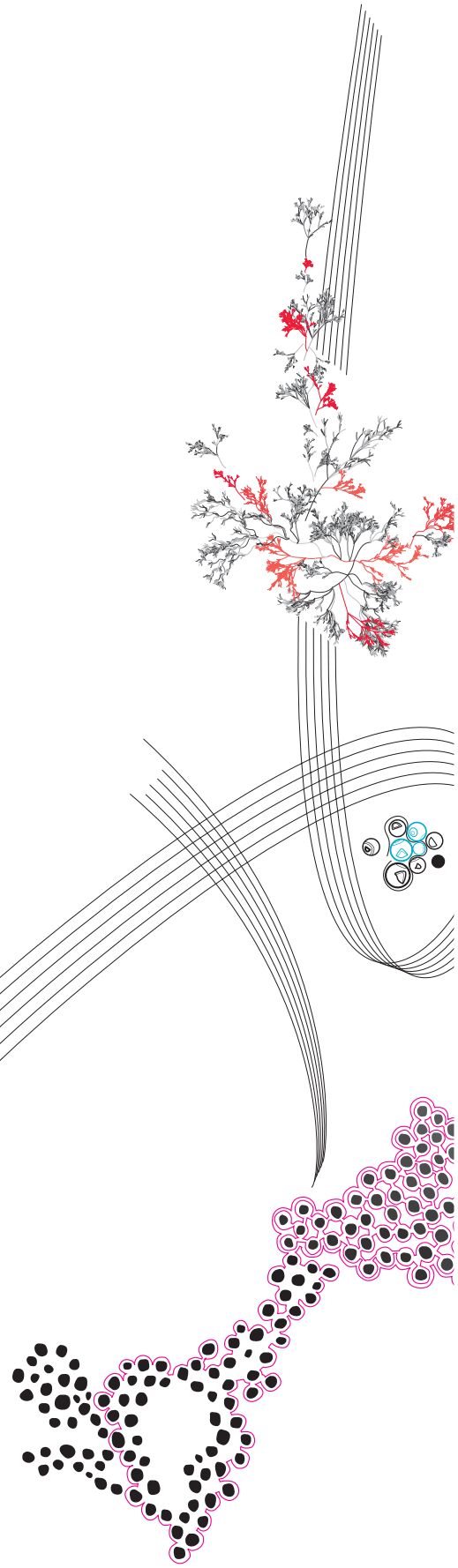


BSc Thesis Applied Mathematics

# Chimney fire prediction in IJsselland

Lars Noah van Kasteren

Supervisors: M.C. van Lieshout, C. Lu

January, 2024

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science

UNIVERSITY OF TWENTE.

## Preface

First of all, I want to send a lot of thanks to Niels Peters and Paul Visscher with their fast response and help when I had a problem with my data. Without their fast help, there would not be a correct paper here, if there would be one at all. Many, many thanks for that.
Secondly, I would like to thank my supervisors, Marie-Colette van Lieshout and Changqing Lu. Their help to my bachelors assignment was much needed and given a lot. Be it with explaining concepts, finding mistakes in my data or just with ideas on how to go further, it was of great value.
Furthermore, I would like to thank my family. Especially the last few weeks have been very busy with this, with a few obstacles appearing. I would specifically like to thank my mother for the litres of hot chocolate and hours of talking about the thesis and the extra proofreading.

# Chimney fire prediction in IJsselland

L. N. van Kasteren*

January, 2024

**Abstract**

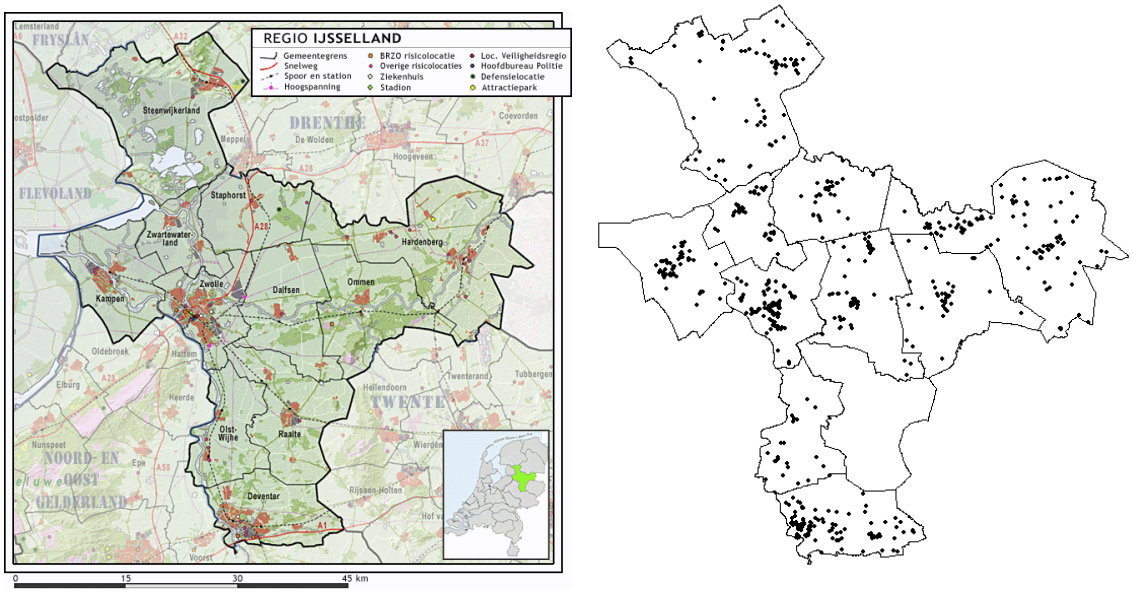
In this paper, we develop a combined machine learning and statistical modelling process to predict chimney fire risk in IJsselland. We first use random forests and permutation importance techniques to find the explanatory variables. Secondly, we design a Poisson point process model and use logistic regression estimation to estimate the parameters. Afterwards, we verify the results using residuals. In all phases, we compare the results to the results from the Twente region, which is very similar to IJsselland. All the modelling is done on data collected by the Twente and IJsselland Fire Brigade.

*Keywords*: Chimney fire, random forest, permutation importance, Poisson point process, logistic regression estimation.

## 1 Introduction

Averaged over the whole year, there are almost 5 chimney fires per day in the Netherlands [1]. This is a large number, especially considering most of those are in the same few months in the winter, which means the average over the winter months would be even higher. Therefore, it is of great importance to investigate how the fire brigade could predict, prevent and prepare for those chimney fires. Research has been done on the safety region Twente, in which a prediction model was made for that region [8]. The goal of this research is to see whether the approach for the Twente region also extends to the IJsselland region. Similar to that research, we will first use machine learning with random forests [5] and permutation importance techniques [11][6], to find out which variables would be important for the prediction model. Afterwards, we will use the most important variables to fit a generalized linear Poisson model, using logistic regression estimation [3] on the spatio-temporal domain our research is based on, namely the IJsselland region and the years 2011-2021. This will give optimal orders for the prediction function, as well as confidence intervals for the estimator values. With the fitted model, we will run a test prediction on the last year of our data, to compare this to the actual results. The temporal predictions will include confidence intervals as well. After that, we do a residual analysis to further validate our model. Finally, we give a conclusion and some suggestions for future research.

