PROCEEDINGS

Edited By: Giada Adelfio and Antonino Abbruzzo

GRASPA 2023

GRASPA-SIS BIENNAL CONFERENCE

The Researcher Group for Environmental Statistics of The Italian Statistical Society

TIES EUROPEAN REGIONAL MEETING

The International Environmetrics Society

Palermo, 10-11 July, 2023

Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo

Proceedings of the GRASPA 2023 Conference
Palermo, 10-11 July 2023
Edited by: Giada Adelfio and Antonino Abbruzzo
-
Palermo: Università degli Studi di Palermo.

ISBN: 979-12-210-3389-2

Sponsored by:

Contents

Preface

The international conference GRASPA2023, held in Palermo, Italy, 10-11 July 2023, is organised by the Department of Economics Business and Statistics of the University of Palermo. GRASPA has become a permanent working group of the Italian Statistical Society (SIS) since May 2013. GRASPA2023 is the biennial conference of the Italian research group for Environmental Statistics GRASPA-SIS and the major event on Environmental Statistics in Italy. This conference endorses co-operation among statisticians, academics as well as practitioners from government and independent environmental agencies, and it allows sharing of research interests related to the development and the use of statistical methods in environmental sciences. Moreover, GRASPA2023 is the 2023 European regional conference of The International Environmetrics Society (TIES).
Brian Reich (North Carolina State University) and Frederic Schoenberg (University of California Los Angeles) are the keynote speakers, and six invited tracks on various statistical and environmental topics are the conference's core. More than thirty contributed papers, presented in an extensive poster session, and informal moments of discussion and interaction between the participants will enrich the conference. Moreover, a Best Poster Award Committee will evaluate all posters on their design, clarity of the presentation and scientific content.
The book of short papers of GRASPA2023 includes abstracts of the keynote, and papers of invited and contributing authors, listed alphabetically according to the conference sessions and the first author's family name. Notably, extended versions of a selection of invited and contributed papers will be considered for publication in a Springer book series' Springer Proceedings in Mathematics & Statistics', with the associated Electronic ISSN 2194-1017 and Print ISSN 2194-1009.
We want to thank all the reviewers for their constructive comments on the papers and the members of the organising team and the scientific committee. Special acknowledgement goes to the generous partners for their overall sponsorship.
Please enjoy the technical program, enjoy Palermo, and continue studying new statistical methods and approaches, hopefully contributing to fostering Environment protection.

Palermo, July 10th, 2023

Giada Adelfio and Antonino Abbruzzo
Editors

5. Spatio-temporal modelling for urban

5.1 - Stochastic reconstruction of a spatio-temporal Hawkes process with isotropic excitation: an application to road accidents
5.2 - ARPALData: an R package for retrieving and analyzing air quality and weather data from ARPA Lombardia
5.3 - Mechanistic spatio-temporal modeling of infectious diseases and crime data on urban environments

Stochastic reconstruction of a spatio-temporal Hawkes process with isotropic excitation: an application to road accidents

Pierfrancesco Alaimo Di Loro[1,*] and Marco Mingione[2]

[1] *Department GEPLI, Libera Università Maria Ss. Assunta (LUMSA); p.alaimodiloro@lumsa.it*
[2] *Department of Political Sciences, Roma Tre University; marco.mingione@uniroma3.it*
[*] *Corresponding author*

Abstract. *We propose a semi-parametric procedure to estimate a periodic spatio-temporal Hawkes process model for the road accidents occurred in Rome during 2021. The model's specification includes daily and weekly components able to catch the cyclical occurrence of such events and a spatial intensity function to derive accurate risk-maps of traffic collisions. The model also envisions a spatio-temporal excitation component that, differently from previous attempts, preserves an isotropic behavior in space. Estimation is performed using the stochastic reconstruction algorithm, where ad-hoc boundary correction strategies have been implemented to reduce the well-known bias on the borders. Results confirm the inhomogeneous and clustered pattern characterizing such phenomena and show that* $\approx 1\%$ *of all occurred crashed in the study area are an effect of previous events.*

Keywords. *Semi-parametric; Hawkes process; EM algorithm; Space-time; Road accidents.*

