Journal of Quantitative Criminology
https://doi.org/10.1007/s10940-023-09580-1

ORIGINAL PAPER

A Bayesian Aoristic Logistic Regression to Model Spatio-Temporal Crime Risk Under the Presence of Interval-Censored Event Times

Álvaro Briz-Redón[1]

Accepted: 14 December 2023

Abstract

Purpose Crime data analysis has gained significant interest due to its peculiarities. One key characteristic of property crimes is the uncertainty surrounding their exact temporal location, often limited to a time window.

Methods This study introduces a spatio-temporal logistic regression model that addresses the challenges posed by temporal uncertainty in crime data analysis. Inspired by the aoristic method, our Bayesian approach allows for the inclusion of temporal uncertainty in the model.

Results To demonstrate the effectiveness of our proposed model, we apply it to both simulated datasets and a dataset of residential burglaries recorded in Valencia, Spain. We compare our proposal with a complete cases model, which excludes temporally-uncertain events, and also with alternative models that rely on imputation procedures. Our model exhibits superior performance in terms of recovering the true underlying crime risk.

Conclusions The proposed modeling framework effectively handles interval-censored temporal observations while incorporating covariate and space–time effects. This flexible model can be implemented to analyze crime data with uncertainty in temporal locations, providing valuable insights for crime prevention and law enforcement strategies.

Keywords Bayesian statistics · Censored data · Crime data · Data imputation · Spatio-temporal models · Temporal uncertainty

Introduction

The use of advanced statistical techniques for crime analysis has experienced significant growth in the last decade. In particular, it is of special interest to analyze the distribution of crime in space and time and also to study the presence of space–time interaction. For this reason, different modeling approaches have been developed and adapted to find environmental factors that are associated with a higher risk of crime, as well as to predict crimes in

✉ Álvaro Briz-Redón
alvaro.briz@uv.es

[1] Department of Statistics and Operations Research, Facultad de Ciencias Matemáticas, University of Valencia, C/ Dr. Moliner, 50, 46100 Burjassot, Valencia, Spain

Published online: 11 January 2024

the short-term and the mid-term. Among others, different versions of self-exciting models (Mohler et al. 2011; Zhuang and Mateu 2019), spatial models with a non-linear structure (Briz-Redón et al. 2022b), and Bayesian univariate/multivariate spatio-temporal models (Chung and Kim 2019; Law et al. 2014; Li et al. 2014; Quick et al. 2018) have been proposed.

Crime data often presents singular characteristics that can complicate the analysis, which, in turn, leads to the development of new methodologies. For instance, it is well acknowledged that crime figures are usually underestimated, or even biased (Buil-Gil et al. 2021, 2022). Besides, in the context of spatial, temporal, or spatio-temporal crime datasets, both the spatial and the temporal accuracy are often a matter of concern. Problems with spatial accuracy usually refer to the impossibility of identifying the spatial unit in which the event has occurred (considering, for example, an administrative division of the study area). In particular, if the spatial location of the event is available in the form of textual information (representing the human-readable address of the location), it is common to have geocoding errors or events that fail to be geocoded. This type of problem has been discussed previously, to establish minimally acceptable geocoding rates (Andresen et al. 2020; Briz-Redón et al. 2020; Ratcliffe 2004). Regarding the issues related to temporal accuracy, it is also usual that for some crime events, we do not observe their temporal location with the desired accuracy (minute, hour, date, etc.). Indeed, this situation takes place in most events if we are dealing with certain types of crime such as property theft (Ashby and Bowers 2013). In this case, what we have is a lower and an upper bound of the temporal location of the event, that is, a temporal interval or window for each of the events. This kind of temporal observation is usually referred to as interval-censored.

The existence of interval-censored event times is a well-known issue in the field of quantitative criminology. Even though it may not have received enough attention, there are different strategies to deal with interval-censored event times in criminal records. The simplest approach is to choose an appropriate time unit that eliminates temporal uncertainty. For example, if the existing uncertainty is at the day level, we can operate at the week level. Although in some cases this approach allows us to eliminate the temporal uncertainty from most (or all) observations, setting the temporal resolution of the analysis based on the uncertainty surrounding the observations is not desirable. For example, it might happen that for some datasets there is no uncertainty at the month level, however, performing a spatio-temporal analysis considering months as the temporal unit may be too coarse to be useful from a practical point of view.

Another way to deal with uncertainty is to perform the imputation of event times, a common strategy that can be performed in several ways. One possibility is to assign to these events the temporal unit (hour, date, etc.) that lies just at the midpoint of the uncertainty time window reported, or a random temporal unit within the window. Considering the initial or the final temporal location of the time window is another option, but these two approaches are typically biased (Ashby and Bowers 2013). However, these imputation-based methods are usually outperformed by the one called the aoristic method, as shown by Ashby and Bowers (2013).

The aoristic method, which has been explored and proposed mainly by criminologist J. Ratcliffe (Ratcliffe and McCullagh 1998; Ratcliffe 2000, 2002), consists of assigning the same weight to each time unit included in the interval that delimits the temporal uncertainty about the crime event. Hence, when one uses the aoristic approach, temporally-uncertain events do not receive a single imputation value, but a probability score for each of the temporal units within which the event is located. Specifically, all the temporal units receive the same score or weight, so that they add up to 1. For this reason, the aoristic method does not entirely correspond to an imputation method. Nevertheless, by following the aoristic procedure, one can, for instance, deduce the temporal distribution of a set of crimes by adding both the number of temporally-certain events and the fractions of temporally-uncertain events corresponding to each date within the period. Even though the aoristic approach allows carrying out some exploratory analyses of crime datasets including interval-censored temporal observations, this may not be sufficient depending on the purpose of the analysis. In particular, if we are interested in describing the temporal or spatio-temporal distribution of a set of crimes that present temporal uncertainty, it is often helpful to introduce covariate information that contributes to the understanding of the crime under analysis. Thus, to be able to explain a pattern of crime events while handling extra information in the form of covariates, it is advisable to explicitly model the temporal uncertainty that the observations present. This goes beyond the classical aoristic method, which aims at allowing an exploratory analysis of the data while taking into account the existence of interval-censored observations.

Indeed, model-based approaches can also be followed to deal with interval-censored observations. For instance, the von Mises distribution or Dirichlet processes, which are typically used for the analysis of circular data, have been recently proposed for the analysis of aoristic data (Mulder 2019). Furthermore, one way to deal with the temporal uncertainty of the observations is to include all of them in a single model and assume a certain probability distribution on the exact temporal location of these observations. This leads, in general, to models that handle missing data under the Bayesian framework. Indeed, Reich and Porter (2015) adopted a Bayesian modeling framework for clustering criminal events which, among other features, enabled them to deal with interval-censored event times. Besides, the model proposed by these authors allowed them to link events that share the same *modus operandi* or offender (in case this information is available).

In this paper, a model-based approach is followed to analyze a burglary dataset recorded in Valencia, Spain, which, as is common with residential burglaries, includes a large proportion of temporally-uncertain events. Specifically, a spatio-temporal logistic regression is proposed to model burglary risk in space and time. The model is estimated within a Bayesian framework, allowing the inclusion of temporally-uncertain events. The aoristic approach is imitated when introducing this uncertainty into the model. Therefore, the objective of the paper is twofold. First, to describe a modeling framework to estimate burglary risk in space and time, while accounting for events with temporal uncertainty. Second, to highlight the suitability of including this kind of crime event in the analysis to get more reliable parameter estimates and attempt to recuperate the underlying spatio-temporal distribution of crime. Specifically, we show by fitting the model to several simulated datasets that the proposed modeling framework can outperform other competing alternatives that rely on imputation procedures.

The paper is structured as follows. "Data" section contains a description of the data used for the analysis, emphasizing the presence of temporal uncertainty in the events. "Methodology" section describes the modeling framework proposed for the estimation of burglary risk in space and time, under the presence of interval-censored temporal observations. Then, "Results" section starts with a simulation study to assess the performance of the proposed model under different scenarios, and compare it against several competing models. The second part of "Results" section is devoted to describing the results obtained for the burglary dataset mentioned above. Finally, "Discussion and Conclusions" section includes a discussion and some concluding remarks.

Data

Study Settings

The case study has been conducted in the city of Valencia, the third most populated city in Spain, with a population size of around 800,000 inhabitants. Specifically, the urban core of the city has been considered for the analysis, excluding some peripheral districts that only represent 5% of the population. Besides, to investigate the spatial distribution of burglaries across the city, the boroughs of Valencia have been considered for analysis. There are 70 boroughs in the study area delimited for the research.

Burglary Data

A dataset provided by the Spanish National Police containing information about 2626 burglaries recorded in the city of Valencia from 1 January 2016 to 31 December 2017 has been used for the analysis. As far as we know, this dataset is exhaustive, in the sense that it contains all the burglaries registered in Valencia by the Spanish National Police during this period, with no missing data due to problems of geolocation. This dataset has already been analyzed by Briz-Redón et al. (2022a) to study the near-repeat phenomenon. In this dataset, the geographical coordinates are available for each of the events, allowing the analysis to be conducted at any desired spatial scale. In contrast, the temporal location of some of the events presents a certain degree of uncertainty. Particularly, the exact date of occurrence of the burglary is known for only the 60.9% of the cases. Further comments on the temporal uncertainty of the data are provided in the following Subsection.