*Email: l.n.vankasteren@student.utwente.nl, student number: s2348330




(A) IJsselland municipalities (B) Place of the fires

FIGURE 1: *Map of the IJsselland region, with their municipalities and cities (a) and the places of the chimney fires (b).*

# 2 Data

## 2.1 Data collection

Via the contact person from the Twente Fire Brigade[1], data was received on all the reported chimney fires in the IJsselland region between January 1st, 2011, and December 31st, 2021. A map of IJsselland can be found in Figure 1a.
In total, the data consisted of a total of 641 different chimney fires, for which the date, time, location and some information of the circumstances of the fire are reported. Sadly, for some of the chimney fires, the buildings do not exist anymore, or the chimney fire reports have missing data in which case their locations cannot be correctly identified. This means that the total number of actual chimney fires comes down to 599 different incidents. The locations of the chimney fires can be found in Figure 1b.
Looking at the figures, we see that the chimney fires are not spread homogeneously over the whole region. Figure 1b shows that chimney fires occur more often in the more densely populated areas, such as the cities Zwolle, Kampen and Deventer.
Additionally, we look at the chimney fires for each month during the given period (Figure 2). As can be seen, the distribution of this is periodic, with more chimney fires happening in the winter months than in the summer months. This is something that could be expected, since, in general, chimneys will be used more often in the winter months.
As explained in [8], there will be many different environmental variables that could have an influence on the occurrence of chimney fires. For the places where they could occur, Figure 1 would suggest that the population density or the urbanity could have a great influence on this. Building types might be another influence on the heterogeneity. Since neighbouring buildings tend to have similar building types, that could explain the clustering as well. For the temporal factors, both seasonal and weather conditions, such as temperature or wind

[1]Paul Visscher, p.visscher@brandweertwente.nl

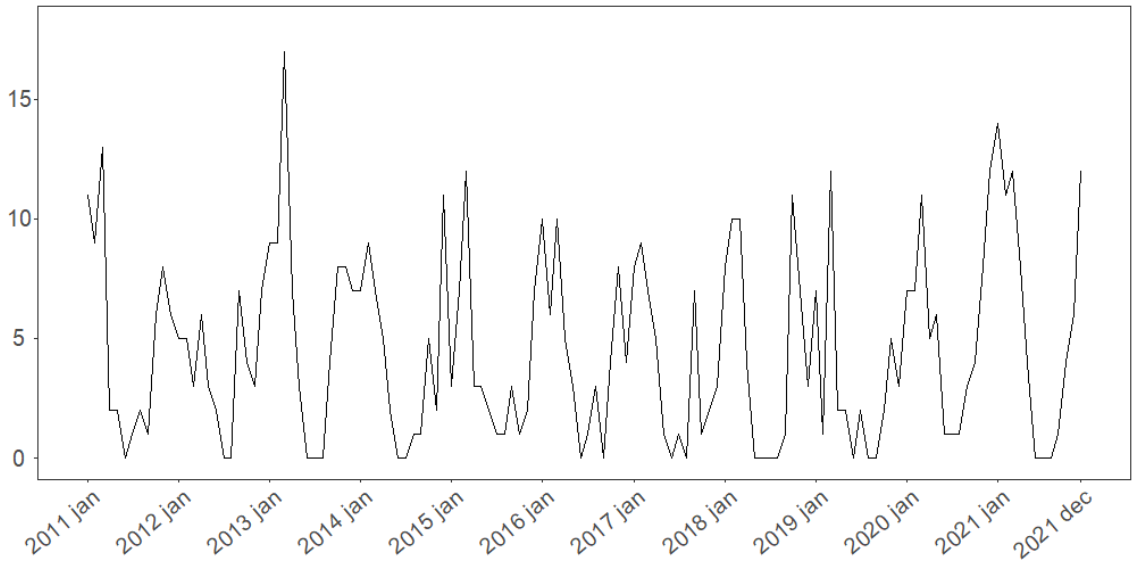



FIGURE 2: *Monthly counts of chimney fires between 2011 and 2021.*

speed, could influence the risk. Chimney types, the existence of chimneys in houses or cleaning of chimneys would supposedly be very good explanatory factors. However, there is no information accessible on those types of factors.
A list of putative explanatory variables can be found in Table 1. Those potential variables were suggested by the Twente Fire Brigade. The variables are very similar to the variables used in the Twente research [8]. Two of the potential variables from that research, the visibility and the presence of a town, are however not recorded for the IJsselland region, so they will not be considered in this research. Some variables, such as population and urbanity, have data that is calculated over 8093 $500m \times 500m$ boxes[2]. Other variables, such as the house build year, are recorded for precise locations, namely that specific house itself. The temporal data consists of daily observations from the Heino weather station. Its location can be found at `https://www.knmi.nl`.

## 2.2 Data pre-processing

To be able to use the data, we have to pre-process the given data. Firstly, since some variables have data that consists of single values (such as building information) and others consist of data from multiple years (such as population information), the average will be taken over the latter such that we have a single value for every spatial variable. This is needed to make sure that the variables are treated similarly during the importance search. Furthermore, the buildings that are not currently in use are filtered out, and we assume that buildings keep their functions during the whole period of interest. Finally, data from leap days in the leap years are excluded from both the fires as from the weather data, to keep the length of the year consistent for the predictions. For the future risk prediction, the prediction from February 28th will be used for February 29th in the years that this would be necessary.

[2]Determined by Centraal Bureau voor de Statistiek (CBS), see column '**Source**' in Table 1

TABLE 1: *Possible explanatory variables with their abbreviations, descriptions and sources.*

| Variable | Abbreviation | Description | Source[3] |
|---|---|---|---|
| $V_{\sigma,1}$ | House | The total number of houses | NIPV |
| $V_{\sigma,2}$ | House_indu | The total number of houses with an industrial function | NIPV |
| $V_{\sigma,3}$ | House_hotl | The total number of houses with a hotel function | NIPV |
| $V_{\sigma,4}$ | House_resi | The total number of houses with a residential function | NIPV |
| $V_{\sigma,5}$ | House_20 | The total number of houses built before 1920 | NIPV |
| $V_{\sigma,6}$ | House_2045 | The total number of houses built between 1920 and 1945 | NIPV |
| $V_{\sigma,7}$ | House_4570 | The total number of houses built between 1945 and 1970 | NIPV |
| $V_{\sigma,8}$ | House_7080 | The total number of houses built between 1970 and 1980 | NIPV |
| $V_{\sigma,9}$ | House_8090 | The total number of houses built between 1980 and 1990 | NIPV |
| $V_{\sigma,10}$ | House_90 | The total number of houses built after 1990 | NIPV |
| $V_{\sigma,11}$ | House_frsd | The number of free standing houses, detached or semi-detached | NIPV |
| $V_{\sigma,12}$ | Resid | The total number of residents | CBS |
| $V_{\sigma,13}$ | Resid_015 | The total number of residents with an age below 15 | CBS |
| $V_{\sigma,14}$ | Resid_1525 | The total number of residents with an age between 15 and 25 | CBS |
| $V_{\sigma,15}$ | Resid_2545 | The total number of residents with an age between 25 and 45 | CBS |
| $V_{\sigma,16}$ | Resid_4565 | The total number of residents with an age between 45 and 65 | CBS |
| $V_{\sigma,17}$ | Resid_65 | The total number of residents with an age of 65 or higher | CBS |
| $V_{\sigma,18}$ | Man | The total number of male residents | CBS |
| $V_{\sigma,19}$ | Woman | The total number of female residents | CBS |
| $V_{\sigma,20}$ | Address | The total number of addresses in the neighbourhood | CBS |
| $V_{\sigma,21}$ | Urbanity[4] | The urbanity of the neighbourhood | CBS |
| $V_{\tau,1}$ | WindSpeed | Daily average wind speed (km/h) | KNMI |
| $V_{\tau,2}$ | Temperature | Daily average temperature (°C) | KNMI |
| $V_{\tau,3}$ | WindChill | Daily average windchill[5] (°C) | |
| $V_{\tau,4}$ | Sunshine | Daily total sunshine duration (h) | KNMI |

Moreover, since some spatial variables are known for precise coordinates, whereas others are only known for areal units (the $500m \times 500m$ boxes), we use kernel smoothing. This will especially be needed for the point process modelling phase in Section 4, such that we get a (smoothed) value for the variables for every location in the IJsselland region. For temporal variables, we assume that their values do not change during the day and will be equal to the daily averages. Therefore, our temporal variables will depend on the date only, and not on the time on a specific day. From Table 1, the following notation will be used in the remainder of this paper:

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

with $(u, t)$ the location-time combination in the spatio-temporal domain.

## 3 Selection of explanatory variables

Not all the variables in Table 1 will be needed for the predictions; some will not have any influence on the chimney fire occurrence, others may be mutually dependent. To find the most important variables, we will use random forests [5] and conditional permutation techniques for variables selection.

