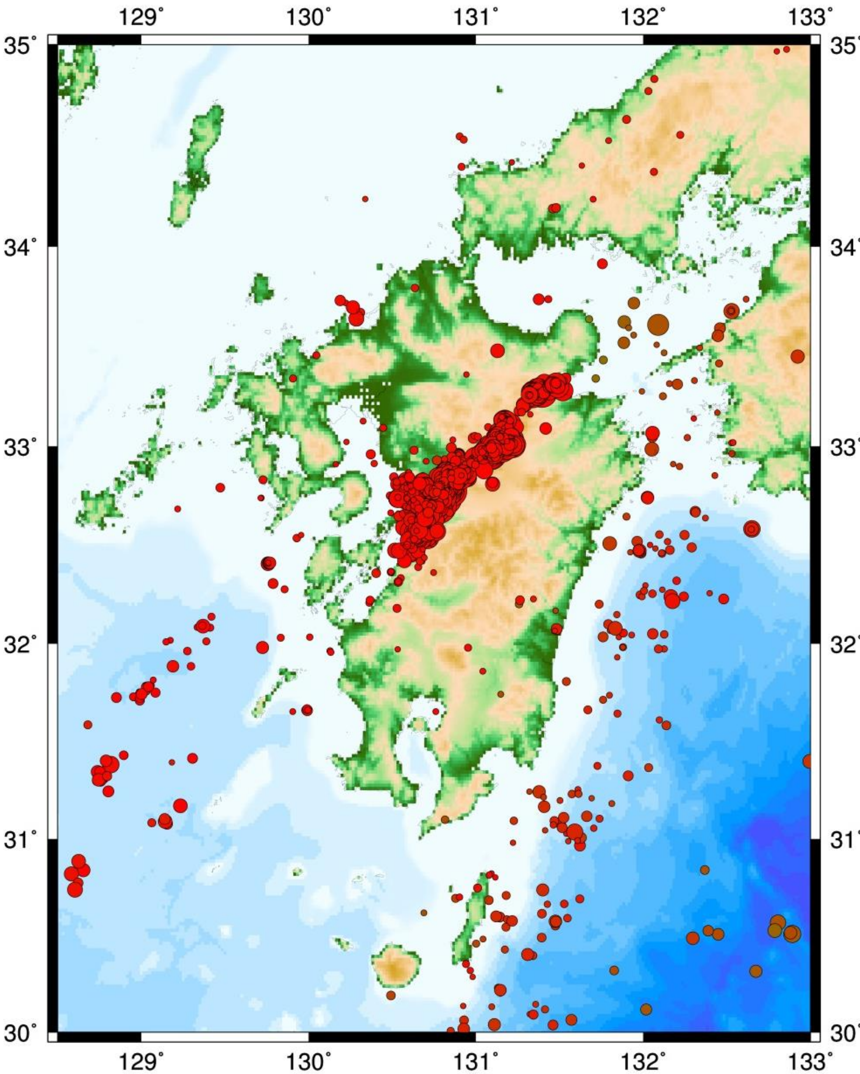
- (Ogata, Omi, Iwata) Bayesian, assuming GR relation for whole range
- (Marsan and Enescu, 2012) Assuming Omori-Utsu formula or ETAS model
- This study: Conditional independence between magnitudes and occurrence times