---

1 Introduction

According to the latest report of the National Statistical Institute, more than 150,000 road accidents occurred during 2021 in the Italian territory, of which 2,875 were deaths within 30 days. Early estimates for the first semester in 2022 (January-June) highlight an increasing – though not alarming – number of road accidents compared to the pandemic period. The major causes of such events can be addressed to distracted driving, failure to give way, and speeding. Although the long-term trend of road accidents in Italy has been decreasing in the last two decades, the social costs of such events (in economic and lives terms) is still huge, making their monitoring essential for the identification crashes hot-spots, and the timely implementation prevention policies. Recent technological developments are finally enabling straightforward and cheap ways of recording the exact space-time location of the vehicles at the moment of the accident, favouring the implementation of advanced statistical techniques to model the dynamics of the accidents occurrences. For example, since 2014, road authorities of the City of Rome are assigned the task to record time and location of every car accident on which their intervention is required. These geo-referenced records are collected monthly and published for public use at `https://dati.comune.roma.it/catalog/dataset?tags=Incidenti&groups=sicurezza-urbana`, in the spirit of the *open-data* [10]. This huge gold-mine of public data has already drawn some attention [7, 3].
The most natural way to model and describe the observed point-pattern is represented by spatio-temporal point-processes [8, 9]. A very interesting challenge in road-accident modeling is understanding whether the occurrence of a road accident increases the risk of other events in its close proximity [1] and, if so, quantifying the number of subsequent crashes that have been triggered by the first occurrence. This sort of cascading effect might be due to the direct of effect of the original crash or its indirect consequences:

increased traffic congestion, lane reduction, etc. This particular dynamic is known as *self-excitation* and it is the defining property of the Hawkes process [13]. It has been largely used in the literature to characterize the clustered point-patterns arising from earthquakes [20, 24] or financial shocks [4, 14]. Its application has been recently extended to many more contexts: crimes [19, 22], infectious diseases [6, 5], and finally road accidents [17, 15].
For estimation purposes, we here take the lead from the stochastic-reconstruction of the spatio-temporal Hawkes process of [22] and propose alternative boundary correction and smoothing techniques. The work is similar in spirit to [15], who applies the model to road-accidents occurring on a single (uni-dimensional) one-way road. We work on the full Euclidean space and smooth the excitation function in such a way that its isotropy is preserved. More details are provided in Section 2. We include an original application on the road accidents occurring on the Rome surface in 2021, as described in Sections 3. Finally, Section 4 provides a final discussion.

2 The model

We seek to model the occurrence of traffic collisions over a spatio-temporal domain $\mathcal{Q} = \mathcal{D} \times [0,T]$, where $\mathcal{D} \subseteq \mathbb{R}^2$ denotes the spatial dimension. We assume that the number of car-crashes $N(B \times [t_1,t_2])$, where $B \subset \mathcal{D}$ and $[t_1,t_2] \subset [0,T]$ is the result of a simple and (locally) finite spatio-temporal point process, that can be defined through suitable specification of the conditional intensity function $\lambda_c(\boldsymbol{s},t) = \lim_{d\boldsymbol{s},dt \to 0} \frac{\mathbb{E}\left[N(d\boldsymbol{s},dt)\middle|\mathcal{H}_t\right]}{d\boldsymbol{s} \cdot dt}$, where $\mathcal{H}_t = \{(\boldsymbol{u},\tau)\}_{\tau<t}$ is the history of he process up to time t (t excluded). As suggested in Section 1, point patterns arising from road accidents present an inhomogeneous and clustered pattern: (i) events are not spread uniformly along the observed region $\mathcal{Q}$, mainly because the risk of collision is affected by environmental factors; (ii) events may exhibit a self-excitation behaviour due to the sudden and unexpected slowdowns or other consequential traffic events.
Here, we consider the space-time separable, periodic and non-parametric version of the Hawkes process model proposed by [22]. Its conditional intensity function has the following form:

$$\lambda_c(\boldsymbol{s},t) = \mu_0 \cdot \mu_s(\boldsymbol{s}) \cdot \mu_t(t) + A \cdot \int_0^t \int_{\mathcal{D}} g_s(\boldsymbol{s}-\boldsymbol{u}) \cdot g_t(t-\tau) N(d\boldsymbol{u} \times d\tau). \tag{1}$$