## 3.1 Methods

### 3.1.1 Random forests

Random forests [5] are widely used in machine learning. They usually consist of enormous amounts of decision trees, where all those trees are generated on a subset of the data by random sampling with replacement from subsets of the list of possible variables. Using the random forests, a computer can be "trained" to fit certain explanatory variables (see Table 1) to the response variable, in this case the chimney fires.
Consider a dataset $D$ with $n$ observations and $m$ explanatory variables. Let $x_i^{(j)}$ and $y_i$ with $i = 1, ..., n$ and $j = 1, ..., m$ denote the j-th explanatory variable and the response variable, respectively, for the $i$-th observation, with $\boldsymbol{x}_i$ the vector of all the $x_i^{(j)}$.
As said before, a random forest consists of many (say, in this case, $E$) decision trees. To generate a tree $e$ for the forest, at each node of the tree, a (predetermined) number of explanatory variables are selected randomly from all possible variables. The subset of $D$ that consists of those variables, we call $B(e)$ (the bag). One of those variables, say $x^{(j)}$, is used to split the node into two subsets, $B_1(e)$ and $B_2(e)$, in such a way that the residual sum of squares

$$\sum_{(\boldsymbol{x}_i, y_i) \in B_1(e)} (y_i - \bar{y}_{B_1(e)})^2 + \sum_{(\boldsymbol{x}_i, y_i) \in B_2(e)} (y_i - \bar{y}_{B_2(e)})^2 \tag{1}$$

is minimized. Here $\bar{y}_{B_1(e)}$ and $\bar{y}_{B_2(e)}$ are the average of $y_i$ over their corresponding subsets. The node splitting keeps going until $e$ satisfies certain conditions, such as a certain depth. This happens for all the trees in the forest (the amount of trees in the forest is predetermined).

### 3.1.2 Permutation importance

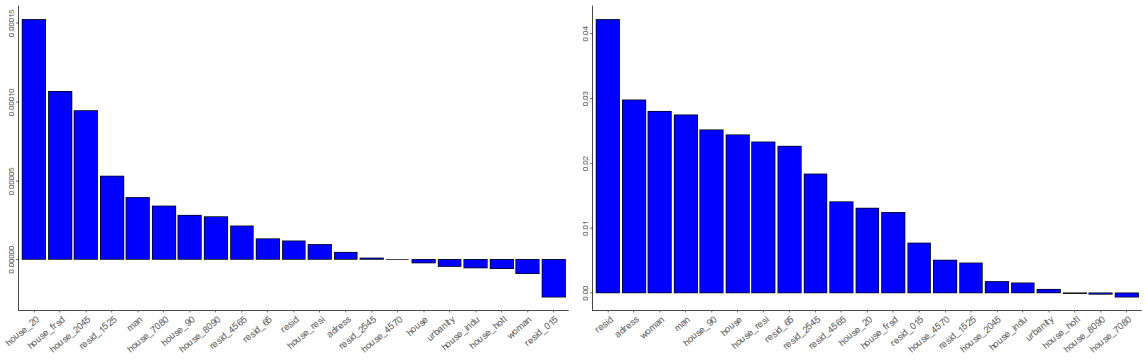
After creating the random forest, we must determine the important explanatory variables. As the name suggests, this is done by permutations.
To measure the importance of a variable $x^{(j)}$ in tree $e$, we permute the values of the observations over the out-of-bag observations of the tree, denoted as $oB(e)$. Only the $x^{(j)}$ values change. We denote the permutation as $\pi_j$. The value of $x_i^{(j)}$ then changes to $x_{\pi_j(i)}^{(j)}$. The increase of the prediction error is

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

Here $\hat{y}_i$ is the prediction in tree $e$ for observation $(\boldsymbol{x}_i, y_i)$ from the original explanatory variables, and $\hat{y}_{i,\pi_j}$ the prediction using the permuted variable. Finally, to determine the importance of $x^{(j)}$ over the whole forest, the average increase in prediction error over all trees is taken, so

$$\frac{\sum_e I(x^{(j)}, e)}{E}, \tag{3}$$

where the higher the prediction error, the more "important" the variable is.
There is however one problem with this way of computing the importance: explanatory variables that are correlated, will influence the results, even if they might not be of real importance for the response variable. One variable might not have any influence, but still get a high score because it is correlated with a variable that has a lot of influence.



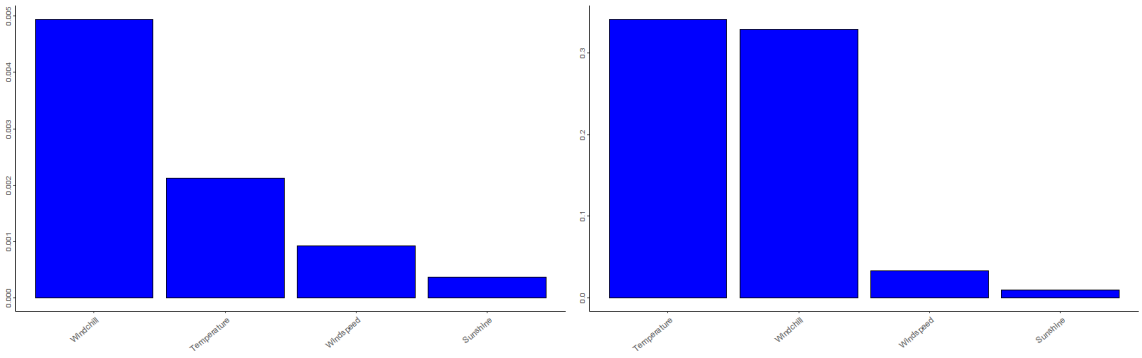
(A) Conditional permutation (B) Non-conditional permutation

FIGURE 3: *The importance (increase in prediction error) (y-axis) for spatial variables using conditional (A) and non-conditional (B) permutation techniques.*

Therefore, we will use the modified version, called the conditional permutation importance [6]. Instead of picking a permutation from the complete subset $oB(e)$ of $D$, the subset is partitioned into different subsets in which each of them shares the same information on all the other variables $\boldsymbol{x}\backslash x^{(j)}$. Then the same operations will be done over those subsets, instead of the whole $oB(e)$.

### 3.2 Variable importance

Since spatial variables consist of a single value in time, but different values for different locations, and the temporal variables are different for each day, but the same for different locations, we separate the variable importance calculations into two different groups, spatial and temporal, for the two kinds of variables.
Some of the $500m \times 500m$ boundary boxes are only partially in the IJsselland region. Since some bigger cities lie on the boundary of IJsselland, those boundary boxes might be biased and therefore inaccurate. Therefore, it was decided to cut those boundary boxes out of the variable importance calculation.
To do the analysis, we constructed one random forest for the spatial variables and another one for the temporal variables. Both were constructed out of 2000 trees. The number of explanatory variables to get from the random selection is set at about 1/3 of the total number of variables, as suggested for regression in the randomForest R package [7]. This results in 7 variables for the spatial permutations and 2 temporal variables. Since we have some highly correlated variables, we decided to use a conditional permutation technique instead of the normal permutation technique from [5]. Following [8], we use the *permimp* R-package [6]. We will put plots of both the non-conditional and conditional permutation techniques in here to compare the results. We will however use the results from the conditional permutation only. The plots of the spatial and temporal permutation importance are found in Figure 3 and Figure 4, respectively. The larger the increase of certain explanatory variables in the plots, the more important they would be for the prediction. Note that the numbers on the y-axes of the plots for temporal and spatial importance are not comparable with each other, since the importance is calculated with separate random forests.




