The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3088–3111
https://doi.org/10.1214/23-AOAS1752

DATA-DRIVEN CHIMNEY FIRE RISK PREDICTION USING MACHINE LEARNING AND POINT PROCESS TOOLS

BY CHANGQING LU[1,a], MARIE-COLETTE VAN LIESHOUT[2,b], MAURITS DE GRAAF[3,c] AND PAUL VISSCHER[4,d]

[1]*Department of Applied Mathematics, University of Twente,* [a]*c.lu@utwente.nl*
[2]*Stochastics, Centrum Wiskunde & Informatica,* [b]*marie-colette.van.lieshout@cwi.nl*
[3]*Innovation Research & Technology, Thales Nederland B.V.,* [c]*maurits.de.graaf@cwi.nl*
[4]*Sector Strategy & Support, Brandweer Twente,* [d]*p.visscher@brandweertwente.nl*

Chimney fires constitute one of the most commonly occurring fire types. Precise prediction and prompt prevention are crucial in reducing the harm they cause. In this paper we develop a combined machine learning and statistical modelling process to predict fire risk. First, we use random forests and permutation importance techniques to identify the most informative explanatory variables. Second, we design a Poisson point process model and employ logistic regression estimation to estimate the parameters. Moreover, we validate the Poisson model assumption using second-order summary statistics and residuals. We implement the modelling process on data collected by the Twente Fire Brigade and obtain plausible predictions. Compared to similar studies, our approach has two advantages: (i) with random forests, we can select explanatory variables nonparametrically considering variable dependence; (ii) using logistic regression estimation, we can fit our statistical model efficiently by tuning it to focus on regions and times that are salient for fire risk.

1. Introduction. During the last decade, the Dutch fire and rescue services have been developing an interest in applying business intelligence to improve their strategy of fire prediction and prevention (NVBR (2010)). To prepare for risk reducing measures, such as essential public awareness campaigns and proper fire staffing and equipment arrangements, accurate predictions are required. In this study we focus on chimney fires, as they occur frequently, rely heavily on environmental factors and impact people's daily life. Collaborating with the Twente Fire Brigade, we conduct a complete risk prediction study for chimney fires. We define fire risk prediction as an occurrence modelling problem, analyze underlying patterns and design appropriate prediction models. Our approach for chimney fire prediction is general and can be transferred to similar fire types, such as kitchen fires.

The literature for fire risk prediction is mostly concerned with wildfires. Overall, the prediction approaches can be divided into two categories: machine learning based approaches (e.g., Rodrigues and de la Riva (2014), Jain et al. (2020), Malik et al. (2021)) and statistical approaches (e.g., Turner (2009), Møller and Díaz-Avalos (2010), Xu and Schoenberg (2011)). Usually, machine learning based approaches do not require prior knowledge but can detect the dependence between fire risk and a large number of environmental variables automatically using specialised learning algorithms, such as logistic regression (Preisler et al. (2004)), support vector machine (Sakr et al. (2010)), decision tree (Stojanova et al. (2012)), random forest (Rodrigues and de la Riva (2014)) and neural network (Satir, Berberoglu and Donmez (2016)). Most machine learning algorithms are applicable for discrete data, whereas hazard maps of fire risk are continuous. Moreover, fire occurrences are usually recorded as

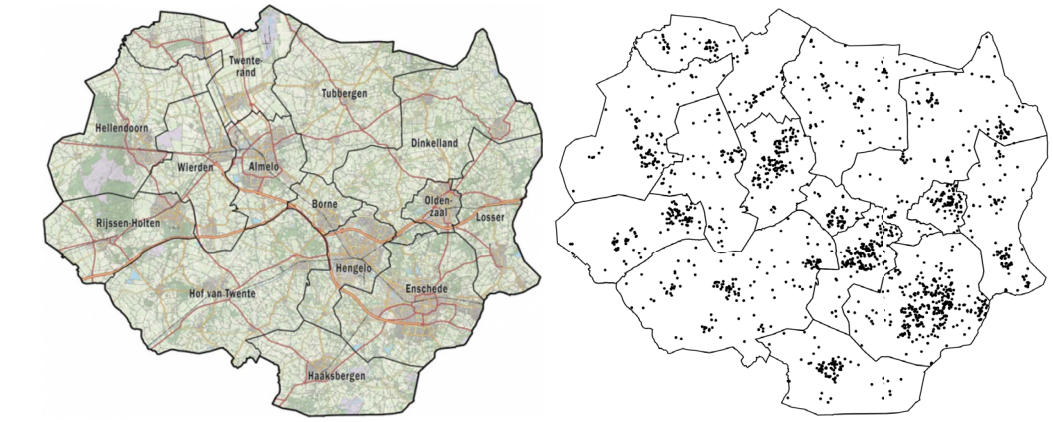
Received October 2022; revised February 2023.
Key words and phrases. Fire prediction, K-function, logistic regression estimation, pair correlation function, Poisson point process, spatiotemporal point pattern, variable importance.

spatiotemporal point patterns. To resolve such discrepancies, certain studies (e.g., Sakr et al. (2010), Stojanova et al. (2012), Satir, Berberoglu and Donmez (2016)) discretized fire incidents into areal unit data and transformed risk prediction from an occurrence modelling problem to a risk scale classification problem (i.e., labelling different amounts of fires to corresponding risk scales). However, doing so may result in a loss of information. In addition, machine learning approaches often require relatively large amounts of data to obtain a satisfactory performance, and due to their 'black box' behaviour, it is difficult to interpret the influence of an explanatory variable on fire risk in terms that are meaningful for practical implementation for fire services.

In contrast, statistical approaches enable us to learn spatiotemporal point patterns directly. Moreover, they are interpretable based on specific, parametric, mathematical models and allow for theoretical confidence intervals. For instance, Hering, Bell and Genton (2009) and Costafreda-Aumedes, Comas and Vega-Garcia (2016) used the K-function to analyze clustered patterns of wildfire data. Ye (2011) and Boubeta et al. (2015) employed Poisson structures to generate hazard maps of fire occurrences. Møller and Díaz-Avalos (2010) considered spatial and temporal explanatory variables in a shot-noise Cox process and fitted it with minimum contrast techniques. Pereira et al. (2013) and Serra et al. (2014) modelled fire occurrences using log-Gaussian Cox processes, which were particularly designed to simulate latent phenomena. Recently, a Bayesian framework was developed and suggested to improve fire prediction (Juan Verdoy (2021)). Pimont et al. (2021) also employed a log-Gaussian Cox process model to predict wildfires but established Bayesian inference for model components using integrated nested Laplace approximation (Rue, Martino and Chopin (2009)). Koh et al. (2023) developed a joint hierarchical model framework by combining extreme-value theory and point processes and studied summer wildfire data for the French Mediterranean basin. For chimney fires School (2018) proposed a log-Gaussian Cox process model to predict the risk based on explanatory variables selected by Pearson correlation coefficients and random effects simulated by Gaussian random fields. Most statistical studies (e.g., Møller and Díaz-Avalos (2010), Pereira et al. (2013)) first determined their model structures and selected and fitted explanatory variables afterward. This procedure does not fully exploit information contained in variables and sometimes requires complex modelling of residuals. In addition, the selection of explanatory variables is conducted either manually (Turner (2009)) or using basic statistical methods (Yang et al. (2015), School (2018)). Some studies (Thurman and Zhu (2014), Yue and Loh (2015), Choiruddin, Coeurjolly and Letué (2018)) applied regularized penalty functions in the estimation of point processes for variable selection. However, a parametric form of the intensity function, including all variables, needs to be specified in advance.

To conduct a data-driven fire risk prediction study, we combine machine learning and point process tools in our modelling process to leverage the advantages of both types of approaches. In a pilot study (Lu et al. (2021)), we used the permutation importance techniques of random forests to select important variables, as they are recommended as relatively accurate machine learning methods for fire prediction (Rodrigues and de la Riva (2014)). Accordingly, we designed a generalized linear Poisson model, based on the selected variables, and fitted it on areal unit data to predict the chimney fire risk in Twente. In this paper we complete the theoretical background and implementation details for the variable selection part and refine the areal unit model to a spatiotemporal Poisson point process model. Since point patterns (cf. Figure 1(b)) are targeted directly, the new model does not depend on the scale of the areal units, leads to a continuous hazard map for fire prediction and enables more detailed data analyses, such as point interaction tests.

Specifically, our contributions are as discussed below. First, we propose a modelling process for fire prediction, which is fully data-driven and meanwhile allows for model interpretability: (i) we use machine learning algorithms in a supporting role to select explanatory



(a) Twente municipalities (b) spatial pattern of fires

FIG. 1. *Map of Twente municipalities* (*a*) *and spatial projection of the chimney fire incidents during* 2004–2020 (*b*).

variables for fire risk nonparametrically so that variable information can be exploited; (ii) we adhere to a statistical Poisson point process model for fire occurrences with parametric structures designed on the selected explanatory variables so that complete model inference can be conducted. Second, in the fitting of the point process model, we efficiently estimate the parameters, using logistic regression estimation (Baddeley et al. (2014)), by tuning it to focus on important parts in space and time and provide theoretical confidence intervals.

The remainder of the paper is organized as follows. Section 2 introduces the fire data we use and some data preprocessing steps. Section 3 elaborates on the selection of important variables. In Section 4 we motivate and fit a Poisson point process model and compare it to the areal unit model proposed in Lu et al. (2021). Section 5 tests for point interactions in the fire data and validates the Poisson model assumption of absence of such interactions. Section 6 reviews our modelling process and explores the role of the dummy intensity in logistic regression estimation. Section 7 provides a conclusion and some ideas on future work.

2. Data. In this section, we introduce the chimney fire data, the relevant environmental variables and the data preprocessing operations.

2.1. *Chimney fires and environmental variables*. Collaborating with the Twente Fire Brigade, we collected the data of all reported chimney fire incidents occurring between January 1, 2004 and December 31, 2020, in the Twente region in the eastern part of the Netherlands (map shown in Figure 1(a)). After a manual check to remove obvious mistakes, the dataset consists of 1759 incidents. Each incident is reported individually with its ID number, location, time and a brief description of the circumstances of the fire and rescue processes.[1] The spatial and temporal projections[2] are plotted, respectively, in Figure 1(b) and Figure 2. It is clearly visible in Figure 1(b) that the spatial distribution of chimney fires is heterogeneous in the sense that most incidents occur in urban areas, especially in cities with a higher population such as Almelo, Hengelo and Enschede (cf. Figure 1(a)). Because neighbouring buildings tend to have identical chimney types, apparent clustering may also arise. Moreover,

[1]Location is recorded in the form of Dutch RD coordinates and in the unit of metre; time is recorded in the unit of day to match the weather data.

[2]Fire incidents on the leap days have already been excluded here, as discussed in Section 2.2.

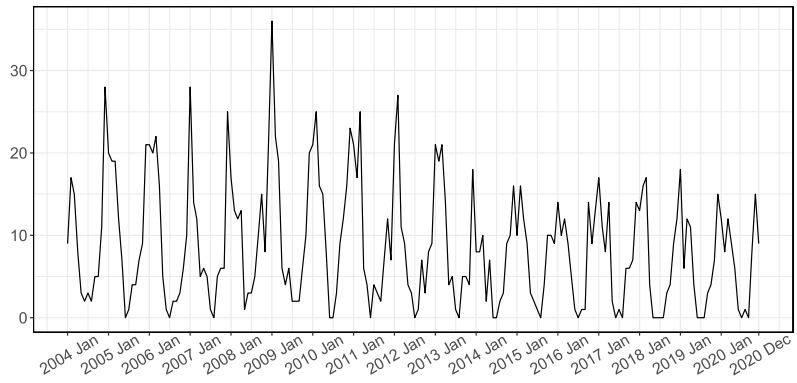


FIG. 2. *Temporal projection* (*monthly counts*) *of the chimney fire incidents* (*y-axis*) *during* 2004–2020.

locations neighbouring in space and time share similar weather conditions. In Figure 2 the temporal distribution of chimney fires is periodic: chimney fires occur more frequently in winter than in summer.