Application to the Kumamoto aftershock sequence data

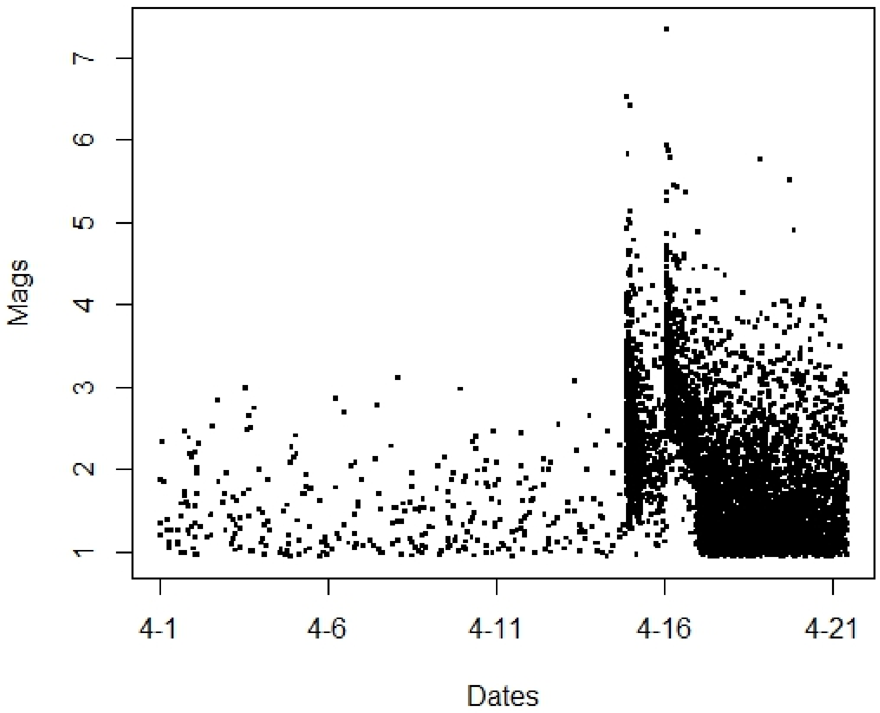
Data Selection

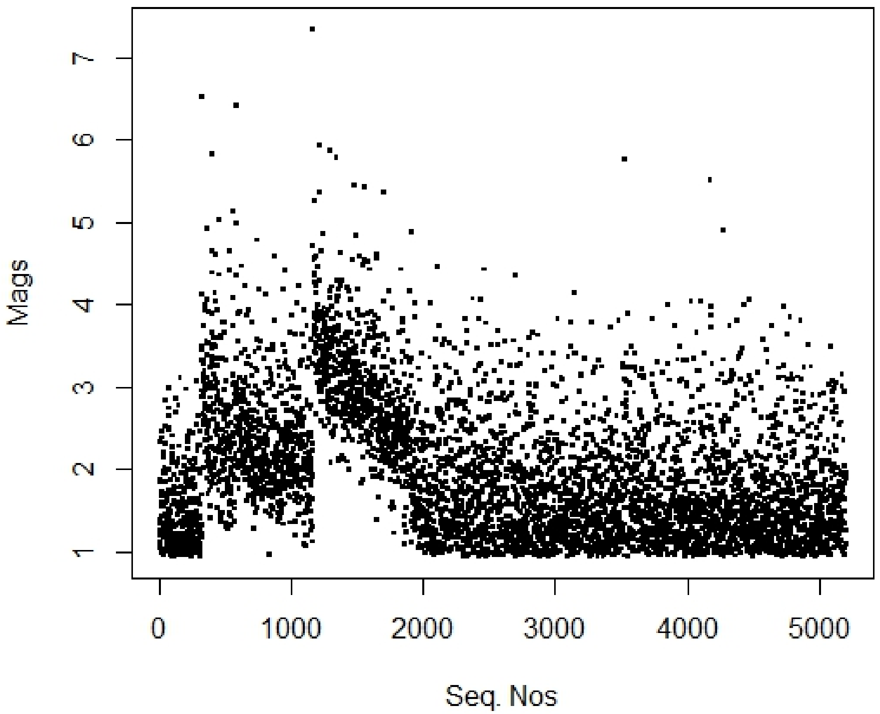
Time: 2016/4/1~2016/4/21
Mag.: 1.0+
Depth: < 100 km
Space: 128° -- 133°
30° -- 35°