Event Time Uncertainty

One important feature of the dataset under study is the presence of interval-censored event times. Specifically, for each burglary, there is a *from date* and a *to date* variable that allow delimiting the temporal location of the event, based on the information available about the burglary (the *from date* represents the last date on the calendar on which the owners can be sure that the home has not yet been burgled, whereas the "to date" is the date on which the owners, the Police, or any citizen has ascertained that the burglary has been committed). Although the original variables including dates are in a YYYY-MM-DD format, they have been transformed into numeric values to ease their use, assigning a value of 1 to 1 January 2016, which represents the start of the study period. In the remainder of the paper, the dates would be considered as numeric values, unless otherwise stated.

Thus, while there is no spatial uncertainty in the data since the coordinates of each dwelling that has been burgled during the period under study are available (of course, for the burglaries that have been notified to the Police), event time uncertainty cannot be overlooked. In a previous study by Briz-Redón et al. (2022a), the temporal uncertainty issue was resolved through the midpoint date imputation method. A preliminary analysis conducted in the context of that study focused on the near-repeat phenomenon, allowed us to conclude that the imputation method (midpoint date or aoristic date) did not have a strong impact on the results. Therefore, the midpoint date method was preferred over the aoristic given its computational convenience.

In the present paper, the aim is to deal with interval-censored events explicitly, without direct imputation of missing event dates, as will be shown in the subsequent Section.

Methodology

Case–Control Study Design

In order to follow a logistic modeling framework (which will be described in "Logistic Regression Model" section), a binary response variable indicating the presence or absence of a burglary event is needed. Therefore, the spatio-temporal locations (with event time uncertainty) of the burglaries available, denoted by $\{\boldsymbol{x}_i, t_i^{from}, t_i^{event}, t_i^{to}\}_{i=1}^{2626}$, are treated as the cases/events, where $\boldsymbol{x}_i$ are the geographical coordinates corresponding to event i, and t_i^{from}, t_i^{event}, and t_i^{to} represent, respectively, the *from date*, the date at which the event actually occurred, and the *to date*. For events with no temporal uncertainty, it holds $t^{from} = t^{event} = t^{to}$. For temporally-uncertain events, we have $t^{from} < t^{to}$ and a missing value for t^{event}. In other words, we only know that $t^{event} \in [t^{from}, t^{to}]$.

To generate the controls (the spatio-temporal locations where no burglary has occurred), the point pattern formed by the locations of all the dwellings in the study area is taken into account. Specifically, this pattern consists of 28,682 locations within the city, which include 382,539 dwelling units (this corresponds to the dwellings registered in Valencia in 2016). A total of 13,130 control locations were sampled with replacement, setting the probability of selection to be proportional to the number of dwellings in the location. Hence, the number of controls was chosen to be five times the number of cases, so a 5:1 ratio of controls to cases was used. The literature suggests that a 4:1 case–control ratio is generally sufficient to carry out a case-control study design (Gail et al. 1976; Hong and Park 2012). The choice of this ratio will affect probability estimates derived from the logistic regression model, so this should be taken into account when performing binary (event/no event) predictions that depend on a cutoff probability. Each of these control locations was assigned a random date from 1 January 2016 to 31 December 2017. No temporal uncertainty has been assumed for the control data, so $t^{from} = t^{event} = t^{to}$ for all these space–time locations. Figure 1a displays the spatial locations of the cases and controls considered for the analysis over the study area. Figure 1b, c show the density of cases and controls over the city of Valencia, respectively, obtained through kernel density estimation (KDE) techniques, which allow us to appreciate that the distributions of crime events and dwelling locations are notably different.

Logistic Regression Model

If $\{\boldsymbol{x}_i, t_i^{from}, t_i^{event}, t_i^{to}\}_{i=1}^{15756}$ denotes the complete set of spatio-temporal locations considered for the analysis, let y_i be a binary variable indicating if each location represents a case (presence of a burglary event) or a control (absence of a burglary event). As usual, we set $y_i = 1$ if i is a case, and 0 otherwise. In order to model the risk of burglary for each spatio-temporal location within the study window, a logistic regression modeling framework is a natural choice. Under the logistic model, the occurrence of a burglary event at spatio-temporal location i is described through a Bernoulli random variable $Y_i \sim Ber(\pi_i)$, where π_i represents the probability that a burglary is actually observed at location i, according to its characteristics. In this paper, we attempt to explain this parameter in terms of several fixed and random (spatial and temporal) effects, leading to the following expression:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_{DoW(i)} + \delta_{w(i)} + \varepsilon_{w(i)} + u_{b(i)} + v_{b(i)} \qquad (1)$$

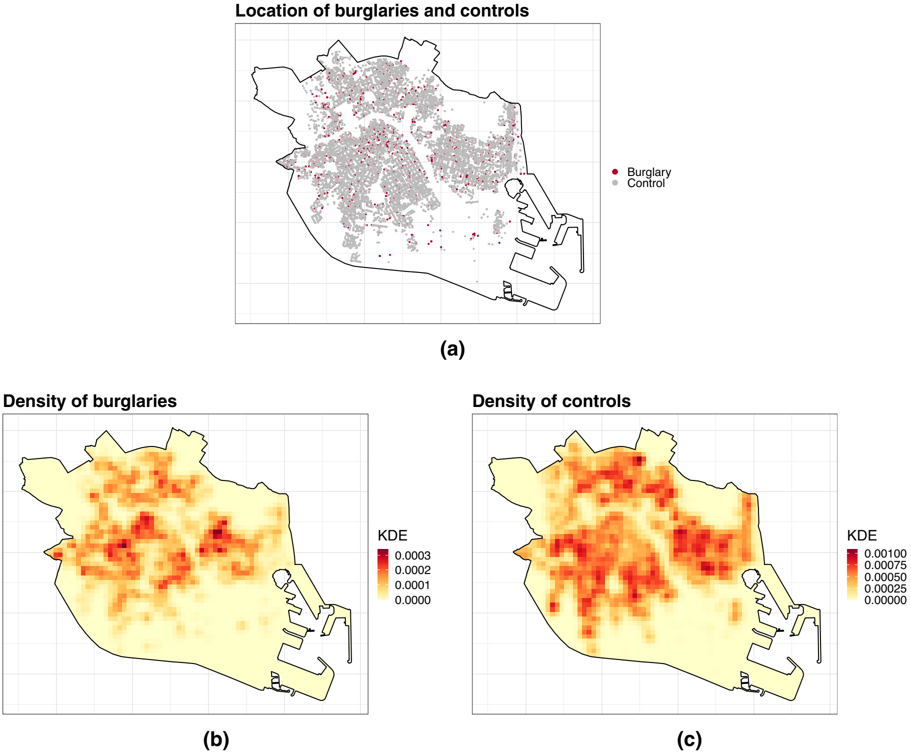


Fig. 1 Location within the study area of the burglaries and controls considered for the analysis (**a**) and kernel density estimation (KDE) values corresponding to the pattern of burglaries (**b**) and controls (**c**)

where each term is defined as follows. First, α is the global parameter of the logistic regression model, which is usually referred to as the intercept of the model. Second, $\beta_{DoW(i)}$ represents the effect of the day of the week corresponding to event i (denoted by $DoW(i)$) on $\text{logit}(\pi_i)$. We note that Monday is taken as the reference level of this variable, so six β_{DoW} parameters are actually estimated, one for each of the remaining days of the week. A vague Gaussian prior, $N(0,\ 1000)$, is assigned to α and the β_{DoW}'s. The rest of the parameters involved in the model represent spatial and temporal random effects, which are outlined in the following lines.

The temporally-structured week effect, δ_w ($w = 1, ..., 104$), is specified through a second-order random walk $\delta_w | \delta_{w-1}, \delta_{w-2} \sim N(2\delta_{w-1} + \delta_{w-2}, \sigma^2_\delta)$, where σ^2_δ is the variance component, whereas an independent and identically distributed Gaussian prior is chosen for the temporally-unstructured week effect, $\epsilon_w \sim N(0, \sigma^2_\epsilon)$ ($w = 1, ..., 104$). The variance components, σ^2_δ and σ^2_ϵ, are assigned a Gamma-distributed prior, $Ga(1,\ 0.5)$, where 1 and 0.5 correspond to the shape and rate parameters of the Gamma distribution. We note that the choice of a Gamma-distributed prior for these parameters is rather arbitrary since other distributions could be considered as well. The choice of a second-order random walk for the temporally-structured week effect, δ_w, allows accounting for temporally-correlated effects in burglary risk that are not explained by day-of-the-week variations. More specifically, the definition chosen for the term δ_w enables us to model the risk of burglary in week w on the assumption that it depends on the risk estimated in the previous two weeks while giving more weight to the risk estimated in the immediately preceding week. Meanwhile, the temporally-unstructured week effect, ε_w, allows capturing weekly variations in burglary risk that are unrelated to what happened in previous weeks.