According to the experts from the Twente Fire Brigade, various environmental factors may influence the risk of chimney fires. Spatially, building types, population density and compositions and urbanity degrees in an area may determine the baseline fire risk. Temporally, season and weather conditions influence the risk as well. Other factors, such as fueling sources, chimney types and dominant human causes, may also affect it; however, their information is not accessible. Table 1 lists 27 putative explanatory variables with specific abbreviations, descriptions and sources. Among them, population and urbanity information are recorded over 6291 predefined 500 m $\times$ 500 m area boxes[3] in Twente, whereas building information contains the precise locations, ages and functions of houses. Temporal data comes in the form of daily weather observations from the weather station (Twenthe) at Twente airport. To assess the influence of small variations in weather among different parts of the Twente region, we also collected the data from two neighbouring weather stations (Heino and Hupsel) outside Twente for analysis. Their locations can be found on https://www.knmi.nl.

2.2. *Data preprocessing*. To unify the data to be used in our modelling process, we perform several data preprocessing operations on environmental variables.

First, we further extract building information at the level of living units, as it may reflect the geographical information of chimneys in a more representative way. For instance, two families living in one semidetached house actually use their individual chimneys. Second, some spatial variables (i.e., population and urbanity information) consist of a list of historical data, whereas others (i.e., building information) are only accessed in a single actual value. To enable similar treatment of the data, we use averaged values over the time period of interest for those variables consisting of a list of historical data so that for all spatial variables a single value is accessible. Moreover, we only consider the buildings that are currently in use and assume that buildings keep their functions and types during the time period of interest. Observing that the age and type information of some buildings are missing, we assign them the label "extra" during the counting process. Finally, we exclude the fire incidents and weather data on the leap days in the leap years, while in future fire risk prediction, we will use the prediction of February 28 to compensate for the missing prediction of the leap day, February 29.

[3]Determined by Statistics Netherlands (CBS, cf. column "Source" in Table 1).

TABLE 1
Putative explanatory variables, with their abbreviations, descriptions and sources. σ: Spatial variables, τ: Temporal variables

Variable	Abbreviation	Description	Source[a]
$V_{\sigma,1}$	House	The total number of houses	NIPV
$V_{\sigma,2}$	House_indu	The number of houses with an industrial function	NIPV
$V_{\sigma,3}$	House_hotl	The number of houses with a hotel function	NIPV
$V_{\sigma,4}$	House_resi	The number of houses with a residential function	NIPV
$V_{\sigma,5}$	House_20	The number of houses constructed before 1920	NIPV
$V_{\sigma,6}$	House_2045	The number of houses constructed between 1920 and 1945	NIPV
$V_{\sigma,7}$	House_4570	The number of houses constructed between 1945 and 1970	NIPV
$V_{\sigma,8}$	House_7080	The number of houses constructed between 1970 and 1980	NIPV
$V_{\sigma,9}$	House_8090	The number of houses constructed between 1980 and 1990	NIPV
$V_{\sigma,10}$	House_90	The number of houses constructed after 1990	NIPV
$V_{\sigma,11}$	House_frsd	The number of freestanding (detached or semi-detached) houses	NIPV
$V_{\sigma,12}$	Resid	The total number of residents	CBS
$V_{\sigma,13}$	Resid_14	The number of residents with an age in the range of 0 till 14	CBS
$V_{\sigma,14}$	Resid_1524	The number of residents with an age in the range of 15 till 24	CBS
$V_{\sigma,15}$	Resid_2544	The number of residents with an age in the range of 25 till 44	CBS
$V_{\sigma,16}$	Resid_4564	The number of residents with an age in the range of 45 till 64	CBS
$V_{\sigma,17}$	Resid_65	The number of residents with an age of 65 or higher	CBS
$V_{\sigma,18}$	Man	The number of male residents	CBS
$V_{\sigma,19}$	Woman	The number of female residents	CBS
$V_{\sigma,20}$	Address	The number of addresses in the neighbourhood	CBS
$V_{\sigma,21}$	Urbanity[b]	The urbanity of the neighbourhood	CBS
$V_{\sigma,22}$	Town	Boolean variable indicating the presence of a town	CBS
$V_{\tau,1}$	WindSpeed	Daily mean wind speed (km/h)	KNMI
$V_{\tau,2}$	Temperature	Daily mean temperature (°C)	KNMI
$V_{\tau,3}$	WindChill	Daily mean wind chill (°C) (calculated from $V_{\tau,1}$, $V_{\tau,2}$)	
$V_{\tau,4}$	Sunshine	Daily sunshine duration (h)	KNMI
$V_{\tau,5}$	Visibility[c]	Daily minimum visibility	KNMI

[a]NIPV: *Nederlands Instituut Publieke Veiligheid*, CBS: *Centraal Bureau voor de Statistiek*, KNMI: *Koninklijk Nederlands Meteorologisch Instituut*. All data is provided by the Twente Fire Brigade, especially for this research.
[b]Categorical, where the urbanity of a neighbourhood is evaluated into levels varying from 1 to 5 (urban–not urban), thus is treated as numerical in random forests.
[c]Categorical, where the minimum visibility distances (0–∞ km) are defined to levels varying from 1 to 80, thus is treated as numerical in random forests.

In addition, particularly for the point process modelling phase in Section 4, in order to obtain the data that records spatial variables for every location in Twente (i.e., density maps) from either areal unit data (i.e., population and urbanity information) or data with precise coordinates (i.e., building information), we employ kernel smoothing. For temporal variables we assume that their values at different times of a day remain invariant and equal to daily mean values. Recalling Table 1, we then introduce the following notation for environmental variables:

(i) $V_{\sigma,i}(u)$: The smoothed value of the ith spatial variable at location u,
(ii) $V_{\tau,i}(t)$: The assigned value of the ith temporal variable at time t,

where (u, t) denotes any location and time combination in the spatiotemporal domain.

3. Selection of explanatory variables. Obviously, not all environmental variables listed in Table 1 are required to model the risk of chimney fires, and some of them are mutually

dependent. To select the most informative ones, in Lu et al. (2021), we performed a nonparametric variable importance analysis using the permutation importance techniques of random forests (Breiman (2001)). For the sake of completeness, we elaborate here on the theoretical background and implementation details.

3.1. *Random forests and permutation importance.* Random forests (Breiman (2001)) are widely used as robust classification and regression methods in many applications, such as risk prediction (Wongvibulsin, Wu and Zeger (2019)) and data mining (Schonlau and Zou (2020)). A random forest is usually composed of hundreds or thousands of decision trees, where each tree is generated on a sampled subset of the data by repeated bagging (i.e., bootstrap sampling with replacement) and trained to fit explanatory variables to the response variable. Afterward, a combined result over all trees will be reported as the final output. In addition, random forests can also be used to assess the importance of a variable by means of permutation importance techniques (Breiman (2001), Strobl and Zeileis (2008), Altmann et al. (2010)). Through randomly permuting the values of a variable over the observations, the importance of a variable is defined as the mean increase of the prediction error over all trees computed on the permuted data compared to that computed on the original data.

More formally, consider a regression problem on a dataset $\mathcal{D}$ with n observations. Suppose that, for each observation in $\mathcal{D}$, there are m explanatory variables. Let $x_i^{(j)}$ and y_i, with $i = 1, \ldots, n$ and $j = 1, \ldots, m$, denote the jth explanatory variable and the response variable, respectively, for the ith observation; $\boldsymbol{x}_i$ is the vector that collects all $x_i^{(j)}$. The construction of a random forest consists of the generation of E decision trees. To generate a tree e, a subset of $\mathcal{D}$, denoted as $\mathcal{B}(e)$, is sampled by bagging. At each node of the tree, a number of explanatory variables are selected randomly from all variables as the candidates. Then one of the candidates, say $x^{(j)}$, is used to split the node into two subsets, $\mathcal{B}_L(e)$ and $\mathcal{B}_R(e)$ (e.g., for a numerical $x^{(j)}$, $\mathcal{B}_L(e) = \{(\boldsymbol{x}_i, y_i) : x_i^{(j)} \le c; (\boldsymbol{x}_i, y_i) \in \mathcal{B}(e)\}$ and $\mathcal{B}_R(e) = \{(\boldsymbol{x}_i, y_i) : x_i^{(j)} > c; (\boldsymbol{x}_i, y_i) \in \mathcal{B}(e)\}$), in such a way that the residual sum of squares

$$\sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{B}_L(e)} (y_i - \bar{y}_{\mathcal{B}_L(e)})^2 + \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{B}_R(e)} (y_i - \bar{y}_{\mathcal{B}_R(e)})^2 \tag{3.1}$$

is minimized. Here $\bar{y}_{\mathcal{B}_L(e)}$ and $\bar{y}_{\mathcal{B}_R(e)}$ denote the mean of the response variables in corresponding subsets. Such a node splitting procedure is continued until tree e satisfies certain preconditions (e.g., reaching the maximum number of levels of a tree). Similar tree generation processes are employed to construct all other trees in the forest. To measure the importance of an explanatory variable, again say $x^{(j)}$ in tree e, we randomly permute its values over the out-of-bag observations of the tree, denoted as $o\mathcal{B}(e)$ (i.e., $o\mathcal{B}(e) = \mathcal{D} \setminus \mathcal{B}(e)$). Write π_j for the permutation. The value of the jth explanatory variable for the response y_i for an observation $(\boldsymbol{x}_i, y_i)$ then changes from $x_i^{(j)}$ to $x_{\pi_j(i)}^{(j)}$ with $(\boldsymbol{x}_i, y_i) \in o\mathcal{B}(e)$, while the values of other explanatory variables are left unchanged. Thus, the increase of the prediction error is

$$I(x^{(j)}; e) = \frac{\sum_{(\boldsymbol{x}_i, y_i) \in o\mathcal{B}(e)} (y_i - \hat{y}_{i,\pi_j})^2}{|o\mathcal{B}(e)|} - \frac{\sum_{(\boldsymbol{x}_i, y_i) \in o\mathcal{B}(e)} (y_i - \hat{y}_i)^2}{|o\mathcal{B}(e)|}, \tag{3.2}$$

where $\hat{y}_i$ denotes the prediction in tree e for observation $(\boldsymbol{x}_i, y_i)$ using original explanatory variables, $\hat{y}_{i,\pi_j}$ denotes the corresponding prediction but using explanatory variables with the jth variable permuted and $|o\mathcal{B}(e)|$ denotes the number of out-of-bag observations for the tree. Finally, the mean increase of the prediction error over all trees, $\sum_e I(x^{(j)}; e)/E$, is used to illustrate the importance of the jth variable, $x^{(j)}$, in the forest.

In practice, the original construction algorithms of random forests tend to bias the variable selection at tree nodes to factorial variables with many categories or continuous variables with

many cut points. Moreover, traditional permutation importance techniques can sometimes be misled by correlated explanatory variables. To address these problems, unbiased random forests (Hothorn, Hornik and Zeileis (2006), Strobl et al. (2007)) and conditional permutation importance techniques (Strobl et al. (2008)) were proposed based on a conditional inference framework of recursive partitioning. Compared to earlier methods, the idea is to partition the variable space in order to obtain groups of observations with similar association patterns instead of groups of observations with merely similar values of the response variable. For instance, to measure the importance of variable $x^{(j)}$ under conditional permutation in tree e, the set of out-of-bag observations, $o\mathcal{B}(e)$, is first partitioned into a grid where each block shares the same information on the remaining variables, $\boldsymbol{x} \setminus x^{(j)}$. Then the same permutation operation as introduced above is applied to each block, and the increase of the prediction error is computed and summed up over all blocks to illustrate the importance of $x^{(j)}$. With significant increments and optimizations, the random forest approach has been very useful to measure variable importance nonparametrically considering variable dependence (Strobl, Hothorn and Zeileis (2009)).

Note that the notation used in this section is only valid here to explain random forests and permutation importance techniques statistically.