(A) Conditional permutation (B) Non-conditional permutation

FIGURE 4: *The importance (increase in prediction error) (y-axis) for temporal variables using conditional (A) and non-conditional (B) permutation techniques.*

### 3.2.1 Spatial importance

Looking at the spatial importance plots, one could see that they are very different for conditional and non-conditional permutation. This is indeed possible to happen, thinking about the high correlation between certain explanatory variables. More residents usually means a higher address density, and more residents also means more women and/or men. This results in much higher values for certain explanatory variables in the non-conditional permutations, as explained at the end of Section 3.1.2. This is the exact reason why we use the conditional permutation importance instead. If we look at the figure with the conditional permutation importance (Figure 3a), we notice that there are 3 explanatory variables that have much higher scores than all the others: freestanding houses, houses built between 1920 and 1945 and houses built before 1920. The other variables are all much closer to each other. Some explanations for this could be: i) freestanding houses in general have chimneys more often than other house types, ii) older houses with chimneys usually have their chimneys built of bricks, which have a higher fire risk, instead of the more recent stainless steel. This would be probable reasons for the high importance of those variables. As we had hoped before, the most important spatial variables in the IJsselland region are very similar to those from the Twente research [8]. The only big difference is that buildings built before 1920 seem to matter a lot more in IJsselland than in Twente. When looking at the data of the two regions, we see that IJsselland also has more houses built before 1920 that are still being used. It is possible that this could have influence on the result, since the Twente region does not have as many houses from this period.

### 3.2.2 Temporal importance

If we look at the plot of the non-conditional importance, Figure 4b, we see that temperature and windchill have very high importance scores compared to sunshine and wind speed. When we look at the conditional permutation importance, we see that the importance of windchill is by far the most dominant. We see that the score for temperature is a lot lower than for windchill, which is probably because windchill depends directly on the temperature. Temperature is, however, still quite high compared to the other two variables. It could be that temperature would not be needed for the future prediction models and windchill alone would be more than enough, but temperature will still be considered as a potential variable for the model. We will, in Section 4.4, do likelihood ratio tests whether the influence of temperature is needed for fitting the separate models. Temperature and

windchill having a high influence on the amount of fires could be explained by the fact that a low temperature or a low windchill both make people feel cold, which could be a reason for people to use their chimneys. Obviously, the more a chimney is used, the higher the probability will be for a chimney fire. Comparing this to the most important variables in the Twente research, we see that we have temperature as second most important variable, instead of wind speed. Next to that, the importance values for both of them are way smaller compared to windchill in the IJsselland region than they are for Twente. The best possible explanation for this would be that wind chill already contains information on both temperature and wind speed, since it is a combination of the two variables. This means that, with windchill being on top, the influence of both of them would already be lower. With certain permutation in the random forests, it is possible that some of that importance is "given" to the windchill instead.

# 4 Poisson point process model

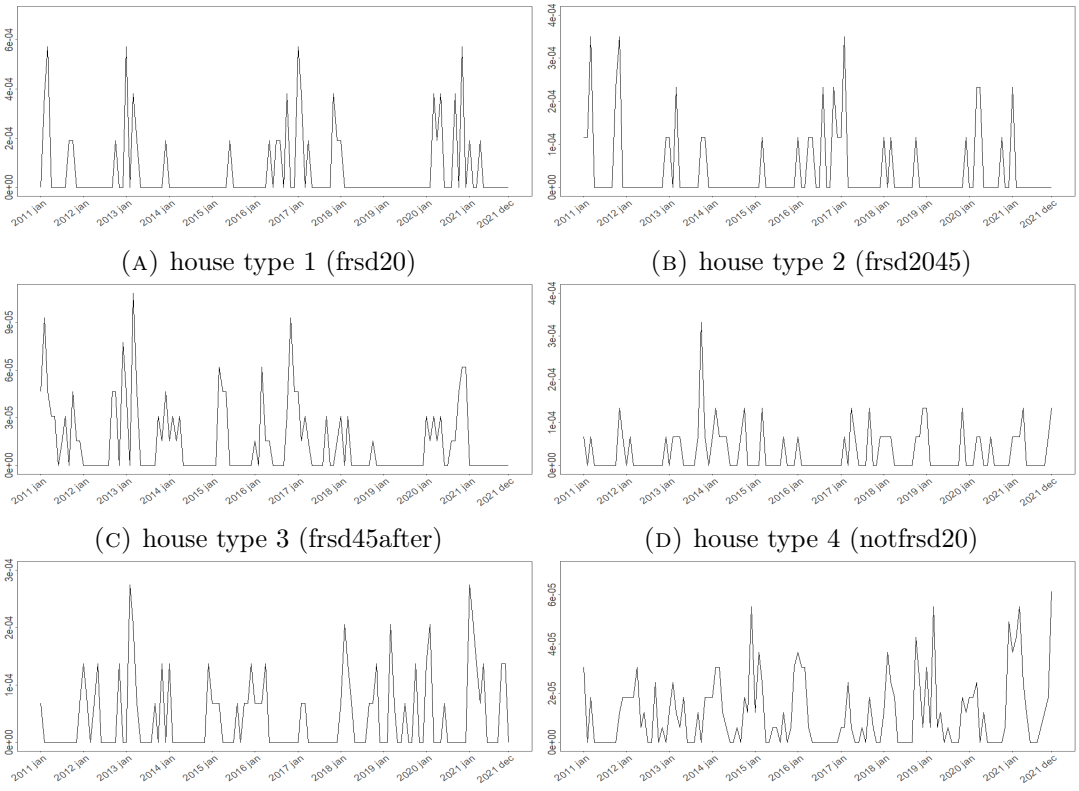
## 4.1 Model motivation

To build a statistical model, we have to look at the influence of the different most important variables on the amount of chimney fires that occur. For this, similar to what happened in [8], we divide all the houses into six different groups:

- Freestanding houses built before 1920 (frsd20)
- Freestanding houses built between 1920 and 1945 (frsd2045)
- Freestanding houses built after 1945 (frsd45after)
- Non-freestanding houses built before 1920 (notfrsd20)
- Non-freestanding houses built between 1920 and 1945 (notfrsd2045)
- Non-freestanding houses built after 1945 (notfrsd45after)

Those six categories contain all the possible houses, so that we do not miss any for the prediction. The names of the variables are put in brackets behind their explanation above.

In Figure 5, we see the monthly intensities for the six different house types (note that the y-axis scales are not the same). We see that for all the specific house types, the chimney fire occurrences are periodic, with the winter months having higher intensities than the summer months. The intensities itself are however very different: freestanding houses have higher fire intensities compared to non-freestanding houses and houses built before 1920 or between 1920 and 1945 have higher intensities compared to houses from 1945 or later as well. The intensities are not perfectly periodic either. Months have different intensities in different years. Therefore, we need to take windchill and temperature into account.
In Figure 6, we see the number of chimney fires in $500m \times 500m$ boxes plotted against the number of houses in that box. The blue line is the locally estimated scatterplot smoothed (loess) curve, with the gray parts around it the 95% envelopes. We notice that for all the specific types, the amount of chimney fires is approximately linear to the number of houses. This does however only hold when the number of houses is not too large. For house type 3, the freestanding houses built after 1945, it stays linear (Figure 6c). For house type 1, 2 and 6, the freestanding houses built before 1920 (Figure 6a), freestanding houses built between 1920 and 1945 (Figure 6b), and not freestanding houses built after 1945 (Figure 6f),



(A) house type 1 (frsd20)

(B) house type 2 (frsd2045)

(C) house type 3 (frsd45after)

(D) house type 4 (notfrsd20)

(E) house type 5 (notfrsd2045)

(F) house type 6 (notfrsd45after)