2016-04-14 21:26 (130.81 32.74) M6.5
2016-04-15 00:03 (130.78 32.70) M6.4
2016-04-16 01:25 (130.76 32.75) M7.3



Application to the recent Kumamoto aftershock sequence data

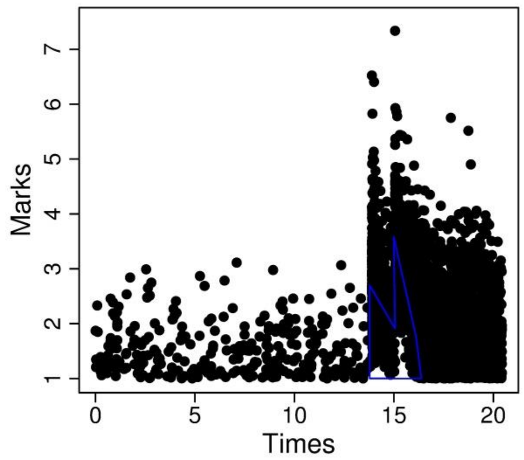




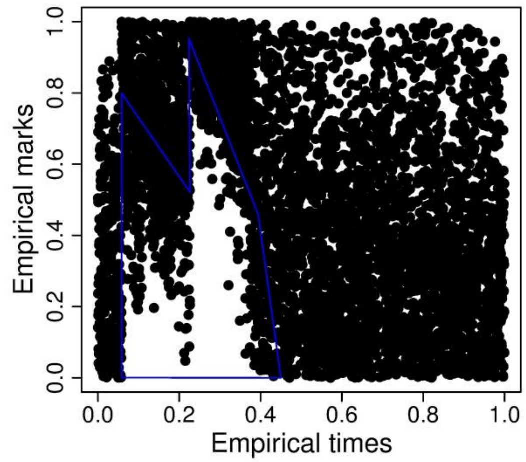
Replenish the missing data

Method (Zhuang et al, 2016)

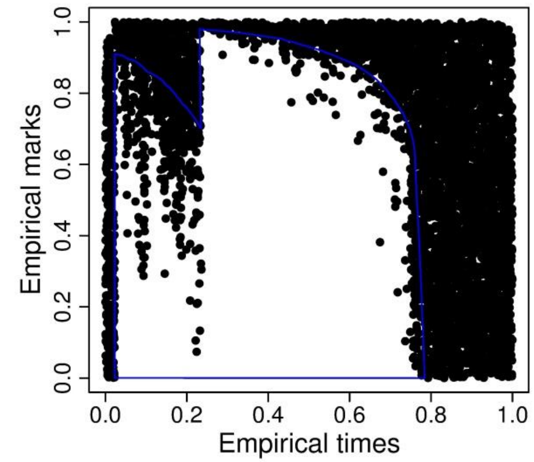
Observed data



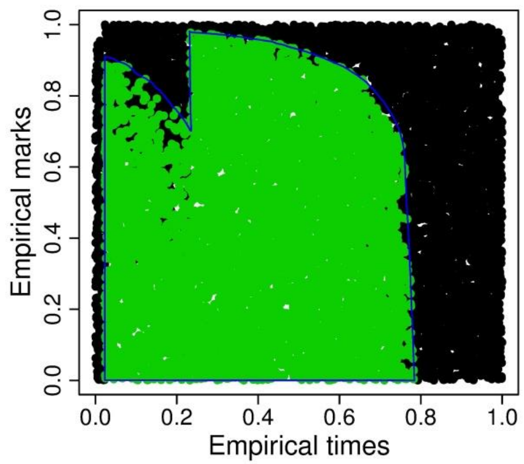
Biscale transformed



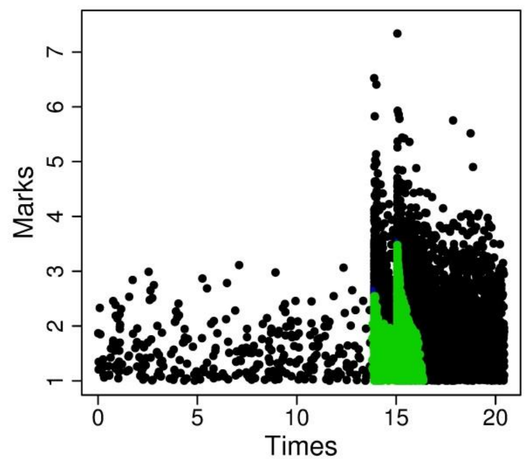
Estimated transformation under complete data