Regarding the spatial random effects, the Besag–York–Molliè (BYM) model has been employed (Besag et al. 1991), under which the conditional distribution of the spatially-structured effect on a borough b, u_b ($b = 1, ..., 70$), is

$$u_b | u_{\tilde{b} \neq b} \sim N\left(\sum_{\tilde{b} \neq b=1}^{n} w_{b\tilde{b}} u_{\tilde{b}}, \frac{\sigma_u^2}{N_b} \right)$$

where N_b is the number of neighbors for area b, $w_{b\tilde{b}}$ is the $(b, \tilde{b})$ element of the row-normalized neighborhood matrix ($w_{b\tilde{b}} = 1/N_b$ if boroughs b and $\tilde{b}$ share a geographical boundary, and 0 otherwise), and σ_u^2 represents the variance of this random effect. The spatially-unstructured effect over the areas, denoted by v_b ($b = 1, ..., 70$), follows a Gaussian distribution, $v_b \sim N(0, \sigma_v^2)$, where σ_v^2 is the variance of the effect. It was assumed a Gamma-distributed prior, $Ga(1,\ 0.01)$, for both σ_u^2 and σ_v^2, where again we are using the shape-rate parameterization of the Gamma distribution. The spatially-structured random effects allow borrowing strength from nearby locations (boroughs), leading to more reliable burglary risk estimates across space. Indeed, by defining u_b in this way we model the burglary risk in borough b under the assumption that it depends on the average risk of burglary in the boroughs that are contiguous to b. On the other hand, the spatially-unstructured effect, v_b, enables us to detect those boroughs that exhibit a behavior that is different from that of their neighbors.

At this point, it should be noted that the choice of the structure formed by the boroughs of Valencia to measure the spatial variation of burglary risk is to some extent arbitrary. In this case, the choice of an administrative unit is convenient because it would allow practitioners to design surveillance strategies in a simple way. Moreover, from a computational point of view, it is also advantageous, since there are not so many spatial units in the study area (there are 70 boroughs in total). In any case, the type of modeling framework proposed could be adapted to other types of partitions, regular or irregular, more or less fine, of the study area under consideration.

Dealing with the Temporal Uncertainty

The logistic regression model represented by (1) implicitly assumes that the day of the week (DoW) and the week within the year are known exactly for each event/control location, i. Then, $DoW(i)$ and $w(i)$ are two known values and the corresponding fixed and random effects can be estimated. If the exact date of occurrence of event i is unknown, $DoW(i)$ is unknown, while $w(i)$ is also unknown unless the *from date* and the *to date* belong to the same week of the calendar. In this scenario, one possibility is to discard all the events with an unknown date for the analysis. This avoids dealing with missing data and corresponds to a complete case analysis (considering only the data records with no missing values). Removing all these observations from the analysis leads to a reduction in both precision and power, which is undesired.

Hence, to include all these temporally-uncertain events within the modeling framework, one can treat each missing date as a random variable, as usually done in the context of Bayesian statistics to deal with missing data. Specifically, following the aoristic approach, a uniform prior is assigned to each date, considering the information provided by the *from* *date* and *to date* variables available, that is, $t_i^{event} \sim U(t_i^{from} - 0.5, t_i^{to} + 0.5)$. Then, in each iteration of the Markov Chain Monte Carlo (MCMC) process, a numeric date is sampled according to this distribution, which is rounded (to the nearest integer) to allow the computation of $DoW(i)$ and $w(i)$, and hence the consideration of all the data available in the estimation of the fixed and random effects of the model. For known dates, since $t^{to} = t^{from}$, all the sampled values for t_i^{event} coincide with the exact (known) date. Moreover, note that subtracting and adding 0.5, respectively, to t_i^{from} and t_i^{to} is necessary to avoid reducing the weight of the extremes of the interval corresponding to each interval-censored event in the prior distribution. Specifically, if we do not subtract 0.5 to t_i^{from}, we would only assign to t_i^{from} the sampled values that lie in the $]t_i^{from}, t_i^{from} + 0.5[$ interval, which has length 0.5, whereas we would assign to $t_i^{from} + 1$ (or any other integer number within the interval, except for t_i^{to}) the sampled values that lie in the $]t_i^{from} + 0.5, t_i^{from} + 1.5[$ interval, which has length 1 (the same argument can be applied to justify that we need to add 0.5 to t_i^{to}).

Thus, by treating t_i^{event} as a random variable, no data is discarded for the analysis. In the remainder of the paper, this model is called the full model. In addition, we can also compute the posterior distribution $p(t_i^{event} = t|D)$ for $t \in \{t_i^{from}, t_i^{from} + 1, ..., t_i^{to} - 1, t_i^{to}\}$ (where D stands for the dataset used to fit the model), to estimate the probability that a temporally-uncertain event has occurred in each of the dates within the associated temporal interval delimited by the Police.

The logistic regression models described in previous lines have been coded in the NIMBLE system for Bayesian inference (de Valpine et al. 2017), based on MCMC procedures. The technical details regarding the MCMC process that was implemented for the different models fitted are provided in "Results" section. The R codes developed for fitting the models described in the paper are available in the repository https://github.com/albrizre/AoristicLogistic.

Model Criticism

In Bayesian analysis, model assessment and comparison are typically performed through some well-known goodness-of-fit measures such as the Deviance Information Criterion (DIC) introduced by Spiegelhalter et al. (2002), or the Watanabe–Akaike Information Criterion (WAIC) proposed by Watanabe and Opper (2010). However, these metrics are only useful to compare models with the same likelihood function, so the complete cases model and the full model described above cannot be compared in terms of these metrics (the likelihood functions differ since each model is fitted to a different dataset).

Therefore, a different strategy is needed. One possibility is to perform model criticism through the analysis of the distribution of the point estimates of the π_i's, denoted by $\hat{\pi}_i$'s, each of which has been computed as the mean of the posterior distribution $p(\pi_i|D)$. Specifically, we first compare the distribution of the $\hat{\pi}_i$'s across models and location types (case vs. control). This enables us to appreciate if each model can discriminate between cases and controls and if we can find any remarkable difference. Second, considering the full model, it is examined if the distributions of the $\hat{\pi}_i$'s corresponding to certain and temporally-uncertain events differ.

In addition, it is also of interest to study the quality of the models as classification tools. For goodness-of-fit purposes, we can employ the in-sample predictive capability of a model. In this study, the F1 score and the Matthews correlation coefficient (MCC) have been chosen for evaluating the in-sample predictive quality of the models. Other well-known metrics, such as accuracy, have been discarded for the analysis because they are unreliable for imbalanced datasets such as ours (because the proportion of controls is much higher than the proportion of cases). Indeed, they tend to provide an over-optimistic estimation of the classifier ability on the majority class (Chicco and Jurman 2020), which in our case corresponds to the class of controls. First, the F1 score is defined as follows:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

where TP is the number of true positives (the model predicts a burglary and a burglary has actually occurred), FP is the number of false positives (the model predicts a burglary but no burglary has actually occurred), and FN the number of false negatives (the model does not predict a burglary but a burglary has actually occurred). The F1 score (Chinchor and Sundheim 1993) ranges from 0 (worst value) to 1 (best value, which indicates that the model classifies all the observations correctly).

On the other hand, the MCC (Matthews 1975) ranges from − 1 (worst value) to 1 (best value, which again indicates that the model classifies all the observations correctly) and is defined as follows:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}},$$

where TP, FP, and FN represent the same as in the definition of the F1 score, and TN is the number of true negatives (the model does not predict a burglary and no burglary has occurred).

In order to label an observation as a positive or a negative, a cutoff probability, $c \in [0, 1]$, needs to be used as a threshold. Then, if $\hat{\pi}_i > c$ the observation is classified as a positive (we predict a burglary), whereas if $\hat{\pi}_i \leq c$, the observation is classified as a negative (we predict no burglary). If we were working with a balanced dataset (with the same number of cases as controls), $c = 0.5$ would be the natural choice. However, the datasets under analysis are imbalanced in favor of controls (by construction, they present a 5:1 ratio of controls to cases). Hence, lower values of c are more suitable, otherwise, most of the observations will be classified as negatives and the classification would be far from optimal. For this reason, an analysis of the F1 score and the MCC as a function of c is performed.

The formula for both the F1 score and the MCC provides a point estimate of the predictive quality of the model. Following the approach proposed by Gilardi et al. (2022) for the analysis of the balanced accuracy of a model, we have also estimated the distribution of these two metrics. Specifically, the sampled values from the posterior distribution $p(\pi_i|D)$ can be used to simulate the F1 score and MCC values and therefore derive the distribution of the two, which allows a more complete comparison of the predictive ability of the models fitted.

Software

The R programming language (R Core Team 2021) has been used for the analysis. In particular, the R packages ggplot2 (Wickham 2016), lubridate (Grolemund and Wickham 2011), nimble (de Valpine et al. 2017), rgdal (Bivand et al. 2019), rgeos (Bivand and Rundel 2020), spatstat (Baddeley et al. 2015), and spdep (Bivand et al. 2008) have been used.

Results

A Simulation Study

Before proceeding to the analysis of the dataset of residential burglaries in Valencia, a simulation study is carried out to test the suitability of the proposed model under certain assumptions about the data. In addition, we compare the proposed model with other competing alternatives: the complete cases model (where temporally-uncertain events are discarded), the midpoint model (where each temporally-uncertain event is assigned to the midpoint date of its associated temporal window), and the random model (where each temporally-uncertain event is assigned to a random date within its associated temporal window). The objective is then twofold: to study whether the full model is able to deal with this kind of temporal uncertainty and recover the true baseline risk, and to compare the results provided by this model against those yielded by the complete cases model, the midpoint model, and the random model.