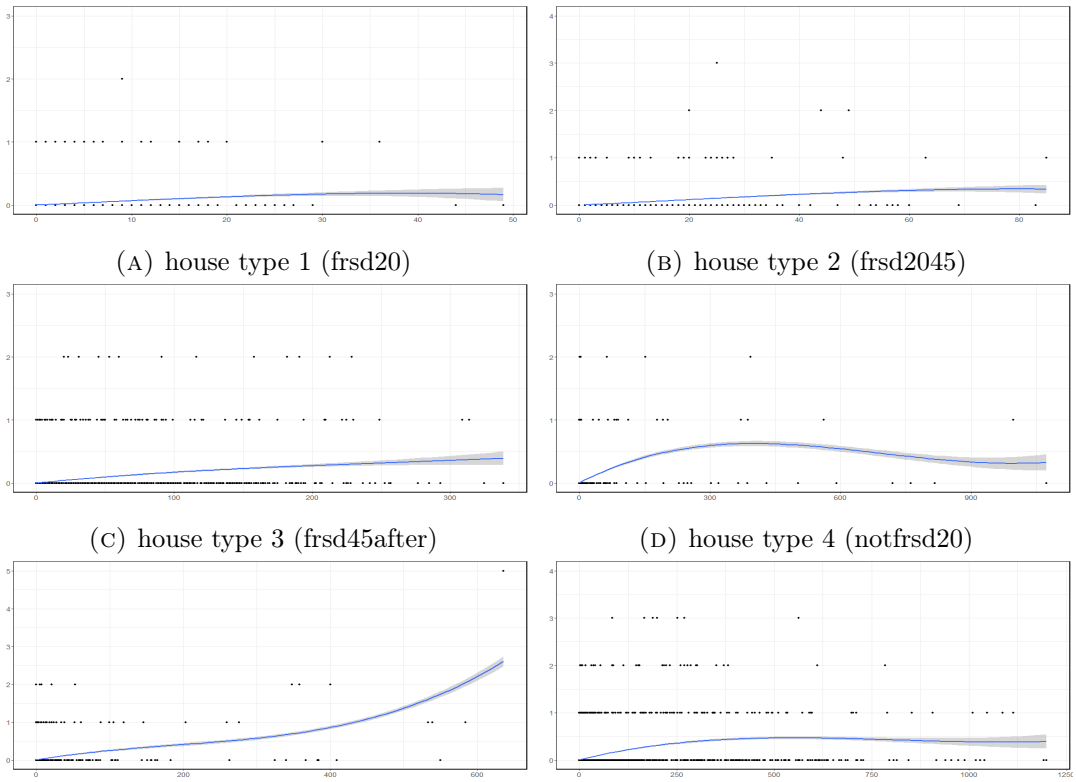
FIGURE 5: *Monthly intensities of chimney fires per house (y-axis) for the six different house types. For the explanation of the variable names, see the list at the start of Section 4.1. Unit is day*$^{-1}$*. Note that the scales of the y-axes are not the same for every figure.*

.

respectively, we see that it goes to a saturated value after a certain number of houses. For the other two house types, the not freestanding houses built before 1920 (Figure 6d), and not freestanding houses built between 1920 and 1945 (Figure 6e), it goes to weird values, by either decreasing or increasing quite a lot, respectively. It is very likely that there is not enough data for the higher values for the number of houses, since, as the plot shows, there are many data points at the lower values, whereas the higher quantities of houses do not have many data points. This is a very probable cause of the weird lines at the higher number of houses for those two house types.

## 4.2 Model structure

For the prediction, we will construct a point process model. We have found explanatory variables for the areal unit model (with the $500m \times 500m$ boxes), but we assume that the explanatory variables for the point process model are the same as for the areal unit model. As we noticed in Section 4.1, the different house types catch chimney fires at a type-dependent rate, which changes over time. This means our prediction model will be a spatial-temporal Poisson point process, with an intensity function of the form $\Lambda(u,t) = \sum_k \lambda_k(u,t)$, where $\Lambda(u,t)$ is for the total fire risk at location $u$ and time $t$, and $\lambda_k(u,t)$ is the fire risk of house type $k$ at location $u$ and time $t$. From the conclusion of Section 4.1,



(A) house type 1 (frsd20)

(B) house type 2 (frsd2045)

(C) house type 3 (frsd45after)

(D) house type 4 (notfrsd20)

(E) house type 5 (notfrsd2045)

(F) house type 6 (notfrsd45after)

FIGURE 6: *Number of chimney fires (y-axis) in boxes with a certain number of houses (x-axis) for the six different house types. The blue curves are the estimated trends using the locally estimated scatterplot smoothing method (loess), including 95% envelopes to show the confidence of the estimates. For the explanation of the variable names, see the list at the start of Section 4.1*

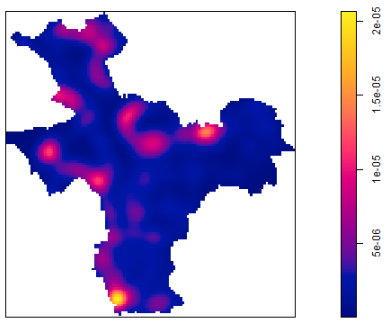
.

where we found out that the amount of fires is linear with the density, as long as the density is not too large, we can define $\lambda_k(u,t) = h_k(u)\vartheta_k(t)$, with $h_k(u)$ the house density of type $k$ at location $u$ and $\vartheta_k(t)$ the risk for a chimney fire for house type $k$ at time $t$. The density maps of the 6 different house types ($h_k(u)$) can be found in Figure 7. They are derived by Gaussian kernel smoothing on the corresponding data of the house locations with a standard deviation of 1500 metres. This is bigger than it is in the Twente research [8], which had a standard deviation of 1000 metres. The reasoning behind this is that the cities in IJsselland are even further away from each other, and mostly consist of very small cities and towns. Since this means there are less data points for the houses, more smoothing is required. Since we have that both seasonal information and temperature ($V_{\tau,2}(t)$) and windchill ($V_{\tau,3}(t)$) influence the chimney fire rate, we will use the following intensity function $\vartheta_k(t)$:

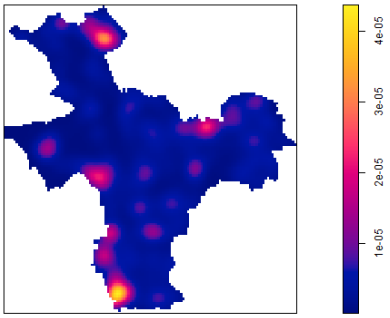
$$\vartheta_k(t) = \exp(Harmonic(t, o_{k,1}) + Polynom(V_{\tau,2}(t), o_{k,2}) + Polynom(V_{\tau,3}(t), o_{k,3}) + Polynom(V_{\tau,2}(t)V_{\tau,3}(t), o_{k,4}). \quad (4)$$

The harmonic part (order $o_{k,1}$) models the influence of seasonal information, and the polynomials with order $o_{k,2}, o_{k,3}$ and $o_{k,4}$ model the influence of temperature, windchill, and their interaction ($V_{\tau,2}(t)V_{\tau,3}(t)$) respectively. The exponent is used to make sure we

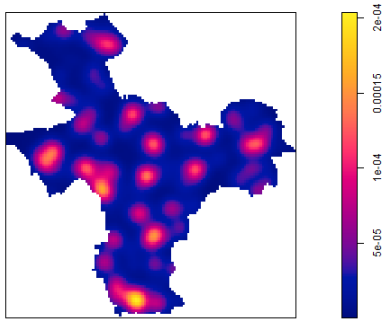
never have a negative intensity function. For parametric representation, the functions $\vartheta_k(t)$ and $\lambda_k(u,t)$ can be described as $\vartheta_k(t,\boldsymbol{\theta}_k)$ and $\lambda_k(u,t,\boldsymbol{\theta}_k)$ respectively, with $\boldsymbol{\theta}_k$ the vector of coefficients in equation (4) for house type $k$. This model is very similar to the model in the Twente research [8], but where they used wind speed, their second most important temporal variable, we use temperature, our second most important temporal variable. This means that in the interaction terms, wind speed is obviously replaced by temperature as well.



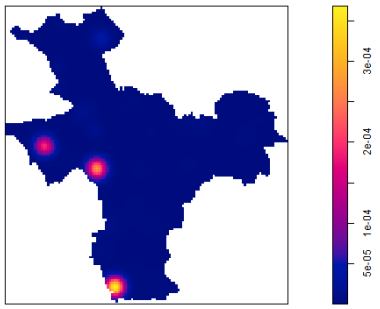

(A) house type 1 (frsd20)



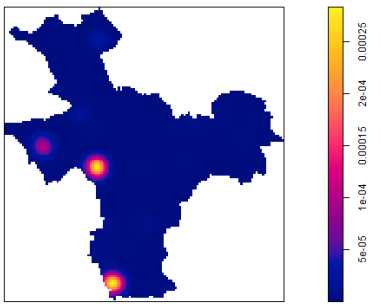

(B) house type 2 (frsd2045)



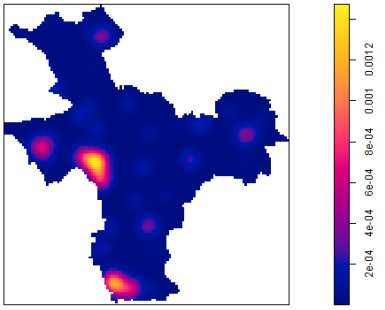

(C) house type 3 (frsd45after)




(D) house type 4 (notfrsd20)




(E) house type 5 (notfrsd2045)




(F) house type 6 (notfrsd45after)

FIGURE 7: *Density maps of the six different house types using Gaussian kernel smoothing. Unit is metre$^{-2}$, the scales of the intensity bars are different for each house type. For the explanation of the variable names, see the list at the start of Section 4.1*

.