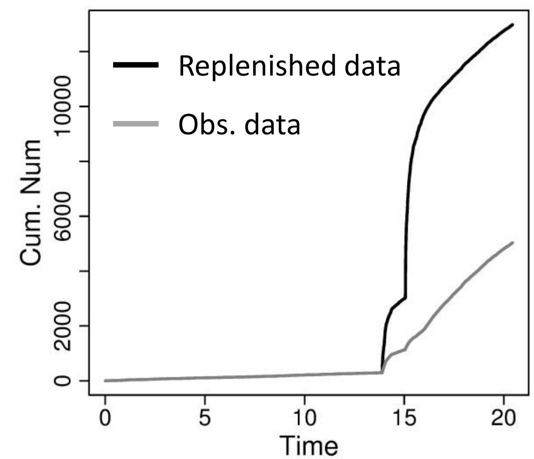
Transformed replenished data



Original replenished data



Cumulative freq. vs time



Consider only the temporal ETAS model

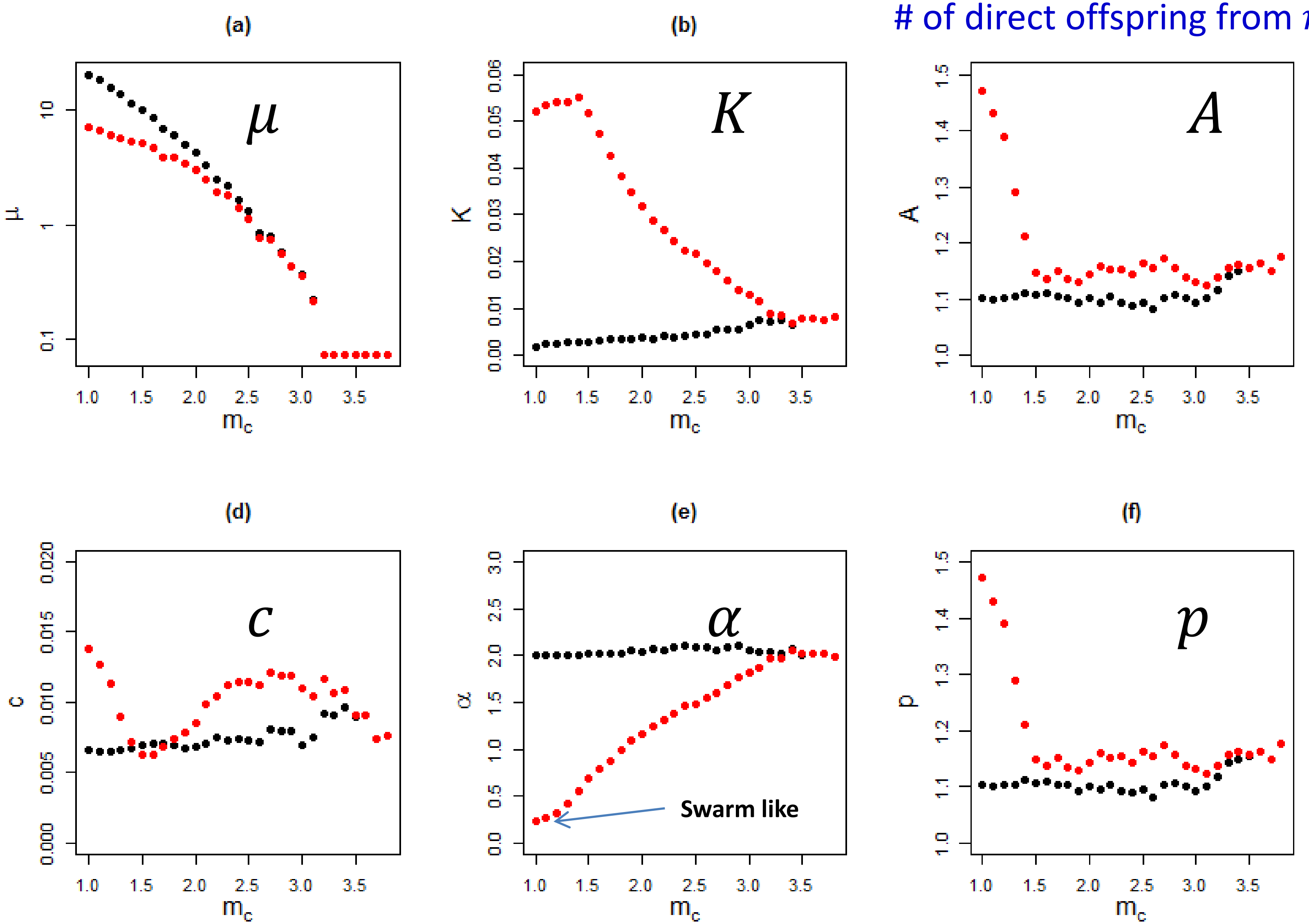
$$\lambda(t) = \mu + K \sum_{i:t_i<t} \frac{e^{\alpha(m_i - m_c)}}{(t - t_i + c)^p}$$ (Ogata 1988, JASA)