Simulation of Cases and Controls

In order to be able to simulate data that resemble the real burglary dataset from Valencia that has motivated the present study under varying levels of temporal uncertainty, we start by considering the set of controls that has also been used for the real data analysis and transform these records into a set of cases and controls through a data-generating process based on the logistic regression model. In particular, depending on the characteristics of record i (the values of the variables available for record i), we simulate a Bernoulli random variable, Y_i, where the probability of generating a crime event, π_i, obeys the following rule:

$$\begin{aligned} Y_i &\sim Ber(\pi_i) \\ \text{logit}(\pi_i) &= -3 + \beta_{DoW(i)} \end{aligned} \tag{2}$$

Therefore, according to the characteristics of record i, we simulate a Bernoulli random variable with probability π_i, which yields a 1 (case) with probability π_i and a 0 (control) with probability $1 - \pi_i$. Specifically, we assume that the probability of simulating a case or control only depends on the day of the week associated with record i. We need to specify some values for the β_{DoW} parameters in order to be able to simulate the data. If β_1 represents the effect for Mondays, β_2 for Tuesdays, and so on, we consider the following choice of values for the β_{DoW}'s: $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7\} = \{1, 1, 1, 1, 2, 2, 2\}$. This implies assigning twice as much weight to Fridays, Saturdays, and Sundays in the specification of the data-generating process that gives rise to cases and controls. In particular, under this choice, the probability that a record whose associated day of the week is between Monday and Thursday yields a crime event is $\frac{\exp(-3+1)}{1+\exp(-3+1)} \approx 0.12$, whereas the same probability for a record whose associated day of the week is between Friday and Sunday is $\frac{\exp(-3+2)}{1+\exp(-3+2)} \approx 0.27$ (we apply the inverse logit transformation to get these values). We note as well that because of how the cases and controls are generated, the spatio-temporal intensity of the cases is constant, that is, for the simulated datasets, crime risk does not depend on the temporal and/or spatial location within the study window, but only on the day within the week, according to (2).

Table 1 Summary of the p_{DoW} values (which denote the probability that an event that actually occurs on the day *DoW* of the week presents some temporal uncertainty) chosen for scenarios 0 to 4

Scenario	p_1	p_2	p_3	p_4	p_5	p_6	p_7
0	0	0	0	0	0	0	0
1	0.4	0.4	0.4	0.4	0.4	0.4	0.4
2	0.3	0.3	0.3	0.3	0.5	0.5	0.5
3	0.2	0.2	0.2	0.2	0.6	0.6	0.6
4	0.4	0.4	0.4	0.4	0.8	0.8	0.8

Addition of the Temporal Uncertainty

Then, the temporal uncertainty is introduced for a proportion of the generated cases following different mechanisms that allow us to reflect that the presence of interval-censored observations may depend on certain factors and that it might be of varying magnitude. Specifically, we simulate five datasets of increasing levels of temporal uncertainty in which the probability that a case presents temporal uncertainty also depends on the day of the week. Hence, if p_{DoW} denotes the probability that an event that actually occurs on day *DoW* of the week presents some temporal uncertainty (p_1 corresponds to such probability for Mondays, p_2 for Tuesdays, and so on), we assume five different scenarios for the p_{DoW}'s, which are summarized in Table 1. As stated above, from scenarios 0 to 4 we assume increasing levels of temporal uncertainty for the simulated cases. In particular, in scenario 0, we assume that there is no temporal uncertainty, so this scenario simply allows us to check that the data have been simulated correctly. In contrast, in scenario 1 we assume that there exists the same level of temporal uncertainty for all days within the week, whereas in scenarios 2, 3, and 4 we assume that the probability that an event presents temporal uncertainty is greater if the event has actually occurred from Friday to Sunday. We note that in scenarios 2 to 4 we assume conditions of increasing complexity in terms of the variability of the p_{DoW}'s within the week, but also in terms of the average probability that a case presents temporal uncertainty.

Thus, in a second step, we implement the following procedure to generate the temporally-uncertain events and the temporal windows associated with these events:

- Given a case, i, of the simulated dataset, we sample from a Bernoulli random variable with probability $p_{DoW(i)}$. If the simulated value is 0, no temporal uncertainty is introduced for the corresponding case. If the simulated value is 1, we continue as follows.
- Assuming that the simulated value in the previous step is 1, the level of temporal uncertainty in days for case i is obtained by sampling from an Exponential distribution of rate $\lambda = 0.2$, and by adding 1 to this sampled value (to ensure that the temporal uncertainty is at least of 1 day, otherwise there would be no uncertainty). Specifically, if ℓ_i denotes the value simulated from the $Exp(\lambda = 0.2)$ distribution for case i rounded to the nearest integer, we consider that the width of the associated time window is $2(\ell_i + 1)$.
- Finally, if t_i represents the actual temporal location for case i, we compute $t_i^{from} = t_i - \omega_i$ and $t_i^{to} = t_i + (2(\ell_i + 1) - \omega_i)$, where ω_i is an integer number sampled from the set $\{0, ..., 2(\ell_i + 1)\}$.

Thus, by following this procedure, we create a temporal window of width $2(\ell_i + 1)$ around t_i which is not necessarily symmetrical with respect to t_i, as it happens with the real data.

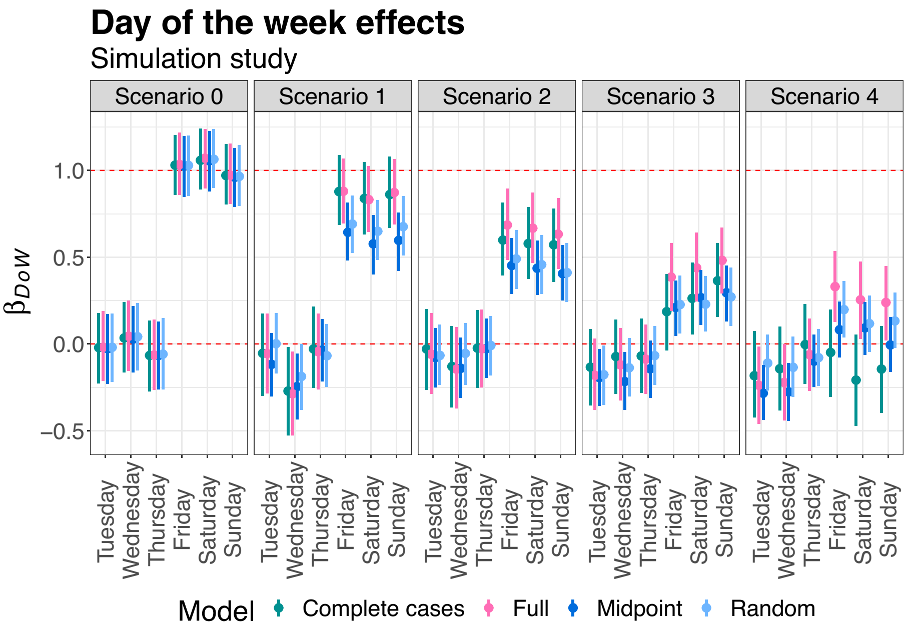


Fig. 2 Day of the week effect estimates yielded by the complete cases model, the full model, and the midpoint model under the different scenarios studied. The dashed red lines located at $\beta_{DoW} = 0$ and $\beta_{DoW} = 1$ correspond to the true underlying parameter values that should be recovered by the model ($\beta_{DoW} = 0$ for Monday, Tuesday, Wednesday, and Thursday; $\beta_{DoW} = 1$ for Friday, Saturday, and Sunday)

We also note that since the expected value of an $Exp(\lambda)$ distribution is $\frac{1}{\lambda}$, the average width (in days) of our simulated time windows is $\frac{1}{0.2} + 1 = 6$.

Model Comparison

We fit the proposed full model, the complete cases model, the midpoint model, and the random model to the five simulated datasets constructed as previously described. Specifically, for each of the models and datasets, the MCMC procedure consisted of 2 chains of length 40,000, with a burn-in period of length 20,000, which resulted in 2000 sampled values per chain after applying a thinning of 10 (which means keeping every 10th value of the chain).

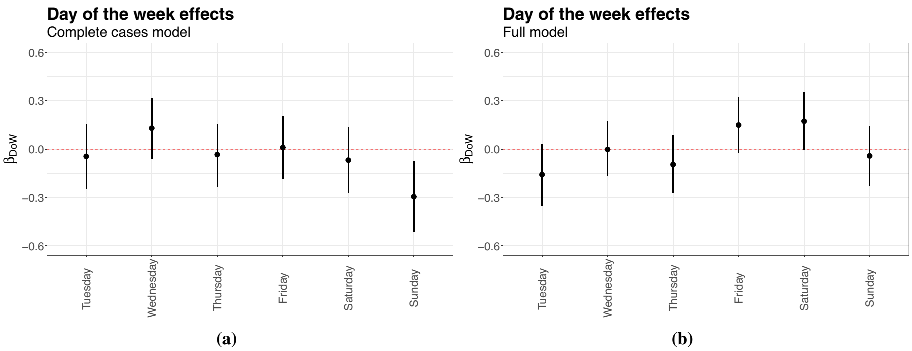
The comparison focuses on studying to what extent these models are able to recover the true values of the β_{DoW} parameters used for the data-generating process. Hence, in all cases, we specify the logistic model as in (1), considering Mondays as the reference value for the days of the week. Therefore, since $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7\} = \{1, 1, 1, 1, 2, 2, 2\}$ has been chosen for the data-generating process, we would expect the estimates of β_{DoW} to be around 0 for Tuesday, Wednesday, and Thursday, and around 1 for Friday, Saturday, and Sunday. Figure 2 shows the day of the week effects estimated with the four models under comparison, considering the simulated datasets associated with scenarios 0 to 4. First, we notice that all models perform well under scenario 0 (the point estimates are very close to $\beta_{DoW} = 0$ and $\beta_{DoW} = 1$, depending on the day of the week), as expected since the simulated dataset corresponding to this scenario presents no temporal uncertainty. Then, from scenarios 1 to 4, we observe that the estimates of β_{DoW} for Friday, Saturday, and Sunday progressively decrease for the four competing models, even though some differences across models arise. In general, the proposed full model is the one that provides the estimates of β_{DoW} for these days that stay closer to the true value $\beta_{DoW} = 1$. In addition to this, the complete cases model performs better than the midpoint model and the random model in scenarios 2 and 3. In contrast, under scenario 4, which represents the most extreme one among the scenarios studied, the complete cases model provides worse estimates than these two imputation-based models. We also note that even though the estimates yielded by the full model become lower under a higher level of temporal uncertainty, the credible intervals obtained for Friday, Saturday, and Sunday do not contain 0 even for scenario 4, which would allow us to infer a positive association with a greater probability of a crime event for these days of the week. The midpoint model and the random model, however, present credible intervals under scenario 4 that contain 0, which might lead us to the erroneous conclusion that these days are unrelated to crime risk. Regarding the width of the credible intervals obtained, we can appreciate that having a greater proportion of interval-censored observations causes the credible intervals to widen slightly in the case of the full model. At the same time, the complete cases model tends to present a wider credible interval, which is a consequence of the fact that it discards a large portion of the available data, whereas the midpoint model and midpoint model usually yield a narrower credible interval. This is also something that we expected because these methods do not discard any of the observations and also because the imputation procedure eliminates the uncertainty surrounding the interval-censored observations, with the disadvantage of introducing a greater level of bias in the estimation.