### 4.3 Model fitting methods

For us to be able to fit the prediction model, we are going find the different parameters with their values. To do this, we will apply maximum likelihood estimation for the log-likelihood function of our point process:

$$\sum_{k=1}^{6}\left(\sum_{x_k}\log\lambda_k(x_k,\boldsymbol{\theta}_k)-\int_{IJsselland}\int_{2011-2021}\lambda_k(u,t,\boldsymbol{\theta}_k)dudt\right) \tag{5}$$

We will approximate the equation using logistic regression estimation [3], such that we can adjust the distribution of quadrature (dummy) points for the approximation in such a way that we have more points at important areas to reduce the estimation variance.
The logistic regression estimation is based on the Campbell-Mecke theorem [9]. For a point process $X$ on a bounded space-time domain $W\times T\subseteq\Re^2\times\Re$ with intensity function $\lambda(u,t)$ with $(u,t)\in W\times T$, for any (vector of) real functions $\boldsymbol{f}(u,t)$ defined on $W\times T$ with $\boldsymbol{f}(u,t)\lambda(u,t)$ absolutely integrable, the theorem reads

$$E\left(\sum_{x\in X}\boldsymbol{f}(x)\right)=\int_W\int_T\boldsymbol{f}(u,t)\lambda(u,t)dudt \tag{6}$$

where $x$ runs through the points of $X$. For our case, $\lambda$ depends on a parameter vector $\boldsymbol{\theta}$ as well, so we write $\lambda(u,t,\boldsymbol{\theta})$. Both sides of the Campbell-Mecke theorem can be estimate to obtain estimating equations for $\boldsymbol{\theta}$.
To find the estimating equations, we use the following vector function

$$\boldsymbol{f}(u,t)=\frac{\partial}{\partial\boldsymbol{\theta}}log\left(\frac{\lambda(u,t,\boldsymbol{\theta})}{\lambda(u,t,\boldsymbol{\theta})+\rho(u,t)}\right)=\left(\frac{\rho(u,t)/\lambda(u,t,\boldsymbol{\theta})}{\lambda(u,t,\boldsymbol{\theta})+\rho(u,t)}\nabla\lambda(u,t,\boldsymbol{\theta})\right) \tag{7}$$

with $\nabla\lambda(u,t,\boldsymbol{\theta})$ the gradient with respect to $\boldsymbol{\theta}$ and $\rho(u,t)$ is a positive-valued function defined on $W\times T$. To estimate the integral in (6), we will use a dummy process $D$ on $W\times T$ with intensity function $\rho(u,t)$, independent from $X$. Using $\boldsymbol{f}(u,t)$ from (7) for (6), we approximate the integral by applying the Campbell-Mecke theorem to $D$ as well:

$$E\left(\sum_{x\in D}\frac{\nabla\lambda(x,\boldsymbol{\theta})}{\lambda(x,\boldsymbol{\theta})+\rho(x)}\right)=\int_W\int_T\frac{\rho(x)}{\lambda(x,\boldsymbol{\theta})+\rho(x)}\nabla\lambda(x,\boldsymbol{\theta})dudt. \tag{8}$$

Now we have unbiased estimators for both sides of the Campbell-Mecke equation (6). For a more specific explanation of the theory and proofs, one is encouraged to take a look at [8]. Since this is not the point of this research, we will continue as if this has been made clear.
For all six house types, we will use logistic regression estimation to find the intensity functions. For each house type $k$, we have $\boldsymbol{\theta}_k$, as defined in Section 4.2, as the vector with the coefficients of the harmonic and polynomial functions. The occurring chimney fires, as stated in that same section, will be considered a point process $X_k$ with intensity function $\lambda_k(u,t,\boldsymbol{\theta}_k)$. We also have the dummy point process $D_k$ with intensity function $\rho_k(u,t)$. With $n$ being the amount of parameters in $\boldsymbol{\theta}_k$, the estimated equation that will be solved is the following:

$$s_i(X_k,D_k,\boldsymbol{\theta}_k)=\sum_{x\in X_k}\frac{\lambda_k(x,\boldsymbol{\theta}_k)C_i(x)}{\lambda_k(x,\boldsymbol{\theta}_k)+\rho_k(x)}-\sum_{x\in D_k}\frac{\rho_k(x)C_i(x)}{\lambda_k(x,\boldsymbol{\theta}_k)+\rho_k(x)}=0 \tag{9}$$

for $i=1,...,n$, and with $C_p(x)$ being the $p$-th temporal covariate (the harmonic components and the windchill and temperature terms with the different order in (4) at point $x$.

### 4.4 Parameter estimation

To estimate all the parameters $\boldsymbol{\theta}_k$, we first have to specify the intensity function $\rho_k(u,t)$ for the dummy point process $D_k$. The point of our dummy point process is to approximate the integral in (6) by (8). For a good approximation, we need the dummy point process to have more points in regions that have a higher chimney fire occurrence and risk than in the other regions, such that the approximation in those areas is more accurate. Looking at Figure 6, we see that areas with more houses of a given type, experience more chimney fires in those type of houses. Therefore, we need more dummy points in the areas with a higher house density of the specific type, which we will do using the density $h_k(u)$ of house type $k$ at location $u$ from Figure 7. If we look at Figure 5, we see that there are more chimney fires in winter than in summer, so we want to assign higher values to $\rho_k(u,t)$ in the colder seasons as well. Using this, we get the following intensity function $\rho_k(u,t)$ for $D_k$:

$$\rho_k(u,t) = r_k h_k(u) \left(0.75 + 0.25 \left(\sin\left(\frac{2\pi}{365}t + \frac{\pi}{2}\right)\right)\right), \tag{10}$$

with $r_k$ being a multiplication factor that is used to ensure that $\rho_k(u,t)$ is at least four times the size of $\lambda_k(u,t,\boldsymbol{\theta}_k)$ for every $(u,t) \in W \times T$, as suggested in [3][6]. After this, we need to generate dummy points for $D_k$. This can be done using the R-package *spatstat* [4]. Using the generated dummy points and the observations of $X_k$, we can then use the R-package *stats* [10] to estimate the model parameters $\boldsymbol{\theta}_k$ using the function for fitting generalized linear models with the *logit* link function as explained in Section 4.3. To find the optimal function orders $o_{k,1}, o_{k,2}, o_{k,3}, o_{k,4}$ in (4) simultaneously, we look for the combination that gives the lowest Akaike information criterion (AIC, see [2]) for a certain set of ranges[7]. Furthermore, since windchill already contains some information on temperature, we also do likelihood ratio tests over the best models with and without the temperature and interaction terms.

We separate the data in two sets: the first set contains all the data from 2011 till 2020, and the second set contains the 2021 data. With this, we can use the 2021 data for testing our predictions, and the other data to actually fit the models of each house type. With the likelihood ratio tests, we find out that three out of the six house types will indeed suffice using only the windchill variables. For house type 2, freestanding between 1920 and 1945, and house type 3, freestanding after 1945, temperature was found to be needed as well. For house type 6, not freestanding after 1945, both interaction and temperature are needed. If we compare this with the Twente model, we see that the IJsselland model functions have fewer harmonic terms, but more polynomial terms. Furthermore, looking at the parameter estimates for the Twente research [8], we see that, in general, the estimated parameter values for the harmonic functions for the Twente model are bigger than for the IJsselland model, especially for the lower order cosines, which give the higher intensities for winter. Those two difference could be explained, since, in general, temperature and windchill contain some harmonic characteristics as well, where they are usually higher in summer months and lower in winter months. This means that both models contain the influence of the seasons, but in different ways. The estimated temporal functions from (4)

[6] The $r_k$'s are given by 40, 40, 20, 20, 16 and 8 respectively.