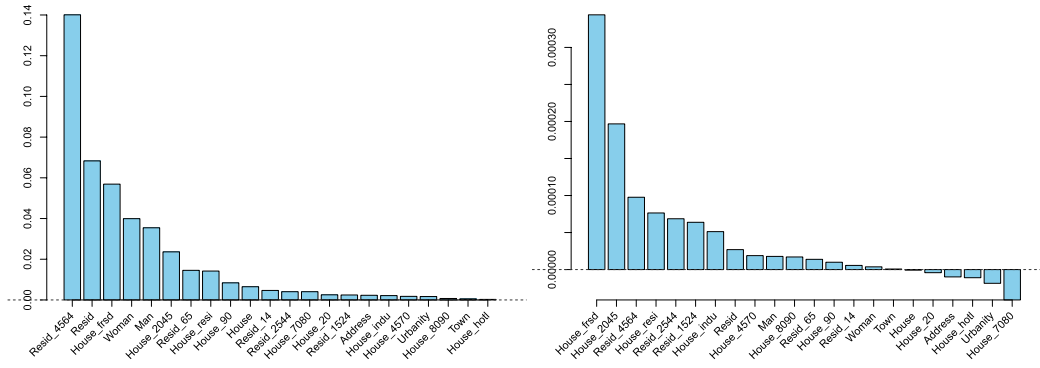
3.2. *Variable importance analysis*. In Lu et al. (2021), we measured the importance of spatial and temporal variables separately on areal unit data. The main reason for this is that spatial variables are available as a single actual value for every area box while temporal variables are available as historical time series.

Specifically, for spatial variables we grouped the incidents by the 6291 predefined 500 m $\times$ 500 m area boxes and merged them into a large data frame consisting of 6291 rows and 23 columns. Each row indicates an area box, and the 23 columns refer to the number of incidents occurring in that box (response variable) as well as the values of 22 spatial environmental variables. We found that some boundary area boxes lie only partially in Twente. However, the population statistics for such boxes do not distinguish between Twente and the neighbouring regions. To avoid bias, we filtered out the data of these boundary boxes in our variable importance analysis. For temporal variables we grouped the incidents in 6205 days and merged them into a data frame consisting of 6205 rows and six columns. Each row indicates a specific day, and the six columns refer to the number of incidents on that day (response variable) as well as the values of five temporal environmental variables.

In the analysis we constructed two unbiased random forests of 2000 trees to fit spatial and temporal variables to the number of incidents in corresponding data tables, respectively, and set the proportion of the number of input variables that are randomly sampled as the candidates at a tree node to approximately 1/3 (i.e., spatial: seven candidates, temporal: two candidates). Considering high correlations between certain variables (e.g., population and urbanity information), we used conditional permutation importance techniques (Strobl et al. (2008)), instead of the traditional ones (Breiman (2001)), to suppress the importance scores biasing toward correlated variables. Our implementation of variable importance analysis was based on the R-package *party* (Hothorn et al. (2006), Strobl et al. (2007)). We used the method proposed in Debeer and Strobl (2020) to compute the conditional permutation importance, which provides a faster computation and shows more stable results than the original implementation (Strobl et al. (2008)).

The variable importance results,[4] under conditional permutation for spatial and temporal variables, are plotted in Figure 3 and Figure 4, respectively. For comparison, we also plot

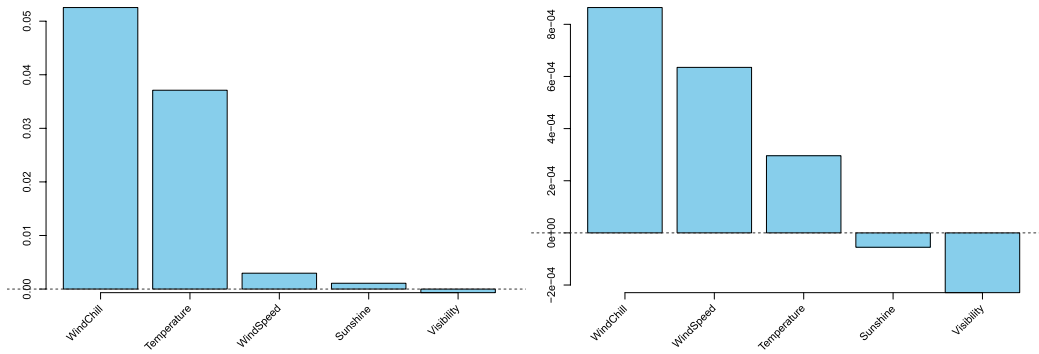
[4]The importance scores for spatial and temporal variables are not comparable, as the two random forests were fitted on spatial and temporal variables separately.



(a) traditional permutation (b) conditional permutation

FIG. 3. *Importance* (*y-axis*) *obtained for spatial variables using traditional* (*a*) *and conditional* (*b*) *permutation techniques*.

the results under traditional permutation. In all plots the y-axis refers to the increase of the squared prediction error when traditional or conditional permutation on a variable is applied. A large increase indicates that the variable is very important for correct predictions, while a decrease indicates that the variable has no influence on or even hampers the prediction. If we compare the results under traditional and conditional permutation, we see that the bias caused by correlated variables is suppressed well. For instance, the number of residents aged between 45 and 64 and the total number of residents have the largest importance scores under traditional permutation, with the number of freestanding houses coming third. However, the latter has the highest variable importance score under conditional permutation. A possible explanation could be that freestanding houses are mostly occupied by families whose adult members are aged between 45 and 64, which implies a strong positive correlation among all three categories. Moreover, other age groups, such as the elderly or families with young children, may be less inclined to use their chimneys. The importance scores of the two variables concerning residents decrease a lot under conditional permutation. Similar observations hold for temporal variables as well. Both wind chill and temperature obtain high importance scores under traditional permutation, although they are correlated. Conditional permutation detects the underlying correlation and suppresses temperature, as wind chill is defined in terms of both temperature and wind speed information. However, since wind speed still obtains a relatively large importance score, even under conditional permutation, we will further



(a) traditional permutation (b) conditional permutation

FIG. 4. *Importance* (*y-axis*) *obtained for temporal variables using traditional* (*a*) *and conditional* (*b*) *permutation techniques*.

consider it in the modelling phase. The additional analysis to assess the influence of small weather variations among different parts in Twente indicates that these variations can be ignored.[5]

Overall, according to the variable importance under conditional permutation, *the number of buildings constructed between 1920 and 1945* and *the number of freestanding houses* are the most important spatial variables, and *wind chill* and *wind speed* are the most important temporal variables. The results could have the following explanations: (i) most freestanding houses contain chimneys, whereas other types do not; (ii) chimney pipes in old buildings tend to be made of brick rather than metal, which increases the risk to catch fires; (iii) strong wind can fuel a fire; (iv) wind chill reflects people feeling cold, thus inducing them to use their fires and chimneys.

4. Poisson point process model. In this section we refine the areal unit model for chimney fire prediction proposed in Lu et al. (2021) to a spatiotemporal point process model. First, we review the motivation for the nested model structure designed in Lu et al. (2021) using the selected variables from Section 3.2. Second, we define a Poisson point process model based on the same structure. Afterward, we present model fitting and selection procedures and propose theoretical confidence intervals for both model parameters and predicted fire intensities. Finally, we compare the new point process model to the areal unit model and illustrate the former's advantages.

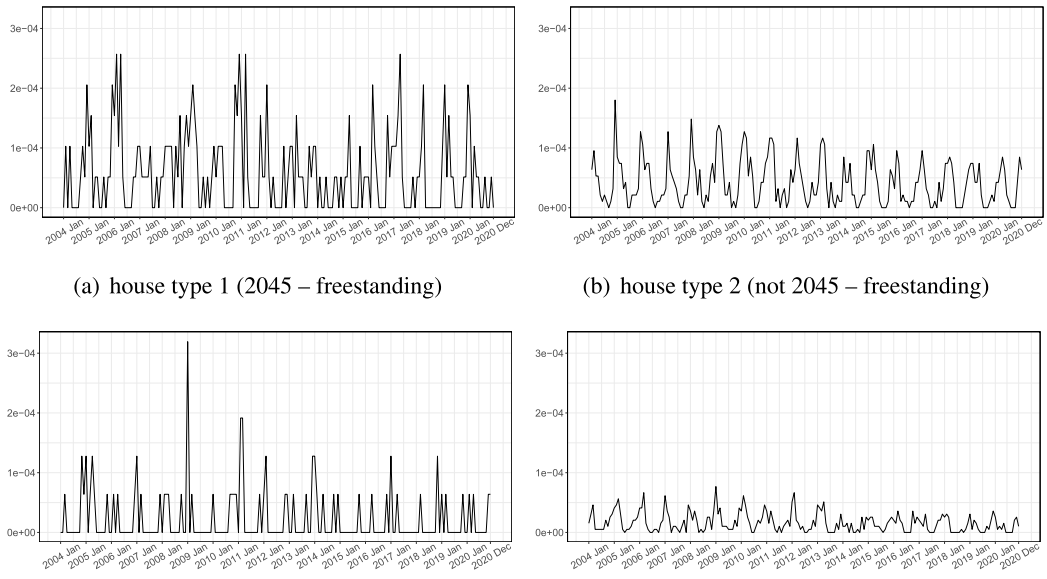
4.1. *Model motivation*. In preparation for an appropriate model structure for chimney fire prediction in Lu et al. (2021), we performed a preliminary investigation on the relations between the selected explanatory variables and chimney fire occurrences using areal unit data.

Recalling the selected variables in Section 3.2, we first divided the houses in Twente into four house types depending on their ages (i.e., whether they have been constructed between 1920 and 1945 or not) and on whether they are freestanding or not and plot the monthly intensities of chimney fires per house for different house types separately. Figure 5 shows that, generally, chimney fire occurrences in all house types are periodic, with incidents concentrated in the colder seasons. However, different house types run different risks of chimney fires: the intensities for old houses (i.e., House_2045) and freestanding houses are higher than others. Moreover, note that the patterns are not perfectly periodic; the amplitudes of the peaks vary per year. To model such varying patterns, we need to take wind chill and wind speed into account.

In addition, we plot the numerical relations between chimney fire occurrences and the number of houses of different types in Figure 6. To approximate the underlying trends in the scatter plots, we apply the locally estimated scatter plot smoothing method (Cleveland, Grosse and Shyu (1992)) and plot the estimated trends. It is interesting to find that, for each house type, chimney fire occurrences increase approximately linearly in the number of houses of that type. However, this observation only holds when the number of houses is not very large. If it is too large, the number of chimney fires tends to a saturated value.

To summarize, we draw the following conclusions: (i) chimneys in different house types catch fires at a type-dependent rate, and the rate is influenced by both seasonal information and temporal explanatory variables; (ii) chimney fire occurrences in the houses of a specific type are approximately proportional to the number of houses of that type.

[5]For details we refer to the Supplementary Material (Lu et al. (2023)).



(a) house type 1 (2045 – freestanding)

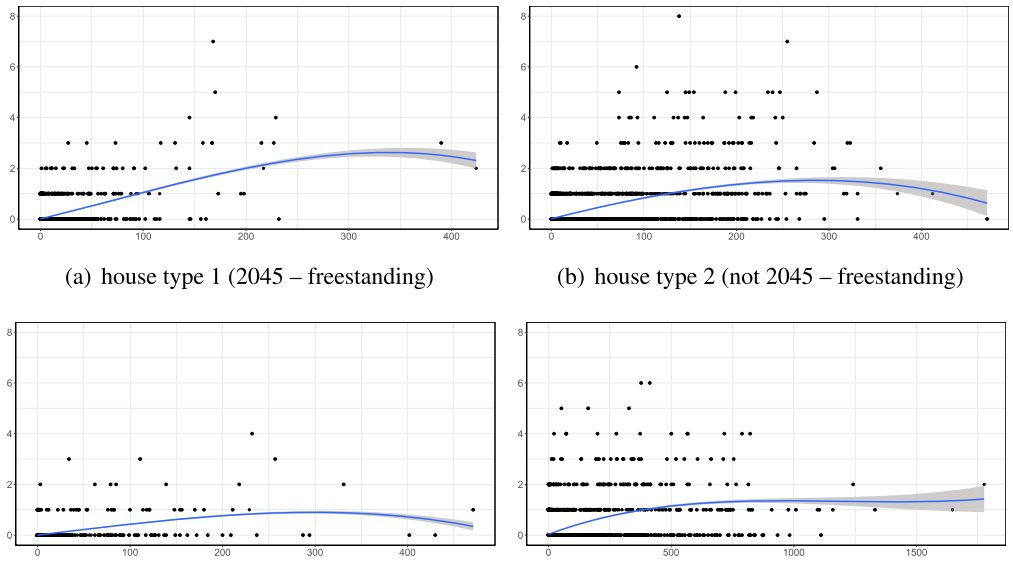
(b) house type 2 (not 2045 – freestanding)

(c) house type 3 (2045 – not freestanding)

(d) house type 4 (not 2045 – not freestanding)

FIG. 5. *Monthly intensities of chimney fires per house* (y*-axis*) *for different house types during* 2004–2020.