In conclusion, the results obtained from this simulation study enable us to conclude that the full model performs better than other competing models for dealing with interval-censored crime events, while the complete cases model could outperform some imputation-based models under moderate levels of temporal uncertainty. Despite this, the estimates provided by the full model under the presence of very high levels of temporal uncertainty are less reliable and might be biased. Anyhow, it is also worth noting that the simulation

Table 2 Parameter estimates (Est.) yielded by the two models considered, along with the upper (Up.) and lower (Lo.) bound of the 95% credible intervals associated with each parameter

Complete cases model				Full model			
Parameter	Est.	Lo.	Up.	Parameter	Est.	Lo.	Up.
α	− 2.18	− 2.33	− 2.02	α	− 1.74	− 1.88	− 1.60
$\beta_{Tuesday}$	− 0.04	− 0.25	0.15	$\beta_{Tuesday}$	− 0.16	− 0.35	0.03
$\beta_{Wednesday}$	0.13	− 0.06	0.31	$\beta_{Wednesday}$	− 0.00	− 0.17	0.17
$\beta_{Thursday}$	− 0.03	− 0.24	0.16	$\beta_{Thursday}$	− 0.09	− 0.27	0.09
β_{Friday}	0.01	− 0.19	0.21	β_{Friday}	0.15	− 0.02	0.33
$\beta_{Saturday}$	− 0.07	− 0.27	0.14	$\beta_{Saturday}$	0.17	− 0.01	0.35
β_{Sunday}	− 0.29	− 0.51	− 0.08	β_{Sunday}	− 0.04	− 0.23	0.14
τ_{δ}	1365.67	383.35	3021.07	τ_{δ}	375.73	127.99	909.31
τ_{ε}	28.06	12.07	69.79	τ_{ε}	54.19	16.50	222.21
τ_{u}	2.45	1.24	4.78	τ_{u}	2.11	1.16	3.55
τ_{v}	71.92	8.50	372.65	τ_{v}	185.21	14.55	1231.34

Each τ parameter corresponds to the precision of the corresponding random effect, being defined as the inverse of its variance. For instance, $\tau_{\delta} = 1/\sigma_{\delta}^{2}$ is the precision of the temporally-structured randomeffect



(a)

(b)

Fig. 3 Day of the week effect estimates yielded by the complete cases model (**a**) and the full model (**b**). The dashed red line is located at the value $\beta_{DoW} = 0$, which represents the absence of an effect (Color figure online)

study carried out is far from exhaustive. Adapting the simulation study to the characteristics of a dataset of our interest, as we have done here, seems an advisable strategy.

Analysis of the Valencia Burglary Dataset

Complete Cases Analysis Versus Full Analysis

A major objective of the article is to compare the results derived from using the complete cases model (discarding temporally-uncertain events) and the full model, considering the burglary dataset recorded in Valencia during the years 2016 and 2017. Specifically, the goal is to check whether the complete cases model gives rise to biased estimates of the parameters and whether the full model can capture the real temporal distribution of the events. We will also compare the results provided by the full model with some exploratory data analyses based on the aoristic approach. As in the case of the models employed for conducting the simulation study, the MCMC procedure for the two models consisted of 2 chains of length 40,000, with a burn-in period of length 20,000, and a thinning of 10. The convergence analysis was performed through the visual inspection of the density plots and the trace plots associated with the parameters of the model. The Supplementary Material of the paper contains these plots for the main parameters of the models (Supplementary Figures 1–4).

Thus, Table 2 summarizes the results in terms of the point and interval estimates of the parameters involved in the two models, allowing direct comparison. First, we note that the estimate of α is smaller for the complete cases model. This is a consequence of the fact that the dataset considered for fitting this model has a smaller proportion of cases than the one used for the full model, which makes the estimate of the baseline probability of event occurrence also lower. In any case, the α parameter has no major relevance in terms of interpretation.

It is of greater interest to analyze the effects of the days of the week, as represented by the β_{DoW} parameters (which represent the variation in risk in comparison to Mondays, the reference level). In this case, the differences are notable and of great relevance from a practical point of view. As shown in Fig. 3, according to the complete cases model, crime risk is notably lower on Sundays, whereas the highest estimate of crime risk corresponds

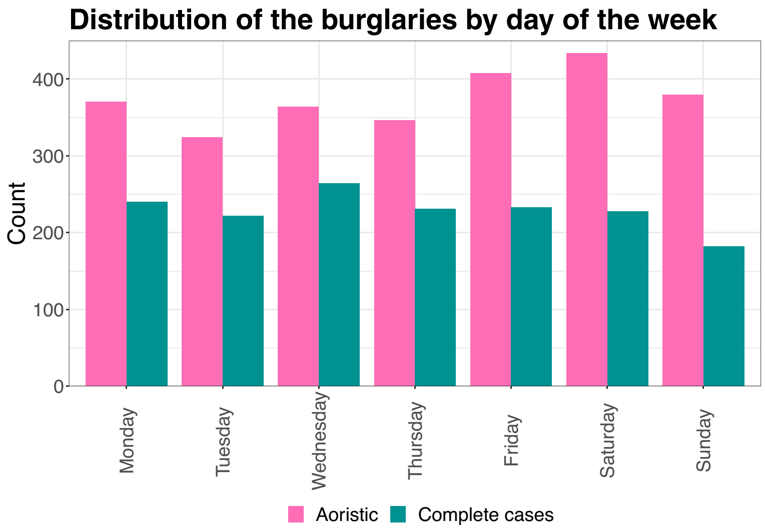


Fig. 4 Distribution of the burglaries by day of the week, considering only the events for which the exact date is known (complete cases analysis), and the whole available dataset by performing an aoristic analysis of the temporally-uncertain events

to Wednesdays. In contrast, the full model yields Fridays and Saturdays as the two days with high burglary risk (even though the lower bound of the 95% credible interval associated with β_{Friday} is slightly below 0), while Sundays do not display low burglary risk. Therefore, the results differ markedly depending on the model considered. In fact, the differences between models can be understood if an aoristic analysis of the distribution of residential burglaries by day of the week is performed, as shown in Fig. 4. It can be observed how the fact of considering temporally-uncertain events in the analysis causes the proportion of crime to reach the highest values on the weekend, especially on Saturdays. In other words, this suggests that there is a higher proportion of temporally-uncertain events that cover (partially or totally) the weekend, possibly because on these days of the week part of the population stays in a second residence, or simply because of the changes in daily routines during the weekend, which could facilitate the action of burglars in certain time slots. This type of plausible assumption could be studied by considering an hour-level analysis, although it could also complicate the model estimation. In any case, the complete cases model entirely misses this type of information, which is recovered by the full model, despite the presence of interval-censored data. In the Supplementary Material of the paper (Supplementary Figures 5–6) we also provide a correlation analysis of the MCMC samples associated with the posterior distributions of the β_{DoW} parameters. In general, we find moderate positive associations, which suggests that the day of the week effects are correlated between them and hence that considering a multivariate prior distribution for these effects could be beneficial for increasing the convergence speed of the MCMC process.