where $\mu_s(\cdot),\mu_t(\cdot)$ are the spatial and temporal background intensities, $g_s(\cdot),g_t(\cdot)$ are the spatial and temporal excitation functions, $\mu_0, A > 0$ are two real-valued parameters that have the role of *relaxation coefficients*. The background is a function describing the space-time varying general risk of a car accident. The spatial term is static along time and accounts for differing risk of collision due to low visibility, higher presence of intersections, etc. The temporal term is constant through space and it is obtained as the product of three further functions $\mu_t(\cdot) = \mu_d(t) \cdot \mu_w(t) \cdot \mu_{tr}(t)$, where $\mu_d(\cdot),\mu_w(\cdot)$ model the daily and weekly periodicity and $\mu_t(\cdot)$ represents the long-term trend. Daily and weekly seasonality is an ubiquitous feature of traffic data reflecting daily 'rush-hour' commuting patterns, and weekly differences such as between the 5-day working week to the 2-day weekend. The excitation component describes the increase in risk occurring in the close space-time proximity of a previous accident. [22] considers a potentially anisotropic excitation in space, whose magnitude depends on the whole separation vector $\boldsymbol{s}'-\boldsymbol{s}' = (x'-x, y'-y)$ with $\boldsymbol{s} = (x,y)$. We retain the space-time separability but modify this expression by making it dependent on the euclidean distance only, i.e. $g(\boldsymbol{s}'-\boldsymbol{s}, t'-t) = g_s(||\boldsymbol{s}'-\boldsymbol{s}||) \cdot g_t(|t'-t|)$.
Since pursuing a completely non-parametric estimation of such a complex model may suffer of a number of numerical and statistical complications, the background and excitation functions are forced to have average equal to 1 and integral equal to 1, respectively. Therefore, the two *relaxation coefficients* μ_0 and A are introduced to regulate the overall level of the background and excitation components.
The non-parametric estimation of every component in (1) can be obtained through the *stochastic reconstruction* algorithm [24, 21].

2.1 Stochastic reconstruction

Let $\{(\boldsymbol{s}_i,t_i)\}_{i=1}^N$ be an observed point pattern over region $\mathcal{D}\times\times[0,T]\subset\mathbb{R}^2\times\mathbb{R}^+$. The stochastic reconstruction relies on a combination of weighted kernel averaging, to smooth separately each component, and the Expectation-Maximization (EM) algorithm, to update the weights at each iteration until convergence.

The background components are smoothed by averaging each single event $(\boldsymbol{s}_i,t_i)$ with weights proportional to the relative importance of the background intensity on the overall intensity in that same location:

$$\phi(\boldsymbol{s}_i,t_i)=\mu_0\frac{\widetilde{\mu}_s(\boldsymbol{s}_i)\cdot\widetilde{\mu}_d(t_i)\widetilde{\mu}_w(t_i)\widetilde{\mu}_t(t_i)}{\lambda_c(\boldsymbol{s}_i,t_i)},\quad \text{where } \phi(\boldsymbol{s}_i,t_i)=\phi_i forall\ i. \tag{2}$$

The reconstruction of the excitation components is obtained by weighted kernel smoothing of all the observed pairwise space and time lags, i.e. given two arbitrary points $\boldsymbol{p}=(\boldsymbol{s},t)$ and $\boldsymbol{p}'=(\boldsymbol{s}',t')$ the lag vector is $\boldsymbol{p}-\boldsymbol{p}'=(||\boldsymbol{s}-\boldsymbol{s}'||,t-t')$. Each pair contributes with a weight proportional to the relative importance of its specific inter-event excitation with respect to the overall intensity on the *excited* location:

$$\rho(\boldsymbol{s},t,\boldsymbol{s}',t')=\begin{cases}\frac{A\cdot\widetilde{g}_s(||\boldsymbol{s}'-\boldsymbol{s}||)\cdot\widetilde{g}_t(|t'-t|)}{\lambda_c(\boldsymbol{s}',t')} & t<t'\\ 0 & t\geq t',\end{cases}\text{, with } \rho(\boldsymbol{s}_i,t_i,\boldsymbol{s}_j,t_j)=\rho_{ij}\quad\forall\, i,j. \tag{3}$$

Theoretical justification to the weighting scheme is based on [8] and it is further discussed in [22].