[7] $o_{k,1} : 1-4, o_{k,2} : 1-5, o_{k,3} : 1-5, o_{k,4} : 1-5$.

for the six house types are given by:

$$\begin{aligned}\vartheta_1(t) =&\exp(\theta_{1,1} + \theta_{1,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{1,3}\sin\left(\frac{2\pi}{365}t\right) + \theta_{1,4}\cos\left(\frac{4\pi}{365}t\right) + \theta_{1,5}\sin\left(\frac{4\pi}{365}t\right)\\ &+ \theta_{1,6}V_{\tau,3}(t))\end{aligned} \tag{11}$$

$$\begin{aligned}\vartheta_2(t) =&\exp(\theta_{2,1} + \theta_{2,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{2,3}\sin\left(\frac{2\pi}{365}t\right) + \theta_{2,4}\cos\left(\frac{4\pi}{365}t\right) + \theta_{2,5}\sin\left(\frac{4\pi}{365}t\right)\\ &+ \theta_{2,6}V_{\tau,2}(t) + \theta_{2,7}V_{\tau,3}(t) + \theta_{2,8}V_{\tau,3}^2(t) + \theta_{2,9}V_{\tau,3}^3(t))\end{aligned} \tag{12}$$

$$\begin{aligned}\vartheta_3(t) =&\exp(\theta_{3,1} + \theta_{3,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{3,3}\sin\left(\frac{2\pi}{365}t\right) + \theta_{3,4}\cos\left(\frac{4\pi}{365}t\right) + \theta_{3,5}\sin\left(\frac{4\pi}{365}t\right)\\ &+ \theta_{3,6}\cos\left(\frac{6\pi}{365}t\right) + \theta_{3,7}\sin\left(\frac{6\pi}{365}t\right) + \theta_{3,8}V_{\tau,2}(t) + \theta_{3,9}V_{\tau,2}^2(t) + \theta_{3,10}V_{\tau,2}^3(t)\\ &+ \theta_{3,11}V_{\tau,3}(t) + \theta_{3,12}V_{\tau,3}^2(t) + \theta_{3,13}V_{\tau,3}^3(t) + \theta_{3,14}V_{\tau,3}^4(t) + \theta_{3,15}V_{\tau,3}^5(t))\end{aligned} \tag{13}$$

$$\begin{aligned}\vartheta_4(t) =&\exp(\theta_{4,1} + \theta_{4,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{4,3}\sin\left(\frac{2\pi}{365}t\right) + \theta_{4,4}\cos\left(\frac{4\pi}{365}t\right) + \theta_{4,5}\sin\left(\frac{4\pi}{365}t\right)\\ &+ \theta_{4,6}V_{\tau,3}(t) + \theta_{4,7}V_{\tau,3}^2(t) + \theta_{4,8}V_{\tau,3}^3(t))\end{aligned} \tag{14}$$

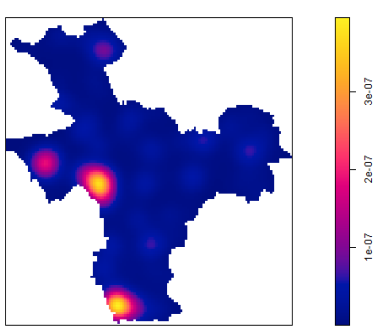
$$\begin{aligned}\vartheta_5(t) =&\exp(\theta_{5,1} + \theta_{5,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{5,3}\sin\left(\frac{2\pi}{365}t\right) + \theta_{5,4}\cos\left(\frac{4\pi}{365}t\right) + \theta_{5,5}\sin\left(\frac{4\pi}{365}t\right)\\ &+ \theta_{5,6}V_{\tau,3}(t) + \theta_{5,7}V_{\tau,3}^2(t))\end{aligned} \tag{15}$$

$$\begin{aligned}\vartheta_6(t) =&\exp(\theta_{6,1} + \theta_{6,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{6,3}\sin\left(\frac{2\pi}{365}t\right) + \theta_{6,4}V_{\tau,2}(t)\\ &+ \theta_{6,5}V_{\tau,3}(t) + \theta_{6,6}V_{\tau,3}^2(t) + \theta_{6,7}V_{\tau,3}^3(t) + \theta_{6,8}V_{\tau,3}(t) * V_{\tau,2}(t))\end{aligned} \tag{16}$$

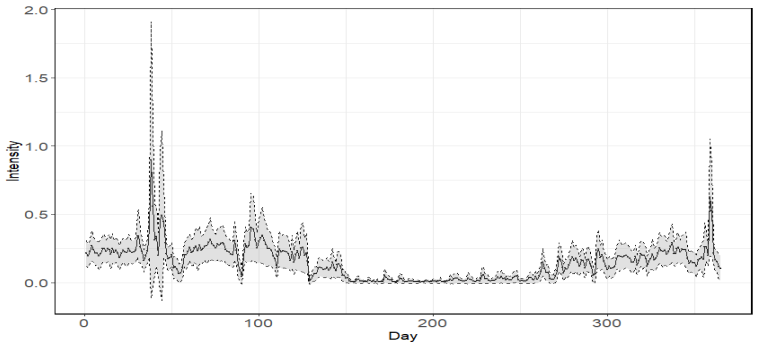
The estimates of all the model parameters can be found in Table 2. For 2021, we show the predictions for both space and time domain, as well as the actual data for that year, in Figure 8.

TABLE 2: *Parameter estimates for the Poisson point process models with intensity function as in (11)-(16) with their 95% confidence interval (CI). The 'e' stands for base 10, so e1 is $10^1$.*