$A = \frac{Kc^{1-p}}{p-1} = \int_0^\infty \frac{K}{(t+c)^p} dt$: expected # of direct offspring from m_c

Influence of short-term aftershock missing on estimating the ETAS model

$$A = \frac{Kc^{1-p}}{p-1}$$

# of direct offspring from m_c

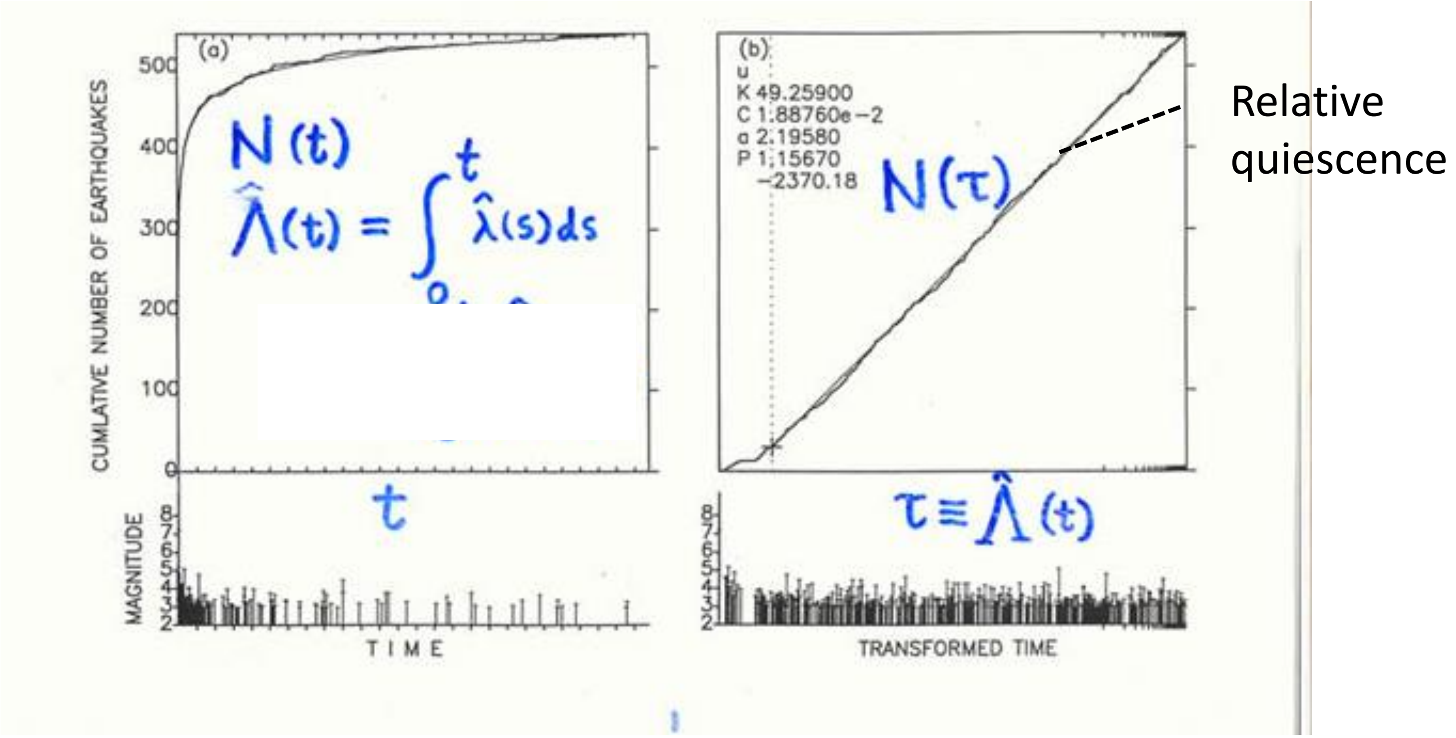


• • • • • • Replenished data

• • • • • • Observed data