Once the only fixed effect of the model (day of the week) has been analyzed, the estimates of the temporal and spatial random effects given by both models are compared. First, Fig. 5 shows the estimates of the temporally-structured random effect, δ_w. Although the overall behavior of the temporal trend captured by this effect is similar for both models, certain differences arise. For instance, the complete cases model determines a peak in burglary risk at the beginning of the study period, which is not determined by the full

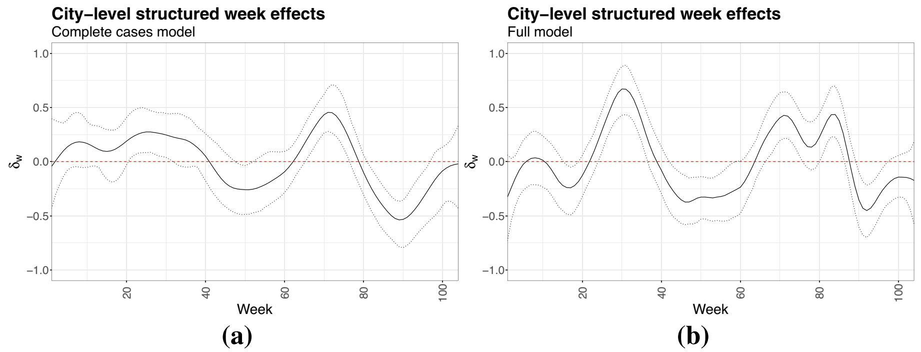


Fig. 5 Week effect estimates (structured component) yielded by the complete cases model (**a**) and the full model (**b**). The dotted lines delimit the 95% credible interval associated with each point estimate. The dashed red line is located at the value $\delta_w = 0$, which represents the absence of an effect (Color figure online)

model. Besides, the full model detects a double peak in crime risk around weeks 70 to 90, whereas the complete cases model locates a single peak within that period. Finally, the peak detected around week 30 by both models is notably higher in the estimation provided by the full model. Indeed, the full model can detect more variability at the week level than the complete cases model. This can be guessed from Fig. 5, but also by comparing the estimates of the precision of the random effect δ_w. As shown in Table 2, the estimate of τ_δ is notably smaller in the case of the full model, which confirms that the random effect δ_w captures more variability in the latter model, since $\tau_\delta = 1/\sigma^2_\delta$.

The aoristic analysis of the distribution of residential burglaries by week allows us to verify, once again, that the full model adequately captures the temporal distribution of burglaries. Thus, as shown in Fig. 6, the presence of temporally-uncertain events is notably higher in the summer months (July and August), when most residents enjoy holiday periods, increasing the likelihood that homes will be empty for days or even weeks. Besides, the aoristic analysis also reveals a peak in burglary counts during May 2017, which can be assumed to be the consequence of a separate process from the one corresponding to the summer peak. This peak in May 2017 is the only one detected by the complete cases model, since the proportion of temporally-uncertain events is quite low during this period, as shown in Fig. 6. Paradoxically, the complete cases model indicates that the summer period of 2017 is a low-risk period, but this is only the consequence of data underrepresentation because of the presence of many temporally-uncertain events within these summer months.

Finally, Fig. 7 enables us to compare the spatial random effects estimates resulting from both models. In this case, the differences are slight, being more noticeable in some neighbors located around the city center. In other words, these results suggest that the presence of temporally-uncertain events does not follow a markedly spatial pattern for the dataset analyzed. Indeed, as shown in Fig. 8a, the scatter plot of the spatial effect estimates yielded by the two models indicates that these are highly correlated (the value of the Pearson correlation coefficient is 0.93 in this case). Nevertheless, as shown in Fig. 8b, accounting for the existence of interval-censored observations makes that the spatial effect estimates corresponding to certain boroughs vary considerably in terms of their position within the set of

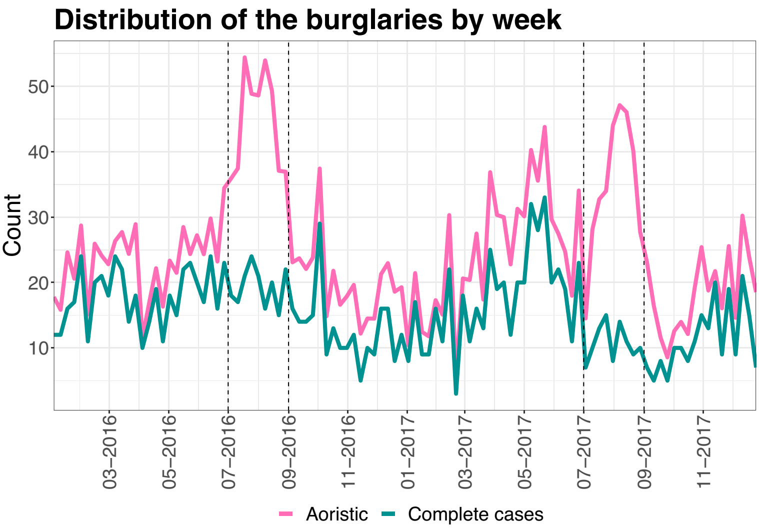


Fig. 6 Distribution of the burglaries by week, considering only the events for which the exact date is known (complete cases analysis), and the whole available dataset by performing an aoristic analysis of the temporally-uncertain events. The dashed lines delimit the summer months (July and August) of both years

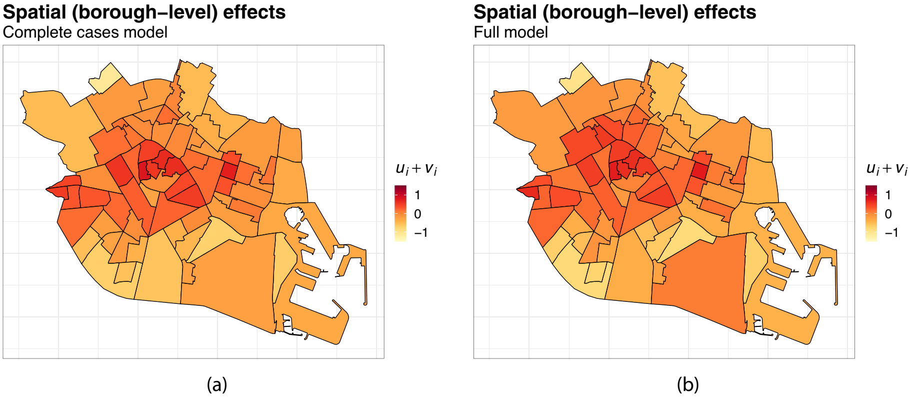


Fig. 7 Choropleth maps of the spatial (borough-level) effect estimates (computed as $u_i + v_i$) yielded by the complete cases model (**a**) and the full model (**b**)

estimates. This has potential implications in practice if one is interested in prioritizing the establishment of preventive measures in the boroughs that present the greatest estimates.

Event Time Imputation

The main advantage of the proposed full model is that it allows the inclusion of temporally-uncertain events in the analysis, following the aoristic approach. In this way, it avoids reducing the sample size (as occurs in the complete cases model) and prevents the potential error involved in imputing the time of the event. In addition, another advantage of the model is that

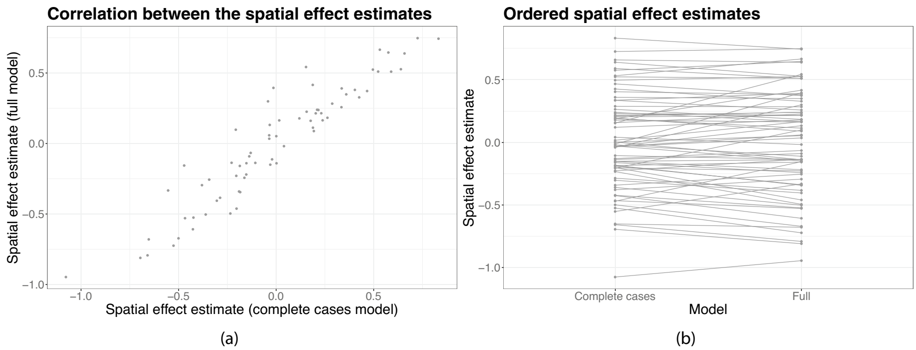


Fig. 8 Scatter plot of the spatial effect estimates (computed as $u_i + v_i$) yielded by the complete cases model against those of the full model for the 70 boroughs within the study window (**a**), and spaghetti plot obtained from the same sets of estimates, where the estimates corresponding to each particular borough are connected by a segment (**b**)

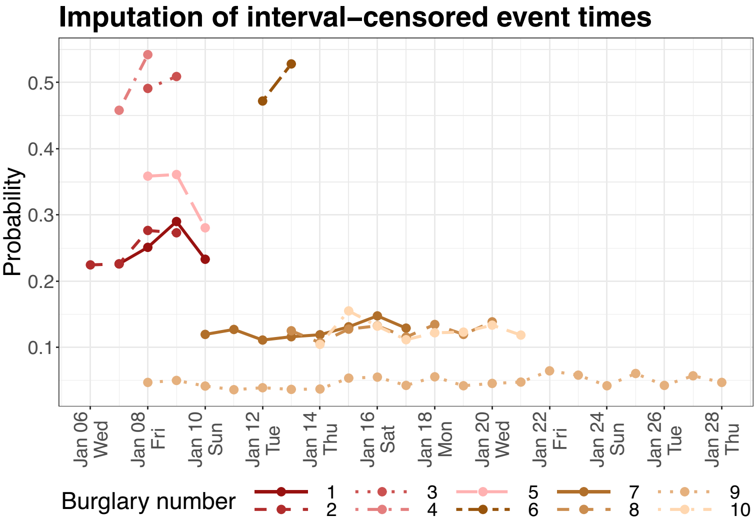
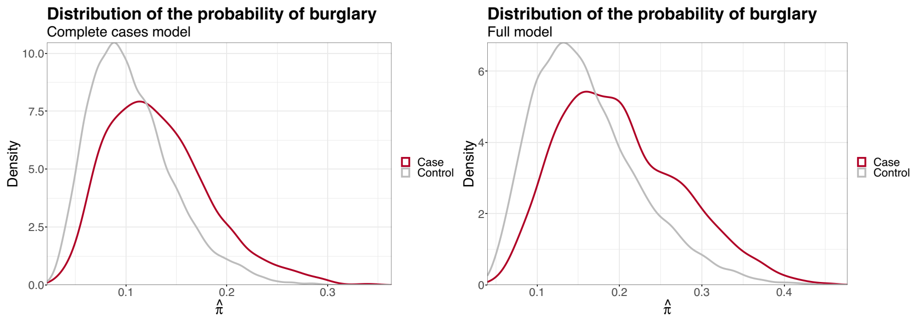