The *isotropic* smoothing of the spatial excitation function deserves further discussion. Indeed, we must consider that the larger the distance and the wider is the circle on which two points at the same distance may lie. Therefore, the naive smoothing of inter-event distances is not representative of the real intensity at each specific distance, as such intensities are spread across circles of different sizes. We can re-weight the contribution of each pair of points by the inverse of the circumference it lies upon, similarly to what it is done when estimating the *pair-correlation* function [9], i.e.:

$$\hat{g}_s(d)\propto\sum_{i,j=1}^{n}\rho_{i,j}\frac{\widetilde{k}_{h_s}(d-(d_{ij}))}{2\pi d_{ij}},$$

where $d_{ij}=||\boldsymbol{s}_j-\boldsymbol{s}_i||$. This procedure has the only drawback of being not very robust for the distance approaching 0 as the variance of the estimate grows indefinitely. However, we can stop at a reasonable small value $\upsilon>0$ and extend the estimate down to 0 linearly without much loss of information.

Finally, it is well known how kernel smoothing is affected by severe boundary issues when the domain is limited. This problem must be tackled separately, and differently, for each component. We consider a *buffering area* $\mathcal{B}\supset\mathcal{Q}$ to overcome this in the spatial and trend components of the background. This amounts to collecting and considering data just outside the domain $\mathcal{Q}$ when smoothing but not when evaluating the likelihood [23]. We consider the *periodic* correction for the remaining background components, as in [15].

Finally, the triggering function have a single natural border in 0. We considered a correction approach based on domain transforms. In practice, the pairs distances $||\boldsymbol{s}_j-\boldsymbol{s}_i||$ and $|t_j-t_i|$ are converted to the log-scale $\log(||\boldsymbol{s}_j-\boldsymbol{s}_i||)$ and $\log(|t_j-t_i|)$, so to eliminate the border, and smoothed. The resulting density is then re-converted into the original scale by applying the jacobian [18, 16].

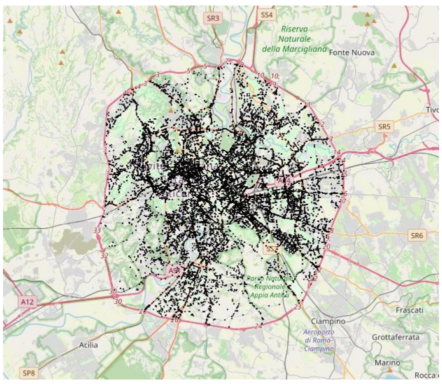
Once all the functions involved in the expression of the conditional intensity function have been estimated and normalized, we must estimate the two relaxation coefficients μ_0 and A. The most natural way to estimate the values that best fit the observed point pattern would be maximum likelihood estimation (MLE), conditionally on the non-parametrically estimated background and intensity functions. The log-likelihood function of $\{(\boldsymbol{s}_i,t_i)\}_{i=1}^n$, realization of a finite and simple spatial-temporal point process over a bounded domain $\mathcal{Q}=\mathcal{D}\times[0,T]$, is known and can be evaluated as:

$$\log(\mathcal{L}_{\lambda_c})=\sum_{i=1}^{n}\log\left(\lambda_c(\boldsymbol{s}_i,t_i,)\right)-\int_0^T\int_{\mathcal{D}}\lambda_c(\boldsymbol{\sigma},\tau)d\boldsymbol{\sigma}d\tau, \tag{4}$$