4.2. *Model structure*. Based on the motivation above, a nested Poisson model structure was designed for chimney fire risk prediction (Lu et al. (2021)). Here we construct a point process model which uses the same structure but learns point patterns (cf. Figure 1(b)) directly. We assume that the explanatory variables for the areal unit model and the point process model are identical.



(a) house type 1 (2045 – freestanding)

(b) house type 2 (not 2045 – freestanding)

(c) house type 3 (2045 – not freestanding)

(d) house type 4 (not 2045 – not freestanding)

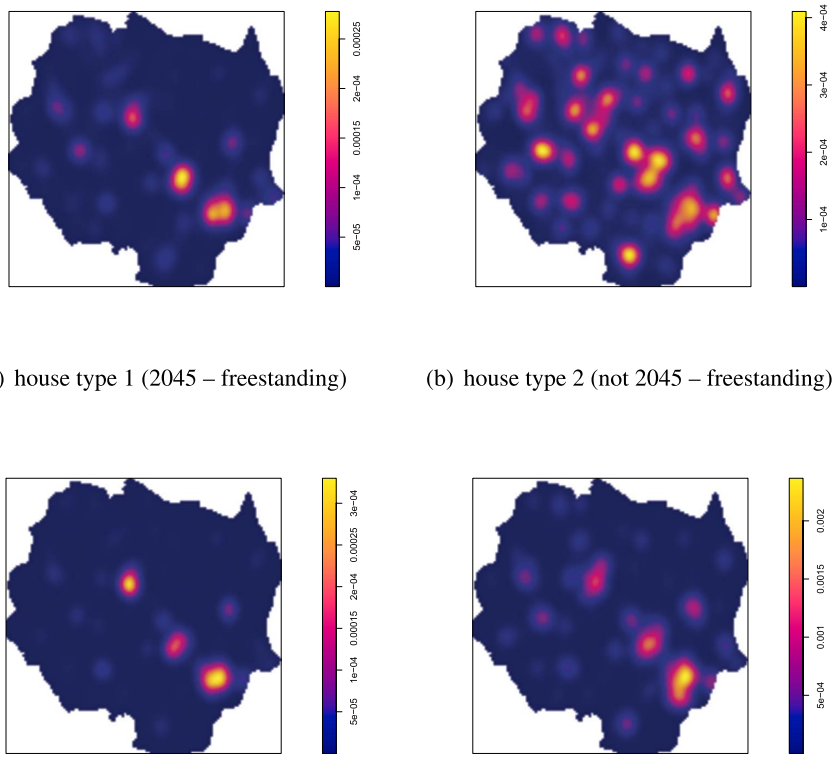
FIG. 6. *Numerical relations between chimney fire occurrences* (y*-axis*) *and the number of houses of different types* (x*-axis*). *The blue curves are the estimated trends using the locally estimated scatterplot smoothing method with parameter* span $= 1$. *The* 95% *envelopes are also included to show the confidence of such estimates*.

Suppose that each house catches chimney fires independently at a type-dependent rate that varies in time. Under this assumption the fire prediction model is a spatiotemporal Poisson point process, defined on the Twente region and the period 2004–2020, with an intensity function of the form $\Lambda(u,t) = \sum_k \lambda_k(u,t)$, where $\Lambda(u,t)$ indicates the overall fire risk at location u at time t and $\lambda_k(u,t)$ indicates the risk at location u at time t for the houses of type k. Furthermore, we define $\lambda_k(u,t) = h_k(u)\phi_k(t)$, where $h_k(u)$ indicates the density of house type k at location u and $\phi_k(t)$ indicates the risk for a house of that type at time t. Such a structure reflects the conclusions reached in Section 4.1, although it relies on the condition that the density of houses at a location is not too large (cf. Figure 6).

More formally, our nested Poisson point process model is defined by its intensity function

$$\Lambda(u,t) = \sum_{k=1}^{4} \lambda_k(u,t) = \sum_{k=1}^{4} h_k(u)\phi_k(t). \tag{4.1}$$

The density maps of houses of four types (i.e., $h_k(u)$), as shown in Figure 7, are derived from corresponding building information (i.e., $V_{\sigma,6}(u)$ and $V_{\sigma,11}(u)$) by smoothing using Gaussian kernels with a standard deviation of 1000 metres.[6] Moreover, since both seasonal information and temporal variables, wind speed and wind chill (i.e., $V_{\tau,1}(t)$ and $V_{\tau,3}(t)$), are taken into



(a) house type 1 (2045 – freestanding) (b) house type 2 (not 2045 – freestanding)

(c) house type 3 (2045 – not freestanding) (d) house type 4 (not 2045 – not freestanding)

FIG. 7. *Density maps of houses of four types. The unit in all plots is metre*$^{-2}$. *Note that the intensity bars have different scales.*

[6]Suggested by the experts from the Twente Fire Brigade, considering the distances between cities in Twente.

account, we propose a temporal intensity function $\phi_k(t)$ for houses of type k of the form

$$\begin{aligned}\phi_k(t) = \exp\big(&\text{Harmonic}(t; o_{k,1}) + \text{Polynom}\big(V_{\tau,1}(t); o_{k,2}\big)\\ &+ \text{Polynom}\big(V_{\tau,3}(t); o_{k,3}\big) + \text{Polynom}\big(V_{\tau,1}(t)V_{\tau,3}(t); o_{k,4}\big)\big),\end{aligned} \tag{4.2}$$

where a harmonic function[7] with order $o_{k,1}$ is employed to model seasonal variations and two polynomial functions with order $o_{k,2}$ and $o_{k,3}$ are used to model information of wind speed and wind chill. In addition, we employ another polynomial function of $V_{\tau,1}(t)V_{\tau,3}(t)$ with order $o_{k,4}$ to allow for interactions between wind speed and wind chill. The exponential function is applied to guarantee that the intensity function stays positive. For purposes of parametric representation, the temporal intensity function $\phi_k(t)$ can also be described as $\phi_k(t;\boldsymbol{\theta}_k)$ (accordingly, $\lambda_k(u,t)$ as $\lambda_k(u,t;\boldsymbol{\theta}_k)$), where $\boldsymbol{\theta}_k$ indicates the vector of coefficients in the harmonic and polynomial functions for house type k.

4.3. *Model fitting.* In Lu et al. (2021), the areal unit model was fitted by maximum likelihood estimation. Moreover, we established theoretical confidence intervals for model parameters and predicted fire intensities based on classical theories: the Fisher information for asymptotic normality for generalized linear models (Fahrmeir and Kaufmann (1985), McCullagh and Nelder (2019)) and the Delta method (Ver Hoef (2012)).

In principle, we could also apply maximum likelihood estimation here to maximize the log-likelihood function of our Poisson point process model, which reads

$$\sum_{k=1}^{4}\Bigg\{\sum_{x_k}\log\lambda_k(x_k;\boldsymbol{\theta}_k) - \int_{\text{Twente}}\int_{\text{Year 2004–2020}}\lambda_k(u,t;\boldsymbol{\theta}_k)\,du\,dt\Bigg\}, \tag{4.3}$$

where x_k denotes combinations of location and time and runs through chimney fire incidents occurring in houses of type k. However, since we have a continuous space-time domain, the integral in the log-likelihood function must be approximated numerically, for example, using quadrature points (Baddeley and Turner (2000)). In our case the number of quadrature points would have to be rather large; thus, we prefer to use logistic regression estimation (Baddeley et al. (2014)) to fit the Poisson point process model, where we can freely adjust the distribution of quadrature points.

The idea behind the logistic regression estimation approach is the well-known Campbell–Mecke theorem (see, e.g., Daley and Vere-Jones (2008)). Consider a point process X that is defined on a space-time domain $\mathbb{R}^2\times\mathbb{R}$ and is locally finite, which means that X places almost surely finitely many points in every bounded subset of $\mathbb{R}^2\times\mathbb{R}$. We assume that X is confined to a bounded domain $W\times T\subseteq\mathbb{R}^2\times\mathbb{R}$ and has intensity function $\lambda(u,t)$ with $(u,t)\in W\times T$. For any vector of real-valued nonnegative or integrable functions $\boldsymbol{f}(u,t)$, defined on $W\times T$, the Campbell–Mecke theorem reads

$$E\Bigg\{\sum_{x\in X}\boldsymbol{f}(x)\Bigg\} = \int_W\int_T \boldsymbol{f}(u,t)\lambda(u,t)\,du\,dt, \tag{4.4}$$

where x runs through the points of X. Suppose that the intensity function λ is parametric and depends on a parameter vector $\boldsymbol{\theta}$, thus can be described as $\lambda(u,t;\boldsymbol{\theta})$. One may estimate both sides of the Campbell–Mecke theorem and then equate them to obtain estimating equations for $\boldsymbol{\theta}$.

The logistic regression estimation approach is based on the components of the vector function

$$\boldsymbol{f}(u,t) = \frac{\partial}{\partial\boldsymbol{\theta}}\log\Bigg[\frac{\lambda(u,t;\boldsymbol{\theta})}{\lambda(u,t;\boldsymbol{\theta})+\rho(u,t)}\Bigg] = \frac{\rho(u,t)/\lambda(u,t;\boldsymbol{\theta})}{\lambda(u,t;\boldsymbol{\theta})+\rho(u,t)}\nabla\lambda(u,t;\boldsymbol{\theta}), \tag{4.5}$$

[7]Harmonic function here refers to cosine and sine functions.

where $\nabla\lambda(u,t;\boldsymbol{\theta})$ denotes the gradient of $\lambda(u,t;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and ρ is a positive-valued function defined on $W \times T$. Here one needs to assume that $\lambda > 0$. To estimate the integral in (4.4) with $\boldsymbol{f}(u,t)$, defined in (4.5), one can use a "dummy" point process D on $W \times T$ that is independent of X and has intensity function ρ. Applying the Campbell–Mecke theorem (4.4) to D, one has

$$E\left\{\sum_{x\in D}\frac{1}{\lambda(x;\boldsymbol{\theta})+\rho(x)}\nabla\lambda(x;\boldsymbol{\theta})\right\} = \int_W\int_T \frac{\rho(u,t)}{\lambda(u,t;\boldsymbol{\theta})+\rho(u,t)}\nabla\lambda(u,t;\boldsymbol{\theta})\,du\,dt. \tag{4.6}$$

Having obtained unbiased estimators for both left- and right-hand side of equation (4.4) with $\boldsymbol{f}$ given by (4.5), one can plug them in to obtain the estimating equation. It solves

$$\boldsymbol{s}(X,D;\boldsymbol{\theta}) = \sum_{x\in X}\frac{\rho(x)/\lambda(x;\boldsymbol{\theta})}{\lambda(x;\boldsymbol{\theta})+\rho(x)}\nabla\lambda(x;\boldsymbol{\theta}) - \sum_{x\in D}\frac{1}{\lambda(x;\boldsymbol{\theta})+\rho(x)}\nabla\lambda(x;\boldsymbol{\theta}) = \mathbf{0} \tag{4.7}$$

over the parameter $\boldsymbol{\theta}$. Note that the subscript x in the two sums runs through the points of X and D, respectively. It is interesting to observe that (4.7) is precisely the derivative of a logistic log-likelihood function

$$l(X,D;\boldsymbol{\theta}) = \sum_{x\in X}\log\left[\frac{\lambda(x;\boldsymbol{\theta})}{\lambda(x;\boldsymbol{\theta})+\rho(x)}\right] + \sum_{x\in D}\log\left[\frac{\rho(x)}{\lambda(x;\boldsymbol{\theta})+\rho(x)}\right]. \tag{4.8}$$

In this way the fitting of a Poisson point process model can be transformed into a maximum likelihood estimation problem of logistic regression.[8] It can be easily implemented, based on the R-package *stats* (Venables and Ripley (2002)), and treated as the fitting of generalized linear models with a *logit* link function. Since the log-likelihood function is concave, the existence and uniqueness of parameter $\boldsymbol{\theta}$ that maximizes the log-likelihood are guaranteed under certain conditions (Silvapulle (1981)).