Fig. 9 Estimates of $p(t_i^{event} = t|D)$ for a selection of 10 temporally-uncertain burglaries occurred in Valencia from 6 January 2016 to 28 January 2016

it makes it possible to perform the imputation of event times, based on the posterior probability of each time unit (in our case, days) contained in the intervals that delimit the uncertainty existing for each event. Hence, Fig. 9 shows the values of $p(t_i^{event} = t|D)$ corresponding to a set of temporally-uncertain burglaries that occurred from 6 January 2016 to 28 January 2016. In each case, t varies from t_i^{from} to t_i^{to}, therefore some probability of occurrence is assigned to each time unit contained in the interval. The values of $p(t_i^{event} = t|D)$ are based on all the information contained in the model, in terms of the fixed and random effects considered. In particular, the temporal uncertainty is connected to the day of the week effect and to the temporal random effect (the spatial random effect, on the other hand, is not influenced by the



(a) **(b)**

Fig. 10 Distribution of the $\hat{\pi}_i$'s corresponding to cases (events) and controls for the complete cases model (**a**) and the full model (**b**)

temporal uncertainty). The values of $p(t_i^{event} = t|D)$ tend to be higher for the days of the week associated with higher risk, which are Friday and Saturday. This is clearly illustrated by some of the events shown in Fig. 9, which exhibit a temporal uncertainty of 2, 3, or 4 days, including all or part of a weekend in the uncertainty window. For other events for which there is more uncertainty (the time window is wider), this effect is not as clear.

Therefore, the modeling approach described could be used as an imputation technique too, by simply considering the value arg $\max_t p(t_i^{event} = t|D)$. It would be necessary to have a dataset in which the events have temporal uncertainty according to Police records and, at the same time, the exact temporal location of these events through an external source of information, to assess the quality of this imputation method.

Model Assessment

The first step to assess the quality of the models has been analyzing the distribution of the $\hat{\pi}_i$'s, which are computed as the mean of the posterior distribution $p(\pi_i|D)$. We recall that π_i represents the probability that observation i is a burglary according to the spatio-temporal characteristics of this observation and the estimated parameters of the model. Thus, Fig. 10 shows the distribution of the $\hat{\pi}_i$'s for both cases and controls, considering the complete cases (Fig. 10a) and the full model (Fig. 10b). It can be observed that both models can distinguish between cases and controls adequately. Indeed, while it is true that the two distributions overlap substantially, the average posterior probability of being a case is considerably greater among the cases rather than among the controls. At the same time, we note that the high level of overlap observed suggests that the ability of the model to classify observations into cases and controls is far from optimal, which suggests that including additional covariate or random effects would be convenient. Anyhow, predicting the occurrence or not of a crime event within a small spatio-temporal location is a really challenging task, so these results are not surprising. It can be also appreciated that the distribution of the $\hat{\pi}_i$'s covers larger values (for both cases and controls) in the case of the full model as a consequence of the greater proportion of cases in this model. This is something that has already been discussed when comparing the α parameters of the models and will be of importance again when evaluating the classification ability of the models, as will be shown later.

Furthermore, Fig. 11 compares the distribution of the $\hat{\pi}_i$'s estimated through the full model for temporally-uncertain and certain (at the date level) events. The distribution of

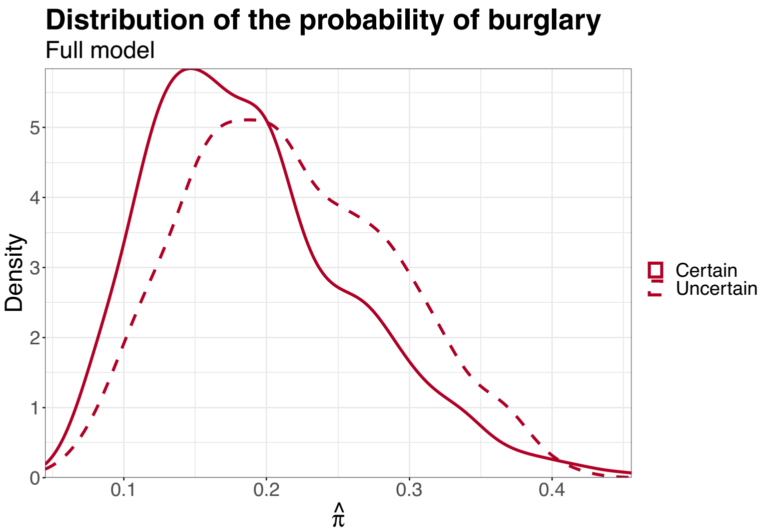


Fig. 11 Distribution of the $\hat{\pi}_i$'s corresponding to temporally-uncertain and certain (exact date known) events for the full model

the $\hat{\pi}_i$'s corresponding to temporally-uncertain events presents a better behavior, in the sense that this distribution is more displaced towards larger estimates of π_i. This analysis allows us to verify that the temporally-uncertain observations have been adequately included in the model since they do not perform worse from the perspective of model fit than those that are not (in fact, they seem to perform better).

Regarding the classification ability of the model, Fig. 12 shows the distribution of the MCC, derived from the sampled values of the posterior distribution $p(\pi_i|D)$. The cutoff probability, c, is varied from 0.05 to 0.35, in steps of size 0.05. Higher values of c are discarded because the number of positive predictions becomes too low. This is a consequence of the fact that the dataset is imbalanced with few case observations (in comparison to the number of control observations), which causes model predictions to be less than 0.5. This is not an issue, we simply have to consider values of c lower than 0.5 (which would be the

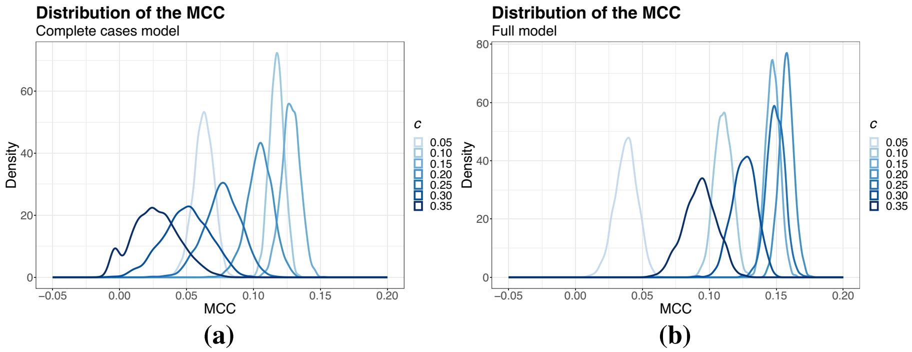
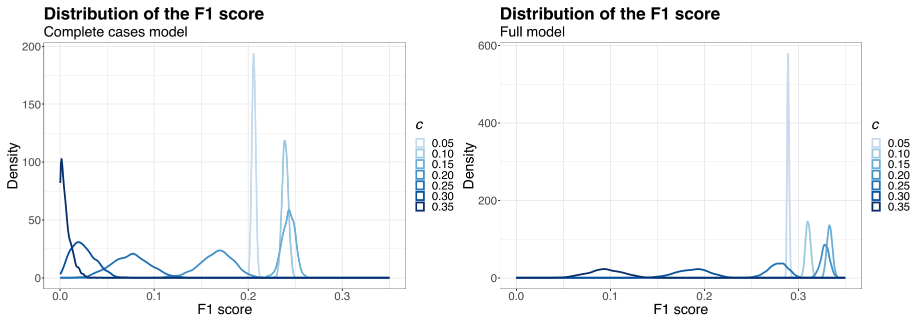


Fig. 12 Distribution of the MCC derived from the sampled values of the posterior distribution $p(\pi_i|D)$, considering the complete cases model (**a**) and the full model (**b**). Several values of the cutoff probability, c, are tested and compared



(a) **(b)**

Fig. 13 Distribution of the F1 score derived from the sampled values of the posterior distribution $p(\pi_i|D)$, considering the complete cases model (**a**) and the full model (**b**). Several values of the cutoff probability, c, are tested and compared

common threshold for a balanced dataset). Figure 12 provides us with two conclusions of interest. First, the optimal values of c are 0.15 and 0.20 for the complete cases and the full model, respectively (testing a finer partition of c values would allow us to more accurately approximate the optimal value of c in each case). The fact that the optimal value of c is larger in the case of the full model is something that we might already expect since the proportion of cases is larger for the full dataset (this has been already discussed given the estimates of the α parameter for each of the models). Second, and more importantly, Fig. 12 enables us to appreciate that MCC values tend to be higher in the case of the full model. Specifically, considering the optimal c values, the MCC ranges from 0.114 to 0.141 (with 95% credibility) in the case of the complete cases model (for $c = 0.15$), whereas it ranges from 0.147 to 0.168 (with 95% credibility) in the case of the full model (for $c = 0.20$). For these choices of c, the resulting confusion matrices are $\begin{pmatrix} \text{TP} & \text{FP} \\ \text{FN} & \text{TN} \end{pmatrix} = \begin{pmatrix} 525 & 1910 \\ 1074 & 11141 \end{pmatrix}$ and $\begin{pmatrix} \text{TP} & \text{FP} \\ \text{FN} & \text{TN} \end{pmatrix} = \begin{pmatrix} 1189 & 3149 \\ 1435 & 9902 \end{pmatrix}$ for the complete cases and the full model, respectively. Similarly, the full model also performs better in terms of the F1 score, as shown in Fig. 13. Specifically, the values of the F1 score are optimal for $c = 0.15$, regardless of the model chosen. In the case of the complete cases model, the F1 score ranges from 0.228 to 0.255 (with 95% credibility), while it ranges from 0.328 to 0.339 (with 95% credibility) in the case of the full model.