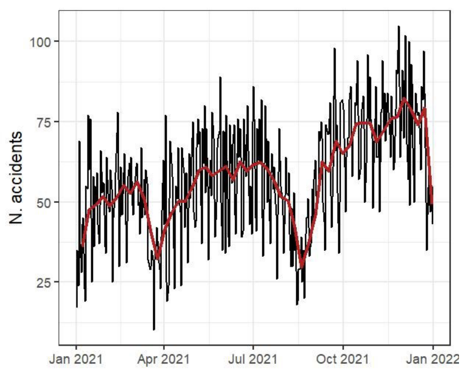
for whatever $\lambda_c(\cdot,\cdot)$. In particular, we consider the conditional intensity function defined in (1) and rely on numerical optimization methods to get the MLE of μ_0 and A.
It is clear how the relaxation coefficients are needed in order to weight the smoothing of the various components, while the smoothed versions of each components are needed in order to get $\log(\mathcal{L}_{\mu_0,A})$ and estimate μ_0 and A. This circularity can be solved as proposed by [22], specifically alternating between finding the optimal relaxation coefficients given background and excitation forms and viceversa, within the EM algorithm. For the sake of brevity, we do not report here the algorithm, but point the interested reader to the original paper of [22]. Convergence can be checked using different criteria. Eventually, after a reasonable number of iterations (which is application-dependent), the log-likelihood increments flatten and a (local) maximum has been found. There is no guarantee that the optimum is global, therefore multiple runs starting from different initial guesses are suggested. One main drawback of this algorithm resides in its computational complexity, that is mostly affected by the smoothing of the excitation function (that scales quadratically with the number of points) and the likelihood maximization. One idea to reduce the burden of the former, is to assume that triggering gets negligible after a pre-specified distance in time and/or space (excitation *tapering*). This can sensibly reduce the number of eligible pairs for the smoothing. In order to make this procedure viable on our large set of data, the core part of the algorithm has been coded in `C++`, exploiting parallelization whenever possible.

3 Application to road accidents in Rome

We estimate the model described in Section 2 to all road accidents that occurred in 2021 inside the Grande Raccordo Anulare, the major ring road surrounding Rome city center, for which the intervention of the local Police Department was required (i.e. does not include all the cases in which the involved subjects achieved a friendly agreement). A total of $n = 21012$ occurred in 2021, with a weekly rate of 404 road accidents, also corresponding to more than 55 each day and 2.4 each hour. Figure 1a shows the purely spatial point pattern through the whole year, while Figure 1b the corresponding daily (and weekly in red) temporal evolution. These highlight a slight increasing trend during the whole study period with evident seasonal behavior and a couple of main hot-spots: one placed in the city center (behind Vatican City) and another one in the South-East part of the city.



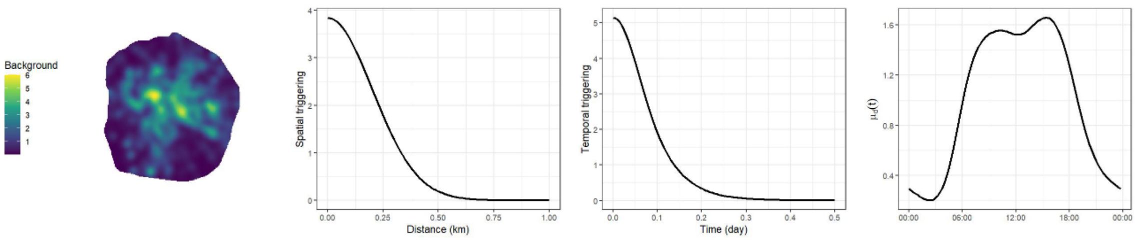
(a)



(b)

Figure 1: (a) spatial locations and (b) time-series of daily (black line) and weekly (red line) road accidents occurred in Roma Capitale during 2021.

Results E-M algorithm converged after 50 iterations in less than 1 hour of computational time, yielding the estimates of $\hat{\mu}_0 = 0.127$ and $\hat{A} = 0.013$. The long-term trend component (not shown for the sake of brevity) highlights seasonal variations, with the major peaks estimated to occur in July, November and January, and major valleys estimated in August and December. This can be mainly explained by the fact that Roman citizens usually rush before summer and Christmas Holidays to finish all their open businesses, and then leave the city soon after for vacation. Figure 2d shows the estimted daily periodic component and highlights a higher risk during the central hours of the day (7 a.m. – 7 p.m.), with the peaks estimated from 8 a.m. to 10 a.m. (rush hours in the morning where people go to work and schools start) and from 3.30 p.m. to 5.30 p.m.. Figure 2a shows the estimated spatial background, highlighting two major risk areas: one placed right behind Vatican City and another one at big crossroad between Via Statilia, Via San Quintino and Via Carlo Emanuele Primo, on the west side of Rome Central Station which is usually very busy during the whole day. Spatial triggering is limited at 1km radius, however the triggering effect is estimated to be practically null over 500m already (see Figure 2b). The estimated temporal triggering is instead visualized in Figure 2c and unsurprisingly decreases as time passes from the occurrence of an event: in particular, the triggering is practically null after 1/4 of a day, i.e. 6 hours.