4.4. *Modelling chimney fire data.* As introduced above, we use logistic regression estimation to fit the intensity functions of our four house types separately. Consider (4.2) for fixed k and write $\boldsymbol{\theta}_k$ as above for the vector of coefficients in the harmonic and polynomial functions. Also, consider the chimney fires occurring in houses of type k as a point process X_k with intensity function $\lambda_k(u,t;\boldsymbol{\theta}_k)$. Denote the corresponding dummy point process by D_k with intensity function $\rho_k(u,t)$. Assuming that we have m parameters in $\boldsymbol{\theta}_k$, the estimating equations then read

$$\begin{aligned} s_p(X_k,D_k;\boldsymbol{\theta}_k) &= \sum_{x\in X_k}\frac{\rho_k(x)C_p(x)}{\lambda_k(x;\boldsymbol{\theta}_k)+\rho_k(x)} - \sum_{x\in D_k}\frac{\lambda_k(x;\boldsymbol{\theta}_k)C_p(x)}{\lambda_k(x;\boldsymbol{\theta}_k)+\rho_k(x)} \\ &= 0, \quad p = 1,\ldots,m, \end{aligned} \tag{4.9}$$

where $C_p(x)$ denotes the pth temporal covariate (i.e., the harmonic components, wind speed and wind chill terms with different orders and the interaction terms of wind chill and wind speed in (4.2)) at point x.

To estimate the parameters in $\boldsymbol{\theta}_k$, we perform a three-step implementation. First, we specify the intensity function $\rho_k(u,t)$ of the dummy point process D_k. Note that the role of the dummy point process is to approximate the integral in the Campbell–Mecke theorem (4.4) by (4.6). In order to efficiently estimate the parameters, we need to put more dummy points in the regions and times with a relatively high number of fire occurrences so that the approximation of the integral in those parts can be more accurate. Consider that, spatially, most chimney

[8]Note that the point process model itself is not a logistic model; only the estimation takes the form of logistic regression. For details we refer to Baddeley et al. (2014).

fires occur in urban areas, and that temporally, there are more chimney fires in winter than in summer; specifically, we use the density map of houses of the given type k (cf. Figure 7) to distinguish urban areas from rural areas and employ a sine function to assign higher $\rho_k(u,t)$ to cold seasons to comply with the observations in Figure 2. Thus, the intensity function $\rho_k(u,t)$ of D_k is given by

$$\rho_k(u,t) = r_k h_k(u)\left(0.5 + 0.25\left(\sin\left(\frac{2\pi}{365}t + \frac{\pi}{2}\right) + 1\right)\right), \tag{4.10}$$

where r_k is a multiplication factor[9] used to ensure that $\rho_k(u,t)$ is at least four times $\lambda_k(u,t;\boldsymbol{\theta}_k)$, as suggested in Baddeley et al. (2014), and $h_k(u)$ is the density of houses of type k at location u. Second, we generate a realisation from the dummy point process D_k using the R-package *spatstat* (Baddeley, Rubak and Turner (2015)). Third, based on the observation of X_k and the realisation of the dummy point process D_k, we use the R-package *stats* (Venables and Ripley (2002)) to estimate model parameters $\boldsymbol{\theta}_k$. In addition, in order to determine the optimal function orders (i.e., $o_{k,1}$, $o_{k,2}$, $o_{k,3}$, $o_{k,4}$ in (4.2)) simultaneously, we apply a grid search to select the combination that yields the smallest Akaike information criterion (Choiruddin, Coeurjolly and Waagepetersen (2021)), given a set of ranges[10] for them. Considering that wind chill already contains some information of wind speed, we also perform likelihood ratio tests over the best models with and without wind speed and the interaction terms.

Similar to Lu et al. (2021), we separate the fire data into two sets: the data on the period 2004–2019 and the data on the year 2020 for fitting and predicting purposes, respectively. Performing the model selection introduced above on the data over 2004–2019, we obtain the models for each of the four house types. We find that wind chill alone is mostly sufficient to obtain accurate risk predictions, except for house type 4 where including the interaction terms provides a better fitting performance. Finally, the temporal intensity functions obtained for the four house types read as follows:

$$\begin{aligned}\phi_1(t;\boldsymbol{\theta}_1) = \exp\Bigg(&\theta_{1,1} + \theta_{1,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{1,3}\sin\left(\frac{2\pi}{365}t\right) + \theta_{1,4}\cos\left(\frac{4\pi}{365}t\right)\\ &+ \theta_{1,5}\sin\left(\frac{4\pi}{365}t\right) + \theta_{1,6}\cos\left(\frac{6\pi}{365}t\right) + \theta_{1,7}\sin\left(\frac{6\pi}{365}t\right) + \theta_{1,8}\cos\left(\frac{8\pi}{365}t\right)\\ &+ \theta_{1,9}\sin\left(\frac{8\pi}{365}t\right) + \theta_{1,10}V_{\tau,3}(t) + \theta_{1,11}V_{\tau,3}^2(t)\Bigg),\end{aligned} \tag{4.11}$$

$$\begin{aligned}\phi_2(t;\boldsymbol{\theta}_2) = \exp\Bigg(&\theta_{2,1} + \theta_{2,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{2,3}\sin\left(\frac{2\pi}{365}t\right) + \theta_{2,4}\cos\left(\frac{4\pi}{365}t\right)\\ &+ \theta_{2,5}\sin\left(\frac{4\pi}{365}t\right) + \theta_{2,6}\cos\left(\frac{6\pi}{365}t\right) + \theta_{2,7}\sin\left(\frac{6\pi}{365}t\right) + \theta_{2,8}V_{\tau,3}(t)\\ &+ \theta_{2,9}V_{\tau,3}^2(t) + \theta_{2,10}V_{\tau,3}^3(t) + \theta_{2,11}V_{\tau,3}^4(t)\Bigg),\end{aligned} \tag{4.12}$$

$$\begin{aligned}\phi_3(t;\boldsymbol{\theta}_3) = \exp\Bigg(&\theta_{3,1} + \theta_{3,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{3,3}\sin\left(\frac{2\pi}{365}t\right) + \theta_{3,4}\cos\left(\frac{4\pi}{365}t\right)\\ &+ \theta_{3,5}\sin\left(\frac{4\pi}{365}t\right) + \theta_{3,6}\cos\left(\frac{6\pi}{365}t\right) + \theta_{3,7}\sin\left(\frac{6\pi}{365}t\right) + \theta_{3,8}V_{\tau,3}(t)\Bigg),\end{aligned} \tag{4.13}$$

[9] Specifically, we set r_k's to 60, 20, 20 and 8 for the four house types, respectively.

[10] $o_{k,1}$: 1–4, $o_{k,2}$: 1–5, $o_{k,3}$: 1–5, $o_{k,4}$: 1–5.

TABLE 2
Parameter estimates for a Poisson point process model defined by the intensity functions of (4.11)–(4.14) and their 95% confidence intervals (CI). The "e" denotes a base of 10

Parameter	Estimate (CI)	Parameter	Estimate (CI)	Parameter	Estimate (CI)
$\theta_{1,1}$	−1.22e1(±3.82e−1)	$\theta_{1,2}$	−2.78e−1(± 4.81e−1)	$\theta_{1,3}$	−1.44e−1(± 2.84e−1)
$\theta_{1,4}$	−1.63e−1(± 2.64e−1)	$\theta_{1,5}$	7.99e−2(± 3.04e−1)	$\theta_{1,6}$	−9.21e−3(± 2.49e−1)
$\theta_{1,7}$	−2.33e−2(± 2.62e−1)	$\theta_{1,8}$	3.05e−1(± 2.19e−1)	$\theta_{1,9}$	1.59e−2(± 2.20e−1)
$\theta_{1,10}$	−7.42e−2(± 3.56e−2)	$\theta_{1,11}$	−6.12e−3(± 3.31e−3)	$\theta_{2,1}$	−1.30e1(± 2.31e−1)
$\theta_{2,2}$	6.40e−2(± 2.74e−1)	$\theta_{2,3}$	1.42e−2(± 1.56e−1)	$\theta_{2,4}$	−3.14e−2(± 1.64e−1)
$\theta_{2,5}$	1.69e−1(± 1.60e−1)	$\theta_{2,6}$	1.48e−1(± 1.24e−1)	$\theta_{2,7}$	−4.54e−2(± 1.27e−1)
$\theta_{2,8}$	−6.81e−2(± 2.64e−2)	$\theta_{2,9}$	2.33e−3(± 3.65e−3)	$\theta_{2,10}$	−8.50e−6(± 2.47e−4)
$\theta_{2,11}$	−2.76e−5(± 1.93e−5)	$\theta_{3,1}$	−1.43e1(± 9.48e−1)	$\theta_{3,2}$	1.57e0(± 1.60e0)
$\theta_{3,3}$	2.25e−1(± 6.00e−1)	$\theta_{3,4}$	−1.19e0(± 1.09e0)	$\theta_{3,5}$	−3.19e−1(± 7.20e−1)
$\theta_{3,6}$	6.32e−1(± 6.06e−1)	$\theta_{3,7}$	1.65e−1(± 5.39e−1)	$\theta_{3,8}$	−1.12e−1(± 5.25e−2)
$\theta_{4,1}$	−1.39e1(± 2.41e−1)	$\theta_{4,2}$	1.49e−1(± 3.05e−1)	$\theta_{4,3}$	6.11e−2(± 1.76e−1)
$\theta_{4,4}$	−2.06e−1(± 1.83e−1)	$\theta_{4,5}$	9.00e−2(± 1.95e−1)	$\theta_{4,6}$	2.77e−2(± 1.61e−1)
$\theta_{4,7}$	7.11e−2(± 1.72e−1)	$\theta_{4,8}$	1.82e−1(± 1.36e−1)	$\theta_{4,9}$	−4.84e−2(± 1.38e−1)
$\theta_{4,10}$	−9.38e−2(± 4.06e−2)	$\theta_{4,11}$	−7.92e−4(± 1.99e−3)	$\theta_{4,12}$	−2.35e−4(± 1.62e−4)
$\theta_{4,13}$	2.99e−3(± 2.10e−3)				

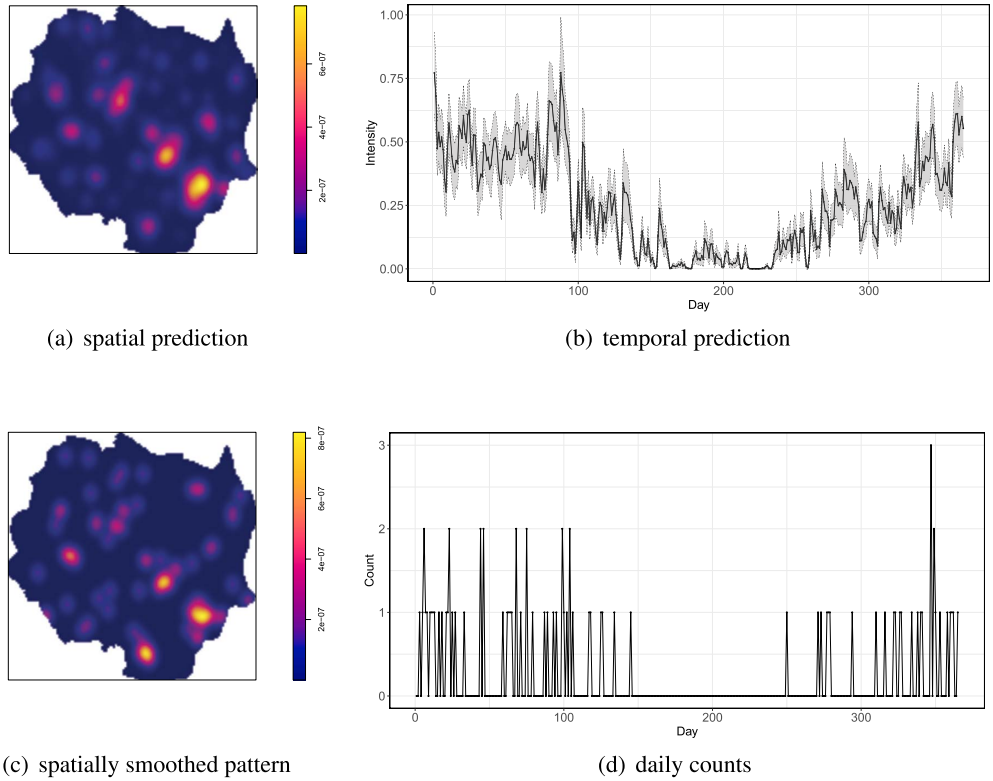
$$\begin{aligned}\phi_4(t;\boldsymbol{\theta}_4) = \exp\Bigg(&\theta_{4,1} + \theta_{4,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{4,3}\sin\left(\frac{2\pi}{365}t\right) + \theta_{4,4}\cos\left(\frac{4\pi}{365}t\right)\\ &+ \theta_{4,5}\sin\left(\frac{4\pi}{365}t\right) + \theta_{4,6}\cos\left(\frac{6\pi}{365}t\right) + \theta_{4,7}\sin\left(\frac{6\pi}{365}t\right) + \theta_{4,8}\cos\left(\frac{8\pi}{365}t\right)\\ &+ \theta_{4,9}\sin\left(\frac{8\pi}{365}t\right) + \theta_{4,10}V_{\tau,3}(t) + \theta_{4,11}V_{\tau,3}^2(t)\\ &+ \theta_{4,12}V_{\tau,3}^3(t) + \theta_{4,13}V_{\tau,1}(t)V_{\tau,3}(t)\Bigg).\end{aligned} \tag{4.14}$$

We provide the estimates of all model parameters (i.e., $\boldsymbol{\theta}$'s) in Table 2. Moreover, we plot the spatial and temporal predictions and the actual realisations[11] for the year 2020 in Figure 8.