Quiescence related to the ETAS model

. **Transformed Time sequence (Ogata, 1992, JGR)**

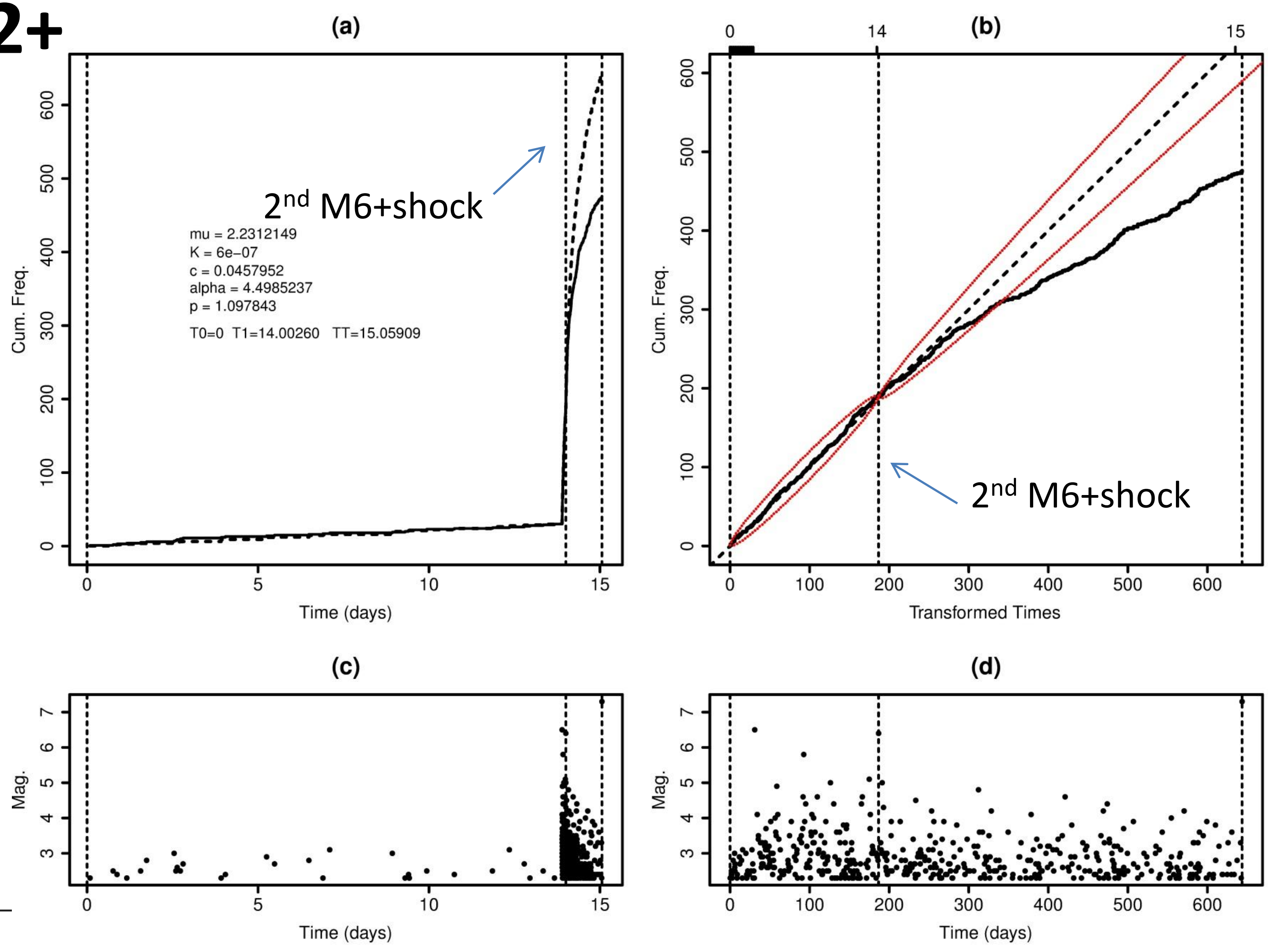


$$t_i \rightarrow \tau_i = \int_0^{t_i} \lambda(u)du$$

If $\{t_i\}$ is the observation of a process determined by conditional intensity $\lambda(t)$, the $\{\tau_i\}$ is a standard Poisson process.