(a) (b) (c) (d)

Figure 2: (a) estimated spatial spatial background intensity. Estimated (b) spatial and (c) temporal excitation components. (d) estimated daily background

4 Discussion

This paper proposes some new strategies to fit semi-parametrically a periodic spatio-temporal Hawkes process while correcting for the boundary biases and preserving isotropy in the spatial excitation component. We provide a specific application to road accidents following previous attempts that were limited due to computational limitations [2]. We here attain improved performances by implementing the core of the algorithm in `C++`, allowing the consideration of data for a whole year.
Nevertheless, there is still large room for improvements as the method is still tailored to euclidean spaces. However, when dealing with road accidents, it should account for the urban road network. [12, 11] consider the linear network in a similar model specification, and we are currently working to combine their ideas and our proposal in an efficient and feasible fashion so that the full application could be adapted to the linear road network of Rome.

References

[1] Benjamin Acker and May Yuan. Network-based likelihood modeling of event occurrences in space and time: a case study of traffic accidents in Dallas, Texas, USA. *Cartography and Geographic Information Science*, 46(1):21–38, 2019.

[2] Pierfrancesco Alaimo Di Loro. Innovative approaches in spatio-temporal model-ing: handling data collected by new technologies. 2021.

[3] Pierfrancesco Alaimo Di Loro, Enrico Ciminello, and Luca Tardella. Hidden Markov Model estimation via Particle Gibbs. *Book of Short Papers SIS 2019*, pages 829–835, 2019.

[4] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

[5] Álvaro Briz-Redón, Adina Iftimi, Jorge Mateu, and Carolina Romero-García. A mechanistic spatio-temporal modeling of covid-19 data. *Biometrical Journal*, 65(1):2100318, 2023.

[6] Wen-Hao Chiang, Xueying Liu, and George Mohler. Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. *International journal of forecasting*, 38(2):505–520, 2022.

[7] Antonio Comi, Luca Persia, Agostino Nuzzolo, and Antonio Polimeni. Exploring temporal and spatial structure of urban road accidents: Some empirical evidences from rome. In *The 4th Conference on Sustainable Urban Mobility*, pages 147–155. Springer, 2018.

[8] Daryl J Daley and David Vere-Jones. An introduction to the theory of point processes. Vol. I. Probability and its Applications, 2003.

[9] Peter J Diggle. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press, 2013.

[10] Gianni Dominici. Per una seconda fase degli open data in italia. *PRISMA Economia-Società-Lavoro*, 2015.

[11] Nicoletta D'Angelo, Giada Adelfio, Antonino Abbruzzo, and Jorge Mateu. Inhomogeneous spatio-temporal point processes on linear networks for visitors' stops data. *The Annals of Applied Statistics*, 16(2):791–815, 2022.

[12] Nicoletta D'Angelo, David Payares, Giada Adelfio, and Jorge Mateu. Self-exciting point process modelling of crimes on linear networks. *Statistical Modelling*, page 1471082X221094146, 2022.

[13] Alan G Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971.

[14] Alan G Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198, 2018.

[15] Kieran Kalair, Colm Connaughton, and Pierfrancesco Alaimo Di Loro. A non-parametric hawkes process model of primary and secondary accidents on a uk smart motorway. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2020.

[16] Gerhard Koekemoer and Jan WH Swanepoel. Transformation kernel density estimation with applications. *Journal of Computational and Graphical Statistics*, 17(3):750–769, 2008.

[17] Zhongping Li, Lirong Cui, and Jianhui Chen. Traffic accident modelling via self-exciting point processes. *Reliab. Eng. Syst. Safe.*, 180:312 – 320, 2018.

[18] James Stephen Marron and David Ruppert. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):653–671, 1994.

[19] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

[20] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.

[21] Jiancang Zhuang. Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(4):635–653, 2006.

[22] Jiancang Zhuang and Jorge Mateu. A semiparametric spatiotemporal hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(3):919–942, 2019.

[23] Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380, 2002.

[24] Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research: Solid Earth*, 109(B5), 2004.