Overall, both spatial and temporal predictions capture the correct trends of fire occurrences. Specifically, in the spatial domain our model learns the urbanity features of the Twente region, which are highly relevant for chimney fires. Since the spatial prediction depends on the distribution of houses of different types, which can only be accessed as actual data, our model displays similar spatial patterns for different years. In the temporal domain, our model captures the periodic pattern on an annual basis and adds weather variable dependent information, which explains the noisier aspect of Figure 8(b) compared to Figure 8(a). A comparison to the areal unit model in Lu et al. (2021) and some ρ-tuning experiments will be specifically discussed in Section 4.6 and Section 6.2. In summary, given appropriate building information, our prediction model can detect areas with higher risks of chimney fires in Twente and, additionally, estimate the risk for specific days based on weather forecast.

4.5. *Confidence interval.* For risk prediction problems, it is important to quantify the uncertainty of the estimates. In the literature Baddeley et al. (2014) demonstrated asymptotic normality for their logistic regression estimators for stationary Gibbs point processes in space when the domain increases. Recently, Choiruddin, Coeurjolly and Letué (2018) proved

[11] Spatial realisation is smoothed by Gaussian kernels with a standard deviation of 1000 metres.



(a) spatial prediction (b) temporal prediction

(c) spatially smoothed pattern (d) daily counts

FIG. 8. *Spatial and temporal predictions* (*a*, *b*) *based on our point process model and actual realisations* (*c*, *d*) *for the year* 2020. *Shadows in* (*b*) *bound the* 95% *confidence intervals. The unit in spatial plots is metre*$^{-2}$; *the unit in temporal plots is day*$^{-1}$.

asymptotic normality also in an increasing-domain setting for spatial point process models that involve explanatory variables. In a companion paper (Lieshout and Lu (2022)), we study strong consistency and asymptotic normality for the logistic regression estimators we use above for spatiotemporal point processes in an infill asymptotic setting.

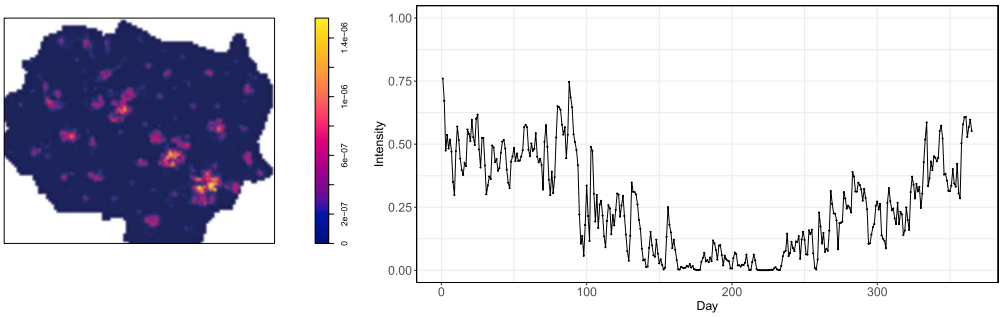
Specifically, under appropriate conditions we prove that the maximiser $\hat{\boldsymbol{\theta}}$ of (4.7), under the true value $\boldsymbol{\theta_0}$, is approximately normally distributed with mean $\boldsymbol{\theta_0}$ and covariance matrix $\boldsymbol{G}$, given by the inverse of the Godambe matrix (Godambe and Heyde (1987)). A plug-in estimator for $\boldsymbol{G}$ is

$$\hat{\boldsymbol{G}} = \left[\int_T \int_W \frac{\lambda_k(u,t;\hat{\boldsymbol{\theta}})\rho(u,t)}{\lambda_k(u,t;\hat{\boldsymbol{\theta}}) + \rho(u,t)} C_p(u,t)C_q(u,t)\,du\,dt\right]_{p,q=1}^{m}, \tag{4.15}$$

assuming that we have m parameters in $\boldsymbol{\theta}$. Approximate confidence intervals for model parameters are then readily obtained.

For chimney fire data modelling, the approximate numerical 95% confidence intervals of model parameters are listed in Table 2. The Delta method (Ver Hoef (2012)) can again be used to calculate approximate confidence intervals for predicted fire intensities. We visualize the temporal risk predictions for the year 2020 in Figure 8(b) as an example.

4.6. *Comparison to the areal unit model.* To compare the performance of the point process model to the areal unit model in Lu et al. (2021), we plot the spatial and temporal predictions for the year 2020 for the latter in Figure 9. Since the spatial predictions from the areal unit model are computed for 500 m × 500 m area boxes, we average them by the volume of these boxes to comply with the scaling level of the point process model.



(a) spatial predictions (b) temporal predictions

FIG. 9. *Spatial* (*a*) *and temporal* (*b*) *predictions based on the areal unit model in Lu et al.* (2021) *for the year* 2020. *The unit in the spatial plot is metre*$^{-2}$; *the unit in the temporal plot is day*$^{-1}$.

Generally, comparing Figure 9 to Figure 8, we see that both areal unit model and point process model capture similar spatial and temporal patterns from the fire data. However, the spatial predictions from the point process model by construction are smoother than those from the areal unit model, as we apply kernel smoothing in the former to obtain the density of a house type at a location, which helps establish spatially continuous predictions. The temporal predictions from the two models are almost identical for every day, because we assume that temporal variables at different times of a day remain invariant and the spatial aggregation is integrated out. Specifically, in terms of the total chimney fire risk predicted for the year 2020, the areal unit model predicts 92.86 and the point process model predicts 92.95, which are very close to each other. The observed actual count of 81 lies in the 95% confidence intervals of both models. However, the areal unit model overestimates many more fire risks for certain area boxes of big city centres (e.g., Enschede) and underestimates for rural area boxes.

To conclude, the point process model generates continuous hazard maps of chimney fires, which avoids the classic modifiable area unit data problem and enables to access fire risks at exact locations in Twente, thus is more user-friendly for fire services. The predictions do not depend on the scale of the area boxes, while only the smoothing bandwidth is inevitably inherited. Moreover, from a computational perspective the logistic regression estimation approach used to fit the point process model is more efficient and captures more accurate data patterns.

5. Model validation.

In this section we validate the independence assumption underlying our Poisson point process model by analyzing the second-order properties of the chimney fire data. We also validate the first-order structure of our model by visualizing the residuals.

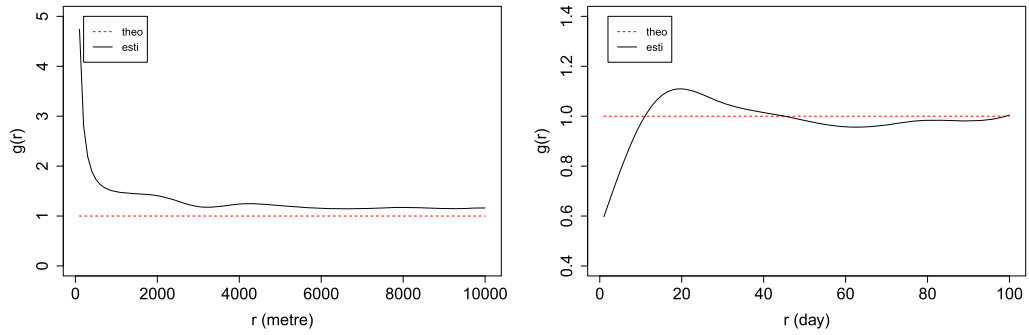
5.1. *Second-order analysis.*

To assess the validity of the Poisson assumption, we perform a second-order analysis on the fire data using two summary statistics: the pair correlation function and the K-function (see, e.g., Lieshout (2019)). The pair correlation function can be used to detect interpoint interactions for specific distances, whereas the K-function is a cumulative statistic that can be used to test the existence of point interactions in data.

5.1.1. *The pair correlation function.*

Consider a point process X defined on a space-time domain $W \times T \subset \mathbb{R}^2 \times \mathbb{R}$ that is locally finite. The second-order product density $\lambda^{(2)}(x, y)$ of the pair (x, y), where $x, y \in W \times T$, is the infinitesimal probability that X places points at x and y. The pair correlation function for the pair (x, y) is then defined as

$$g(x, y) = \frac{\lambda^{(2)}(x, y)}{\lambda(x)\lambda(y)}, \tag{5.1}$$



(a) spatial result (b) temporal result

FIG. 10. *Estimated (solid black line) and theoretical (dashed red line) pair correlation functions for the spatial and temporal projections (cf. Figure 1(b) and 2) of the chimney fire incidents during* 2004–2020 *using the intensity functions fitted by our Poisson point process model.*

where $\lambda(x)$ and $\lambda(y)$ denote the intensities at x and y, and we use the convention that $a/0 = 0$ for all $a \geq 0$. For visualisation purposes we consider pair correlation functions for the projections in space and time separately. Under appropriate stationarity and isotropy assumptions, in d dimensions (spatially, $d = 2$; temporally, $d = 1$), (5.1) is a function of $r = \|x - y\|$ only[12] and can be estimated by

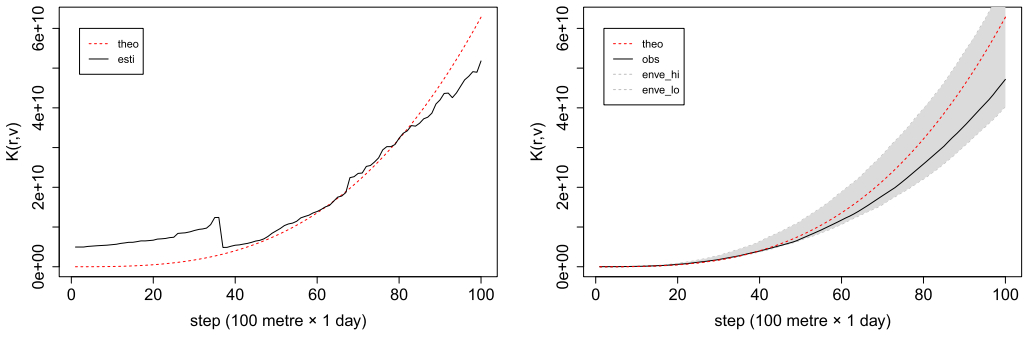
$$\hat{g}(r) = \frac{1}{s_d r^{d-1}} \sum_{x,y \in X}^{\neq} \frac{k_b(r - \|x - y\|)}{\lambda(x)\lambda(y)|W \cap W_{x-y}|}, \tag{5.2}$$

where s_d denotes the surface area of the unit sphere in $\mathbb{R}^d$, $\sum^{\neq}$ denotes the summation over all pairs of distinct points, $k_b(\cdot)$ is an one-dimensional smoothing kernel with bandwidth b, $1/|W \cap W_{x-y}|$ is the spatial edge correction factor (Ohser and Stoyan (1981)) (temporally, $1/|T \cap T_{x-y}|$) and $|A|$ is the volume of $A \subset \mathbb{R}^d$. If there is no interaction, the pair correlation function is 1. At short distances a value of $\hat{g}(r)$ larger than 1 suggests clustering in the sense that points are more likely to occur from one another, whereas a value smaller than 1 suggests inhibition, that is, points tend to mutually avoid each other. If the intensity function $\lambda(x)$ is unknown, a plug-in estimator can be used.