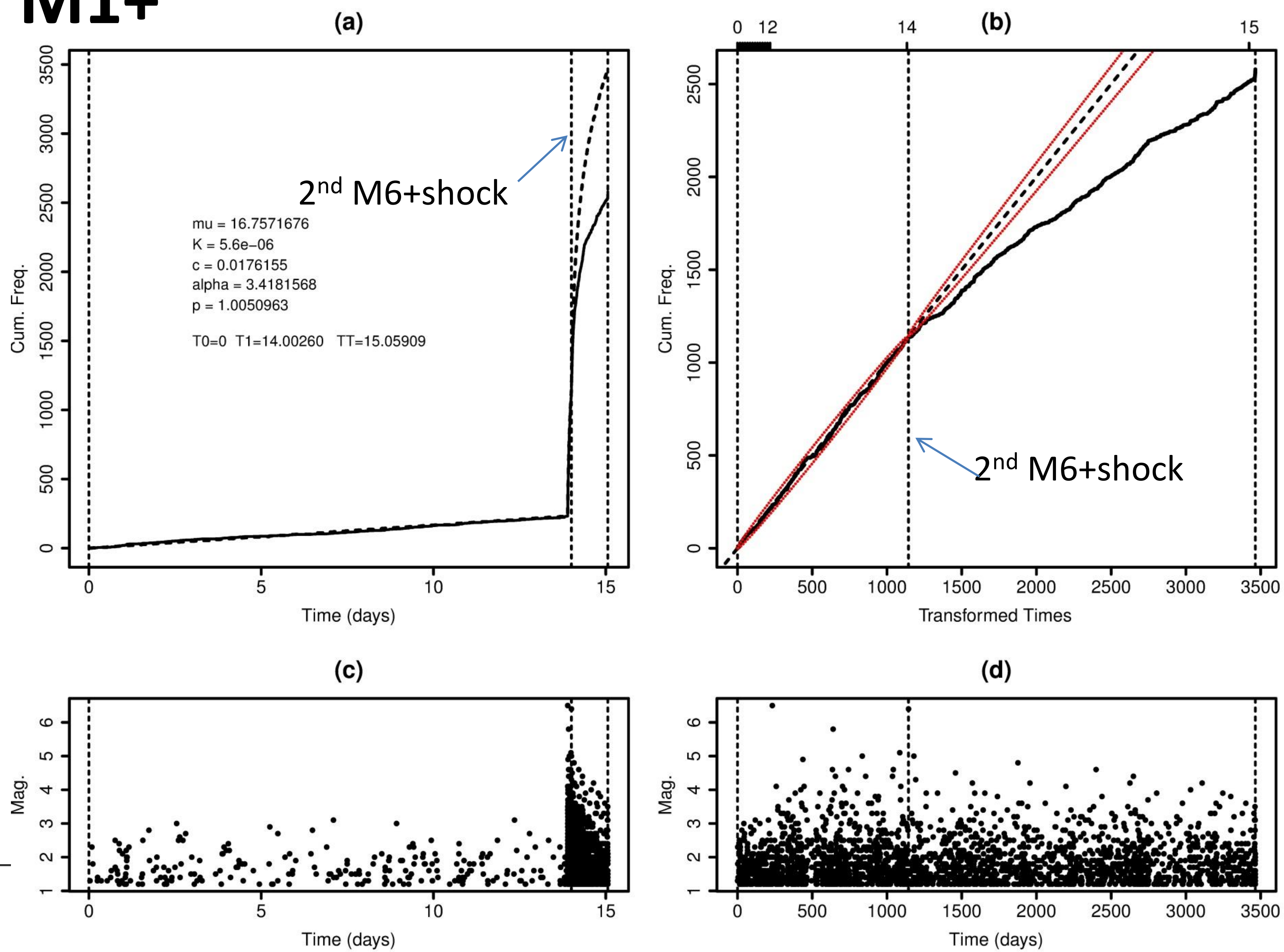
Relative quiescence --- original dataset