Discussion and Conclusions

In this paper, a logistic regression model has been proposed for the analysis of crime data in the presence of temporally-uncertain observations, which are abundant for certain types of crimes. The aoristic method, which allows exploratory analysis of the data in this context, has been taken into account to incorporate such temporal uncertainty into the model. This is a natural approach considering the Bayesian treatment of missing data. The model implemented has allowed us to see how discarding temporally-uncertain observations in the analysis can lead to erroneous conclusions. Although this kind of modeling approach for dealing with interval-censored event observations has already been proposed in the literature (Reich and Porter 2015), this article has the novelty, to the best of the author's knowledge, of following the aoristic approach in a modeling context, while performing a complete comparison of the model proposed with the complete cases counterpart, which would be a typical choice.

There is still room for improvement in the model proposed. Indeed, the model could be enhanced by adding covariate information and interaction terms. For instance, a spatio-temporal interaction random effect or the interaction between the day of the week and the week within the year could be considered. The addition of such terms might lead to a more informative model, so the imputation of event times based on the posterior distribution of t_i^{event} could be more realistic. As a drawback, increasing the complexity of the model by the inclusion of these terms could complicate model estimation, especially if the size of the available crime dataset is not very large. In addition, instead of using the logistic regression model, we could adapt the ideas presented in this paper for dealing with the temporal uncertainty of the events to other modeling frameworks, such as point process models (Mohler et al. 2011; Shirota and Gelfand 2017). This kind of model would allow us to provide more accurate estimates of the risk of crime over space, or to account for temporal self-exciting effects, which are typically observed in crime data. In addition to this, point process models could also allow us to recompute area-level time-varying crime risk estimates by accounting for the existence of interval-censored observations. Essentially, we could employ an approach similar to the one proposed in the paper, in which the temporal uncertainty around each interval-censored observation would be modeled, and then area-level crime risk estimates would be obtained by averaging the estimates of the spatio-temporal intensity of the process on a regular grid that covers the study window.

Another important aspect is that when discussing the results provided by the complete cases and the full model, it has been implicitly assumed that the aoristic analysis of the data gives a true picture of the temporal distribution of the data. However, as pointed out by Mulder (2019), the aoristic analysis might tend to overdisperse the temporal distribution of the events. Thus, by assigning the same weight to each temporal unit within the observation window, we might be assuming too much uncertainty (variability), even though it results in a natural approach if we have no prior knowledge about the true temporal location of the event. The proposed model allows the inclusion of prior information about the events, in the standard way used in Bayesian inference. For instance, a specific prior distribution could be assigned to those events for which there might be some intuition (by the Police or the property owners themselves) about their actual temporal location, or some non-uniform distribution that might be closer to reality could be tested. In fact, future studies could make use of a truncated Normal distribution with the mean located at the mid-point location, or even at a location closer to the start (or end) of the interval. This could reveal, in some cases, whether events tend to cluster temporally in the initial part of the time window, which could be explained in case the burglars have been watching the owners and took advantage of their departure from the home.

Finally, to better assess the potential of the full model for predicting the true temporal location of the temporally-uncertain events, a dataset including both the interval-censored temporal locations recorded by the Police (according to the information provided by the owners or other residents) and the actual (exact) temporal locations derived from other sources would be required. For instance, in the case study conducted by Ashby and Bowers (2013), closed-circuit television camera images were used to determine the exact temporal locations of the events under study. Unfortunately, this kind of dataset is really scarce.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10940-023-09580-1.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

Conflict of interest The author declares no conflicts of interest.

References

Andresen MA, Malleson N, Steenbeek W, Townsley M, Vandeviver C (2020) Minimum geocoding match rates: an international study of the impact of data and areal unit sizes. Int J Geogr Inf Sci 34(7):1306–1322

Ashby MP, Bowers KJ (2013) A comparison of methods for temporal analysis of aoristic crime. Crime Sci 2(1):1–16

Baddeley A, Rubak E, Turner R (2015) Spatial point patterns: methodology and applications with R. CRC Press, Boca Raton

Besag J, York J, Mollié A (1991) Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43(1):1–20

Bivand R, Keitt T, Rowlingson B (2019) rgdal: bindings for the 'geospatial' data abstraction library. R package version 1.4-6

Bivand R, Rundel C (2020) rgeos: interface to geometry engine—open source ('GEOS'). R package version 0.5-3

Bivand RS, Pebesma EJ, Gomez-Rubio V, Pebesma EJ (2008) Applied spatial data analysis with R, vol 747248717. Springer, New York

Briz-Redón Á, Martínez-Ruiz F, Montes F (2022a) Adjusting the Knox test by accounting for spatio-temporal crime risk heterogeneity to analyse near-repeats. Eur J Criminol 19(4):586–611

Briz-Redón Á, Mateu J, Montes F (2022b) Identifying crime generators and spatially overlapping high-risk areas through a nonlinear model: a comparison between three cities of the Valencian region (Spain). Stat Neerl 76(1):97–120

Briz-Redón A, Martinez-Ruiz F, Montes F (2020) Reestimating a minimum acceptable geocoding hit rate for conducting a spatial analysis. Int J Geogr Inf Sci 34(7):1283–1305

Buil-Gil D, Medina J, Shlomo N (2021) Measuring the dark figure of crime in geographic areas: small area estimation from the crime survey for england and wales. Br J Criminol 61(2):364–388

Buil-Gil D, Moretti A, Langton SH (2022) The accuracy of crime statistics: assessing the impact of police data bias on geographic crime analysis. J Exp Criminol 18:515–541

Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom 21(1):6

Chinchor N, and Sundheim BM (1993) MUC-5 evaluation metrics. In: Fifth message understanding conference (MUC-5): proceedings of a conference held in Baltimore, Maryland, August 25–27, 1993

Chung J, Kim H (2019) Crime risk maps: a multivariate spatial analysis of crime data. Geogr Anal 51(4):475–499

de Valpine P, Turek D, Paciorek CJ, Anderson-Bergman C, Lang DT, Bodik R (2017) Programming with models: writing statistical algorithms for general model structures with NIMBLE. J Comput Graph Stat 26(2):403–413

Gail M, Williams R, Byar DP, Brown C et al (1976) How many controls? J Chronic Dis 29(11):723–731

Gilardi A, Mateu J, Borgoni R, Lovelace R (2022) Multivariate hierarchical analysis of car crashes data considering a spatial network lattice. J R Stat Soc Ser A 185(3):1150–1177

Grolemund G, Wickham H (2011) Dates and times made easy with lubridate. J Stat Softw 40:1–25

Hong E-P, Park J-W (2012) Sample size and statistical power calculation in genetic association studies. Genom Inform 10(2):117–122

Law J, Quick M, Chan P (2014) Bayesian spatio-temporal modeling for analysing local patterns of crime over time at the small-area level. J Quant Criminol 30(1):57–78

Li G, Haining R, Richardson S, Best N (2014) Space–time variability in burglary risk: a Bayesian spatio-temporal modelling approach. Spatial Stat 9:180–191

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta (BBA) Protein Struct 405(2):442–451

Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE (2011) Self-exciting point process modeling of crime. J Am Stat Assoc 106(493):100–108

Mulder KT (2019) Bayesian Circular Statistics: von Mises-based solutions for practical problems. PhD thesis, Utrecht University

Quick M, Li G, Brunton-Smith I (2018) Crime-general and crime-specific spatial patterns: a multivariate spatial analysis of four crime types at the small-area scale. J Crim Just 58:22–32

R Core Team (2021) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Ratcliffe JH (2000) Aoristic analysis: the spatial interpretation of unspecific temporal events. Int J Geogr Inf Sci 14(7):669–679

Ratcliffe JH (2002) Aoristic signatures and the spatio-temporal analysis of high volume crime patterns. J Quant Criminol 18(1):23–43

Ratcliffe JH (2004) Geocoding crime and a first estimate of a minimum acceptable hit rate. Int J Geogr Inf Sci 18(1):61–72

Ratcliffe JH, McCullagh MJ (1998) Aoristic crime analysis. Int J Geogr Inf Sci 12(7):751–764

Reich BJ, Porter MD (2015) Partially supervised spatiotemporal clustering for burglary crime series identification. J R Stat Soc A Stat Soc 178(2):465–480

Shirota S, Gelfand AE (2017) Space and circular time log Gaussian Cox processes with application to crime event data. Ann Appl Stat 11(2):481–503

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc Ser B (Stat Method) 64(4):583–639

Watanabe S, Opper M (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 11(12):3571–3594

Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York

Zhuang J, Mateu J (2019) A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. J R Stat Soc A Stat Soc 182(3):919–942