In practice, we employ (5.2) with Gaussian smoothing kernels as k_b and plug in the fitted intensity functions based on the data from all 17 years. We set the possible interaction intervals for the spatial and temporal domain to 10,000 metres and 100 days, respectively, and the smoothing bandwidths to 500 metres and 10 days. The estimated pair correlation functions are plotted in Figure 10. Overall, both the spatial and temporal results show a curve that converges to 1. At smaller distances there is some attraction between the points spatially. Temporally, possibly because of the precision of the observations (daily basis), there is some repulsion and a weak peak at an interval of about 20 days. The latter might also be due to underestimates of the intensity function on a specific pair of days (cf. equation (5.2)).

5.1.2. *The K-function.* Since the pair correlation function shows some point interactions at small distances and time intervals, we perform a joint space-time K-function test on the fire data. The K-function, also known as Ripley's reduced second moment function, of a stationary point process X is defined as the expected number of other points within a given

[12]Spatially, $\|x - y\|$ is the distance between two points; temporally, $\|x - y\|$ is the time interval between two points.



(a) inhomogeneous K function (b) Monte Carlo envelope

FIG. 11. *Estimated* (*solid black line*) *and theoretical* (*dashed red line*) *inhomogeneous* K*-functions for the space-time pattern of the chimney fire incidents during* 2004–2020 *using the intensity functions fitted by our Poisson point process model* (*a*). *Estimated* (*solid black line*) *and theoretical* (*dashed red line*) *inhomogeneous* K*-functions for the chimney fire data and the envelope of the estimated inhomogeneous* K*-functions* (*shadows*) *for the* 99 *Monto Carlo simulations of a Poisson point process with the fitted intensities, both using the intensity functions fitted by kernel estimators* (*b*).

distance of an arbitrary point divided by the intensities. In an inhomogeneous and space-time version, it takes the form of

$$K_{\text{inhom}}(r,v) = \frac{1}{|B|} E\left\{ \sum_{x\in X\cap B} \sum_{y\in X; y\neq x} \frac{1(\|x(u)-y(u)\|\le r, \|x(t)-y(t)\|\le v)}{\lambda(x)\lambda(y)} \right\}, \tag{5.3}$$

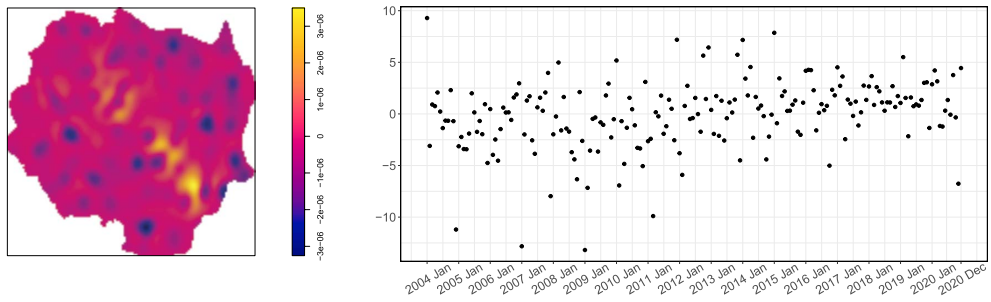
where the distance and time intervals $r, v \ge 0$, B denotes any Borel set, x and y are distinctive points with $x(u)$, $y(u)$ and $x(t)$, $y(t)$ representing their spatial and temporal coordinates and $1(\cdot)$ is the indicator function. Under appropriate weak stationarity assumptions (Gabriel and Diggle (2009)), a pointwise unbiased estimator of the K-function is

$$\hat{K}_{\text{inhom}}(r,v) = \sum_{x\in X} \sum_{y\in X; y\neq x} \frac{1(\|x(u)-y(u)\|\le r, \|x(t)-y(t)\|\le v)}{\lambda(x)\lambda(y)|(W\times T)\cap(W\times T)_{x-y}|} \tag{5.4}$$

with notation as in the previous section. In the absence of interaction, $K_{\text{inhom}}(r,v) = 2\pi r^2 v$. Moreover, when the intensity function $\lambda(x)$ is unknown, a plug-in estimator can be used.

We compute the inhomogeneous K-function over the chimney fire data from all 17 years. To obtain an elegant view, we set the testing sequences of spatial and temporal distances to 10,000 metres and 100 days and divide both of them into 100 step size pairs of 100 metres and one day. We calculate the K-function, using the fitted intensity functions by our Poisson point process model to estimate λ in (5.4), and plot the result in Figure 11(a). As for the pair correlation function, we find some evidence of clustering. To test whether the interaction is statistically significant, we implement a Monte Carlo test (Baddeley, Rubak and Turner (2015)). We generate 99 realisations of a Poisson point process with the fitted intensities (i.e., by our Poisson point process model) and compute their K-functions and that of the actual chimney fire data but now using kernel estimators to estimate the intensity function λ in (5.4). We plot the local envelopes in Figure 11(b). As the empirical K-function of the actual chimney fire data lies completely within the envelope, there is no reason to look beyond a Poisson model.

5.2. *Residual analysis*. In order to further verify the validity of our point process model, we perform a residual analysis to check the remaining structure after the modelling of a Poisson point process with specific forms defined in (4.1), (4.2) and (4.11)–(4.14). We fit the



(a) spatial residuals (b) temporal residuals (monthly counts)

FIG. 12. *Spatial* (*a*) *and temporal* (*b*) *residuals based on the chimney fire data during* 2004–2020. *The unit in the spatial plot is metre*$^{-2}$; *the unit in the temporal plot is month*$^{-1}$.

model using the data from all 17 years and compute spatial and temporal residuals separately. As suggested in Baddeley et al. (2005), the spatial residuals are defined as the difference between the smoothed fitted fire risk and the smoothed actual realisations using Gaussian kernels with a standard deviation of 1000 metres. Temporally, we show the residuals of the difference between the predicted and actual monthly counts.

The results of the residual analysis are plotted in Figure 12. In the spatial domain, no specific pattern is shown, except that our Poisson point process model still overestimates for big city centres (e.g., Enschede and Hengelo), whereas it underestimates a bit for small city centres. This could be expected, as our model depends only on the density of houses of various types, and we observe a saturation pattern of chimney fire occurrences when the density of a house type increases (cf. Figure 6). A possible solution for that is to apply a piecewise function, instead of the linear function, to model the relation between chimney fire occurrences and the density of a house type. In the temporal domain, no particular pattern is apparent.

6. Discussion. In this section we discuss the data-driven modelling process designed for chimney fire prediction, which can be generalized to similar fire types. We also investigate the influence of tuning the dummy intensity ρ to important regions and times in logistic regression estimation on model fitting.

6.1. *Modelling process.* We propose a two-step data-driven modelling process for the risk prediction of chimney fires.

In the first step, we combine machine learning and statistics, where the former is utilized to help select the most important environmental variables and the latter appropriately approximates and explains the relations between explanatory variables and fire incidence. We employ parametric functions to model such relations and fit them using logistic regression estimation so that the influence of a variable on chimney fires can be analyzed by its coefficient in the function model. This combination offers significant advantages. From the machine learning side: (i) compared to naive statistics used for the selection of explanatory variables, random forests can detect both linear and nonlinear correlation between a candidate variable and chimney fire occurrences; (ii) random forests consider the dependence among candidate variables nonparametrically; (iii) through applying the conditional permutation importance techniques, instead of the traditional ones, the bias toward correlated variables is suppressed; (iv) machine learning algorithms can handle the selection with a large number of putative variables, while statistical methods often behave unstable or even fail in such cases. From the statistics side: (i) classic statistical methods for regression modelling with specific mathematical expressions are more interpretable and based on natural assumptions; (ii) statistical

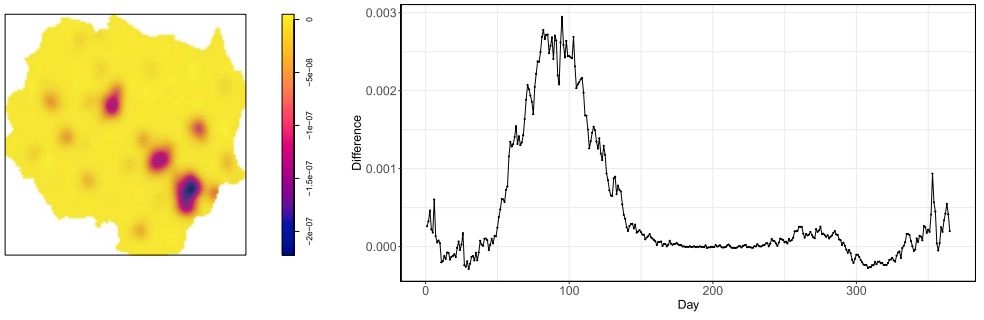
methods also allow for significance testing for detailed analyses and uncertainty quantification. However, due to the fact that the environmental variables relevant to our chimney fire data are not available for every point in space and time, we perform variable selection and model fitting at different levels of data aggregation, which has the disadvantage that the selected explanatory variables might not be consistent. We leave it as an important direction in future work.

In the second step, we use summary statistics, the pair correlation function and the K-function, to detect point interactions. We also perform a residual analysis to validate the model structure. Although in our study a Poisson point process model is sufficient for chimney fire prediction, a hierarchical model (Banerjee, Carlin and Gelfand (2015)), where random effects are incorporated to deal with latent variables, might be more appropriate for other fire types.

6.2. *ρ tuning in logistic regression estimation.* Recall that the role of the dummy point process in logistic regression estimation is to estimate the integral in (4.4). In Baddeley et al. (2014), a rule of thumb was proposed that the intensity function ρ of the dummy point process should be at least four times the intensity function λ of the point process to be estimated; additionally, a data-driven selection of ρ was suggested.

In Section 4.4 we tune ρ to concentrate on the urban areas and cold seasons by (4.10) in order to efficiently obtain more precise estimates for model parameters provided a limited number of dummy points. Here we perform two experiments to verify the influence of this ρ tuning operation in the spatial and temporal domains separately. Specifically, we fit our fire prediction model on the data from all 17 years using logistic regression estimation with and without tuning the dummy intensity ρ and plot the difference of the fitted values. For comparison, we set the expectations of the number of dummy points as equal in both cases. In addition, since the realisations of the dummy point process generated in different runs contain randomness, we perform each experiment 60 times and plot the averaged spatial and temporal differences to reduce the randomness.

The results show that, under the premise of the same amount of calculation, concentrating on important regions and times enables the model to obtain more accurate patterns of chimney fires. Spatially, according to Figure 13(a), the bias caused by the saturation effect shown in Figure 6 is reduced. Temporally, plotted in Figure 13(b), the fitted fire intensities increase in March and April, which are known as the weather tipping seasons and are found more likely to catch chimney fires in School (2018). To summarize, tuning the dummy intensity ρ to regions and times that are salient for fire risk is helpful to estimate the risk more efficiently and capture important underlying patterns.



(a) spatial difference (b) temporal difference

FIG. 13. *Spatial (a) and temporal (b) difference of the fitted intensities using tuned ρ compared to uniform ρ with an equal expected number of dummy points in logistic regression estimation (averaged on 60 random runs). The unit in the spatial plot is metre^{-2}; the unit in the temporal plot is day^{-1}.*

7. Conclusion. In this paper we proposed a data-driven modelling process for chimney fire risk prediction, which can be generalized to studies on other fire types. First, we applied random forests and permutation importance techniques to identify the explanatory variables nonparametrically from a large number of environmental variables. From a practical perspective, our results indicate that prewar detached or semidetached houses run a higher risk of chimney fires, and, therefore, public awareness campaigns should preferentially target the owners of such houses. Second, based on the observed relations between the selected variables and chimney fire occurrences, we defined a Poisson point process model to learn the observed spatiotemporal point pattern in order to predict fire risk. We applied logistic regression estimation for model fitting and provided infill asymptotic confidence intervals. Additionally, to validate our model assumptions we performed a second-order analysis and a residual study. Last but not least, we reviewed our modelling process and showed the advantages of the tuned intensity function of the dummy point process in logistic regression estimation.

For future work the first direction would be to extend the data-driven modelling process of chimney fire risk prediction to similar fire types to test its utility. Additionally, due to data limitations, we performed our nonparametric variable selection on areal unit data, which may not return consistent results on point process data. Hence, it would also be interesting to conduct a study into data-driven variable selection for point process modelling when environmental variables are available for every point in space and time.