M2+



Relative quiescence --- Replenished dataset

M1+

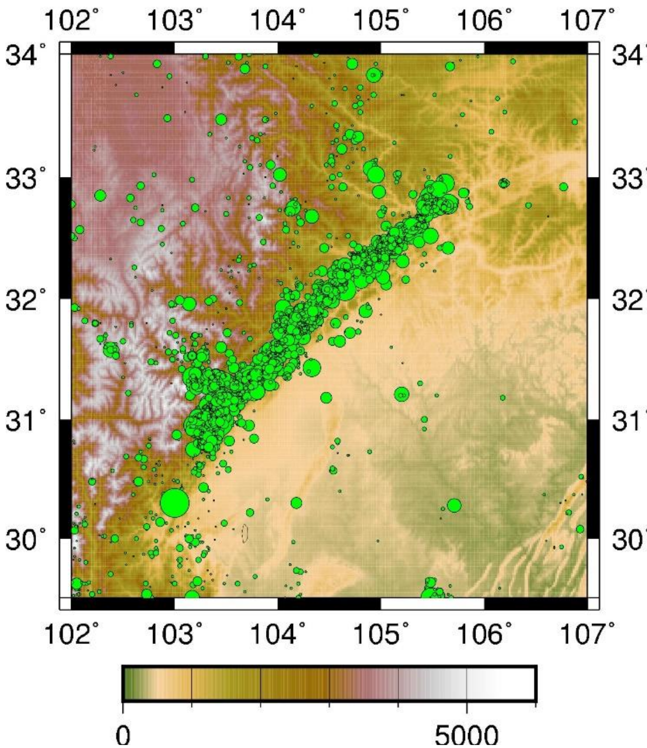


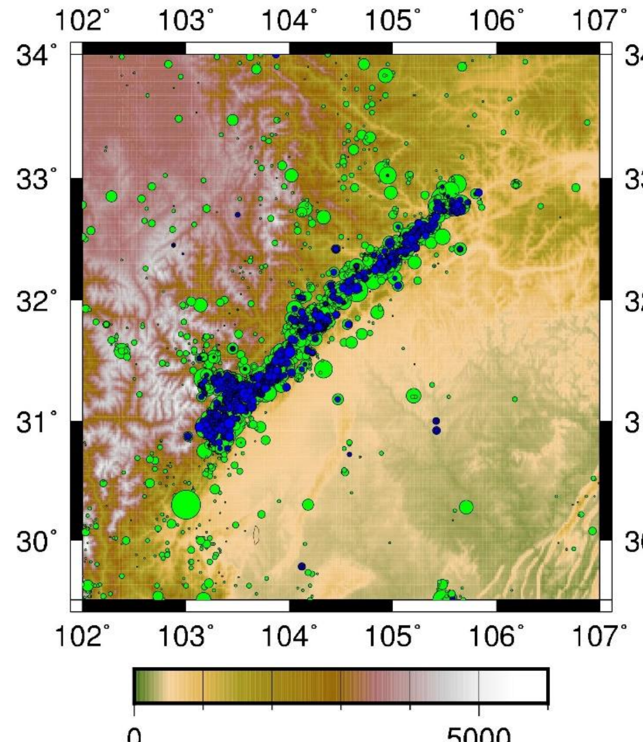
Conclusions

- A method for replenishing missing data in marked temporal point process, which only makes use of the assumption that the marks and occurrence times of events are independent or that the mark distributions has no dramatic changes along the time axis, regardless of how the events interact on the time axis.
- The key point is an iterative algorithm for estimating the missing area in the transform domain according to the parts where data are completely recorded.
- This method is applied to the eruption of the Hakone volcano in Japan and the earthquake catalogue from Southwest China including the aftershock zone of the 2008 M7.8 Wenchuan earthquake

Future Researches

- Extending to higher dimension or spatiotemporal cases.
- If available, more information of the marks or times can be used to improve the replenishment. For example, distribution of earthquake magnitudes is well known (G-R law).





Systems natics

Thank you for listening.