Acknowledgments. We thank the user committee for their valuable input. We also thank Emiel Borggreve, Niels Peters and Etienne Mulder from the Twente Fire Brigade for their help with data cleaning and preprocessing. Marie-Colette van Lieshout is also affiliated with University of Twente. Maurits de graaf is also affiliated with Centrum Winskunde & Informatica. Furthermore, when the project was conducted, Maurits de Graaf was affiliated with University of Twente as well.

Funding. This work was funded by the Dutch Research Council (NWO) for the project "Data Driven Risk Management for Fire Services" (18004).

SUPPLEMENTARY MATERIAL

Supplement to "Data-driven chimney fire risk prediction using machine learning and point process tools" (DOI: 10.1214/23-AOAS1752SUPP; .pdf). Supplement A: Results for Extra Variable Importance Experiments. In the selection of explanatory variables, to test the influence of different hyperparameter settings of random forests (i.e., the number of randomly sampled input variables at each tree node, *mtry*, and the random seed) on variable importance, we performed experiments that construct random forests with multiple configurations. Moreover, to assess the influence of small variations in weather among different parts of the Twente region on variable importance, we performed another analysis on the temporal data from two neighbouring weather stations. We provide the results obtained from these experiments in this supplement.

REFERENCES

ALTMANN, A., TOLOŞI, L., SANDER, O. and LENGAUER, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics* **26** 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

BADDELEY, A., COEURJOLLY, J.-F., RUBAK, E. and WAAGEPETERSEN, R. (2014). Logistic regression for spatial Gibbs point processes. *Biometrika* **101** 377–392. MR3215354 https://doi.org/10.1093/biomet/ast060

BADDELEY, A., RUBAK, E. and TURNER, R. (2015). *Spatial Point Patterns*: *Methodology and Applications with R*. CRC Press.

BADDELEY, A. and TURNER, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Aust. N. Z. J. Stat.* **42** 283–322. MR1794056 https://doi.org/10.1111/1467-842X.00128

BADDELEY, A., TURNER, R., MØLLER, J. and HAZELTON, M. (2005). Residual analysis for spatial point processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 617–666. With discussion and a reply by the authors. MR2210685 https://doi.org/10.1111/j.1467-9868.2005.00519.x

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. MR3362184

BOUBETA, M., LOMBARDÍA, M. J., MAREY-PÉREZ, M. F. and MORALES, D. (2015). Prediction of forest fires occurrences with area-level Poisson mixed models. *J. Environ. Manag.* **154** 151–158. https://doi.org/10.1016/j.jenvman.2015.02.009

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

CHOIRUDDIN, A., COEURJOLLY, J.-F. and LETUÉ, F. (2018). Convex and non-convex regularization methods for spatial point processes intensity estimation. *Electron. J. Stat.* **12** 1210–1255. MR3780731 https://doi.org/10.1214/18-EJS1408

CHOIRUDDIN, A., COEURJOLLY, J.-F. and WAAGEPETERSEN, R. (2021). Information criteria for inhomogeneous spatial point processes. *Aust. N. Z. J. Stat.* **63** 119–143. MR4296145 https://doi.org/10.1111/anzs.12327

CLEVELAND, W. S., GROSSE, E. and SHYU, W. M. (1992). Local regression models. In *Statistical Models in S* 8, 1st ed. Wadsworth & Brooks/Cole.

COSTAFREDA-AUMEDES, S., COMAS, C. and VEGA-GARCIA, C. (2016). Spatio–temporal configurations of human-caused fires in Spain through point patterns. *Forests* **7** 185.

DALEY, D. J. and VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes. Volume II*, 2nd ed. Springer, New York. MR2371524 https://doi.org/10.1007/978-0-387-49835-5

DEBEER, D. and STROBL, C. (2020). Conditional permutation importance revisited. *BMC Bioinform.* **21** 307. https://doi.org/10.1186/s12859-020-03622-2

FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** 342–368. MR0773172 https://doi.org/10.1214/aos/1176346597

GABRIEL, E. and DIGGLE, P. J. (2009). Second-order analysis of inhomogeneous spatio–temporal point process data. *Stat. Neerl.* **63** 43–51. MR2656916 https://doi.org/10.1111/j.1467-9574.2008.00407.x

GODAMBE, V. P. and HEYDE, C. C. (1987). Quasi-likelihood and optimal estimation. *Int. Stat. Rev.* **55** 231–244. MR0963141 https://doi.org/10.2307/1403403

HERING, A. S., BELL, C. L. and GENTON, M. G. (2009). Modeling spatio–temporal wildfire ignition point patterns. *Environ. Ecol. Stat.* **16** 225–250. MR2668734 https://doi.org/10.1007/s10651-007-0080-6

HOTHORN, T., BUEHLMANN, P., DUDOIT, S., MOLINARO, A. and VAN DER LAAN, M. (2006). Survival ensembles. *Biostatistics* **7** 355–373.

HOTHORN, T., HORNIK, K. and ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Statist.* **15** 651–674. MR2291267 https://doi.org/10.1198/106186006X133933

JAIN, P., COOGAN, S., SUBRAMANIAN, S., CROWLEY, M., TAYLOR, S. W. and FLANNIGAN, M. (2020). A review of machine learning applications in wildfire science and management. *Environ. Rev.* **28** 478–505.

JUAN VERDOY, P. (2021). Enhancing the SPDE modeling of spatial point processes with INLA, applied to wildfires. Choosing the best mesh for each database. *Comm. Statist. Simulation Comput.* **50** 2990–3030. MR4322119 https://doi.org/10.1080/03610918.2019.1618473

KOH, J., PIMONT, F., DUPUY, J.-L. and OPITZ, T. (2023). Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. *Ann. Appl. Stat.* **17** 560–582. MR4539044 https://doi.org/10.1214/22-aoas1642

LIESHOUT, M. N. M. VAN (2019). *Theory of Spatial Statistics*: *A Concise Introduction*. CRC Press, Boca Raton, FL.

LIESHOUT, M. N. M. VAN and LU, C. (2022). Infill asymptotics for logistic regression estimators for spatio–temporal point processes. arXiv:2208.12080.

LU, C., LIESHOUT, M. N. M. VAN, GRAAF, M. DE and VISSCHER, P. (2021). Chimney fire prediction based on explanatory environmental variables. In *The* 63*rd ISI World Statistics Congress* 288–291.

LU, C., LIESHOUT, M. N. M. VAN, GRAAF, M. DE and VISSCHER, P. (2023). Supplement to "Data-driven chimney fire risk prediction using machine learning and point process tools." https://doi.org/10.1214/23-AOAS1752SUPP

MALIK, A., RAO, M. R., PUPPALA, N., KOOURI, P., ANIL, V., THOTA, K., LIU, Q., CHIAO, S. and GAO, J. (2021). Data-driven wildfire risk prediction in northern California. *Atmosphere* **12** 109.

MCCULLAGH, P. and NELDER, J. A. (2019). *Generalized Linear Models*, 2nd ed. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

MØLLER, J. and DÍAZ-AVALOS, C. (2010). Structured spatio–temporal shot-noise Cox point process models, with a view to modelling forest fires. *Scand. J. Stat.* **37** 2–25. MR2675937 https://doi.org/10.1111/j.1467-9469.2009.00670.x

NVBR (2010). *De brandweer over Morgen*. Nederlandse Vereniging voor Brandweer en Rampenbestrijding, Arnhem.

OHSER, J. and STOYAN, D. (1981). On the second-order and orientation analysis of planar stationary point processes. *Biom. J.* **23** 523–533. MR0635658 https://doi.org/10.1002/bimj.4710230602

PEREIRA, P., TURKMAN, K., TURKMAN, A., SÁ, A. and PEREIRA, J. (2013). Quantification of annual wildfire risk; a spatio–temporal point process approach. *Statistica* **73** 55–68.

PIMONT, F., FARGEON, H., OPITZ, T., RUFFAULT, J., BARBERO, R., MARTIN-STPAUL, N., RIGOLOT, E., RIVIÉRE, M. and DUPUY, J.-L. (2021). Prediction of regional wildfire activity in the probabilistic Bayesian framework of Firelihood. *Ecol. Appl.* **31** e02316. https://doi.org/10.1002/eap.2316

PREISLER, H., BRILLINGER, D., BURGAN, R. and BENOIT, J. (2004). Probability based models for estimation of wildfire risk. *Int. J. Wildland Fire* **13** 133–142.

RODRIGUES, M. and DE LA RIVA, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environ. Model. Softw.* **57** 192–201.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SAKR, G. E., ELHAJJ, I. H., MITRI, G. and WEJINYA, U. C. (2010). Artificial intelligence for forest fire prediction. In 2010 *IEEE/ASME International Conference on Advanced Intelligent Mechatronics* 1311–1316.

SATIR, O., BERBEROGLU, S. and DONMEZ, C. (2016). Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. *Geomatics, Natural Hazards and Risk* **7** 1645–1658.

SCHONLAU, M. and ZOU, R. (2020). The random forest algorithm for statistical learning. *Stata J.* **20** 3–29.

SCHOOL, M. L. (2018). A log-Gaussian Cox process for predicting chinmey fires at fire department Twente. Master's thesis, Univ. Twente.

SERRA, L., SAEZ, M., MATEU, J., VARGA, D., JUAN, P., DÍAZ-ÁVALOS, C. and RUE, H. (2014). Spatio–temporal log-Gaussian Cox processes for modelling wildfire occurrence: The case of Catalonia, 1994–2008. *Environ. Ecol. Stat.* **21** 531–563. MR3248538 https://doi.org/10.1007/s10651-013-0267-y

SILVAPULLE, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *J. Roy. Statist. Soc. Ser. B* **43** 310–313. MR0637943

STOJANOVA, D., KOBLER, A., OGRINC, P., ŽENKO, B. and DŽEROSKI, S. (2012). Estimating the risk of fire outbreaks in the natural environment. *Data Min. Knowl. Discov.* **24** 411–442.

STROBL, C., BOULESTEIX, A., ZEILEIS, A. and HOTHORN, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **8**.

STROBL, C., BOULESTEIX, A.-L., KNEIB, T., AUGUSTIN, T. and ZEILEIS, A. (2008). Conditional variable importance for random forests. *BMC Bioinform.* **9** 307. https://doi.org/10.1186/1471-2105-9-307

STROBL, C., HOTHORN, T. and ZEILEIS, A. (2009). Party on! *R J.* **1** 14–17.

STROBL, C. and ZEILEIS, A. (2008). Danger: High power!—exploring the statistical properties of a test for random forest variable importance. In *COMPSTAT* 2008—*Proceedings in Computational Statistics* 59–66. Physica-Verlag/Springer, Heidelberg. MR2509600

THURMAN, A. L. and ZHU, J. (2014). Variable selection for spatial Poisson point processes via a regularization method. *Stat. Methodol.* **17** 113–125. MR3133589 https://doi.org/10.1016/j.stamet.2013.08.001

TURNER, R. (2009). Point pattern of forest fire locations. *Environ. Ecol. Stat.* **16** 197–223. MR2668733 https://doi.org/10.1007/s10651-007-0085-1

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York. https://doi.org/10.1007/978-0-387-21706-2

VER HOEF, J. M. (2012). Who invented the delta method? *Amer. Statist.* **66** 124–127. MR2968009 https://doi.org/10.1080/00031305.2012.687494

WONGVIBULSIN, S., WU, K. and ZEGER, S. (2019). Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med. Res. Methodol.* **20**.

XU, H. and SCHOENBERG, F. P. (2011). Point process modeling of wildfire hazard in Los Angeles County, California. *Ann. Appl. Stat.* **5** 684–704. MR2840171 https://doi.org/10.1214/10-AOAS401

YANG, J., WEISBERG, P., DILTS, T., LOUDERMILK, L., SCHELLER, R., STANTON, A. and SKINNER, C. (2015). Predicting wildfire occurrence distribution with spatial point process models and its uncertainty assessment: A case study in the Lake Tahoe Basin, USA. *Int. J. Wildland Fire* **24** 390.

YE, R. (2011). Prediction of forest fires with Poisson models. *Can. J. For. Res.* **27** 1685–1694.

YUE, Y. and LOH, J. M. (2015). Variable selection for inhomogeneous spatial point process models. *Canad. J. Statist.* **43** 288–305. MR3353384 https://doi.org/10.1002/cjs.11244