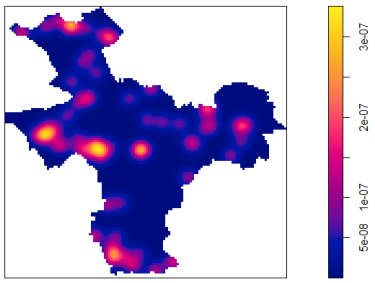
| Parameter | Estimate (CI) | Parameter | Estimate (CI) | Parameter | Estimate (CI) |
|---|---|---|---|---|---|
| $\theta_{1,1}$ | -1.24e1(±0.77e0) | $\theta_{1,2}$ | 3.43e-2(±8.17e-1) | $\theta_{1,3}$ | -7.56e-2(±4.96e-1) |
| $\theta_{1,4}$ | -3.15e-1(±4.42e-1) | $\theta_{1,5}$ | -1.59e-1(±4.34e-1) | $\theta_{1,6}$ | -9.12e-2(±7.73e-2) |
| $\theta_{2,1}$ | -1.43e1(±1.60e0) | $\theta_{2,2}$ | 9.77e-1(±1.32e0) | $\theta_{2,3}$ | 8.07e-2(±6.53e-1) |
| $\theta_{2,4}$ | -8.54e-1(±6.66e-1) | $\theta_{2,5}$ | -2.41e-1(±5.95e-1) | $\theta_{2,6}$ | 2.88e-1(±4.05e-1) |
| $\theta_{2,7}$ | -2.75e-1(±3.41e-1) | $\theta_{2,8}$ | -1.02e-2(±1.29e-2) | $\theta_{2,9}$ | 4.91e-4(±4.35e-4) |
| $\theta_{3,1}$ | -1.39e1(±9.56e-1) | $\theta_{3,2}$ | 3.24e-2(±7.21e-1) | $\theta_{3,3}$ | -2.73e-1(±3.62e-1) |
| $\theta_{3,4}$ | -5.09e-1(±4.54e-1) | $\theta_{3,5}$ | 2.80e-2(±3.86e-1) | $\theta_{3,6}$ | 3.36e-1(±3.29e-1) |
| $\theta_{3,7}$ | -1.61e-1(±3.21e-1) | $\theta_{3,8}$ | 5.40e-2(±2.75e-1) | $\theta_{3,9}$ | -4.44e-2(±4.29e-2) |
| $\theta_{3,10}$ | 3.12e-3(±2.96e-3) | $\theta_{3,11}$ | -1.46e-2(±2.42e-1) | $\theta_{3,12}$ | 2.85e-2(±1.59e-2) |
| $\theta_{3,13}$ | -1.81e-3(±1.45e-3) | $\theta_{3,14}$ | -1.40e-4(±8.27e-5) | $\theta_{3,15}$ | 4.13e-6(±2.53e-6) |
| $\theta_{4,1}$ | -1.34e1(±8.32e-1) | $\theta_{4,2}$ | -3.54e-2(±9.43e-1) | $\theta_{4,3}$ | -3.69e-1(±4.92e-1) |
| $\theta_{4,4}$ | -3.66e-1(±5.22e-1) | $\theta_{4,5}$ | 1.72e-1(±4.85e-1) | $\theta_{4,6}$ | 1.87e-3(±1.01e-1) |
| $\theta_{4,7}$ | 2.00e-4(±8.74e-3) | $\theta_{4,8}$ | -5.78e-4(±6.83e-4) | $\theta_{5,1}$ | -1.28e1(±8.20e-1) |
| $\theta_{5,2}$ | -1.14e-1(±9.92e-1) | $\theta_{5,3}$ | 4.64e-2(±5.63e-1) | $\theta_{5,4}$ | -1.64e-1(±4.73e-1) |
| $\theta_{5,5}$ | 2.65e-1(±4.80e-1) | $\theta_{5,6}$ | -7.96e-2(±6.60e-2) | $\theta_{5,7}$ | -5.77e-3(±6.77e-3) |
| $\theta_{6,1}$ | -1.48e1(±5.97e-1) | $\theta_{6,2}$ | 5.01e-4(±4.03e-1) | $\theta_{6,3}$ | 9.31e-2(±2.53e-1) |
| $\theta_{6,4}$ | 2.00e-1(±1.61e-1) | $\theta_{6,5}$ | -1.97e-1(±1.49e-1) | $\theta_{6,6}$ | -2.10e-3(±4.01e-3) |
| $\theta_{6,7}$ | -3.96e-4(±4.70e-4) | $\theta_{6,8}$ | 3, 35e-4(±8.53e-8) | | |



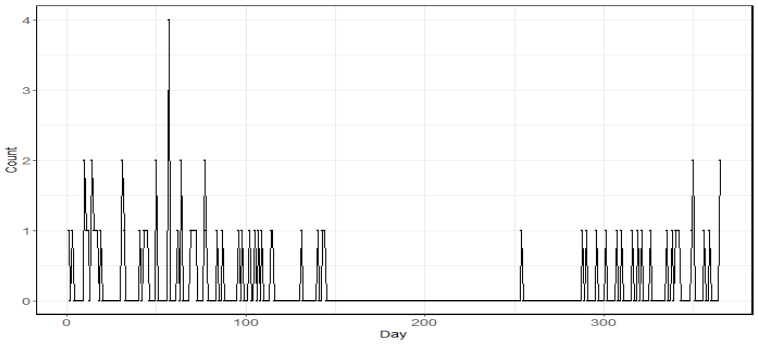

(A) Spatial prediction




(B) Temporal prediction




(C) Smoothed spatial pattern of chimney fires




(D) Daily counts of chimney fires

FIGURE 8: *The prediction and the actual occurrences of chimney fires for 2021. The temporal prediction includes the 95% CI (the dashed grey lines). The unit in the spatial plot is metre$^{-2}$, the unit of the temporal plot is day$^{-1}$. Note that the intensity bars and y-axes of the spatial and temporal plots, respectively, have different scales.*

.

Looking at the figures, we see that both the spatial and temporal predictions follow at least partially the same patterns. For the spatial predictions, we see that we overestimate the bigger cities, Deventer and Zwolle, a bit, whereas we underestimate the smaller cities and towns in IJsselland. At some places with high intensity values for the actual chimney fires, we have smaller intensity values. There are a few places in which the intensity in the prediction is close to zero, even though this is not the case with the smoothed actual pattern. This could happen because the probability of a chimney fire in a region with only few houses is small, but obviously not zero. It is also possible that the assumption for the Poisson point process is not sufficient. Spatial correlation may exist, which would requires higher order analysis.
The year 2021 had very different chimney fire locations compared to the other average over the other years as well, as can be seen in Figure 9. Since there are about 70 chimney fires in 2021, when less densely populated areas have 2 or 3 chimney fires happening, which is possible, the intensity for that area skyrockets already. This could explain the larger amount of higher intensity dots in Figure 8c as well. It makes sense that city centres are overestimated a bit as well, since, looking at Figure 6, we get saturated values after a certain house density. Since this is not implemented in the prediction model, we overestimate a bit for the places with very high house densities. For the temporal prediction, we see that the patterns look very similar; the intensity is much lower in the summer months and a lot higher in the winter months. We notice that the values for the intensity and the actual counts are not the same, but this is easily explained by the probability part: most days in the winter have no chimney fires, but with the days that do, the average will converge to values close to the temporal prediction. Furthermore, we also see the same huge peak in both the prediction and the actual count of chimney fires just after 50 days, which also shows that the prediction looks correct. We see that there are a lot more variations in the prediction values. The reason for this is that the prediction intensity is influenced by periodic functions as well as windchill and temperature. Windchill and temperature differ every day, which results in a lot more flunctuations in the prediction.

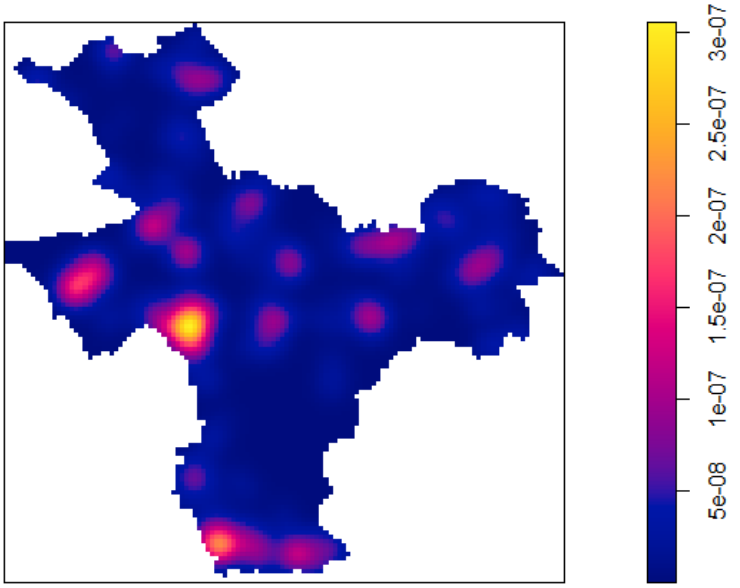



FIGURE 9: *Smoothed spatial pattern of chimney fires for 2011-2020*

### 4.5 Confidence interval

To really be able to provide any useful information with risk, it is very important to give the uncertainty of the estimate as well. Therefore, we give confidence intervals for our predictions. In Table 2, the 95% confidence intervals for the parameter estimation values are added to the estimated values. For the predictions, we use a result from [12] and [8], in which they prove that under certain conditions, the maximiser $\hat{\boldsymbol{\theta}}$ of 9 with true value $\boldsymbol{\theta}$ is approximately normally distributed with mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{G}$ as the inverse of the Godambe matrix. A plug in estimator for $G$ is

$$\hat{\boldsymbol{G}} = \left[ \int_T \int_W \frac{\lambda_k(u,t,\hat{\boldsymbol{\theta}})\rho(u,t)}{\lambda_k(u,t,\hat{\boldsymbol{\theta}}) + \rho(u,t)} C_i(u,t) C_j(u,t) du dt \right]_{i,j=1}^{n} \tag{17}$$

as in [8], for $n$ parameters in $\boldsymbol{\theta}$. Using the Delta method from [13], we can then obtain approximate 95% confidence intervals for the temporal predictions, which can be seen in Figure 8b for the year 2021.

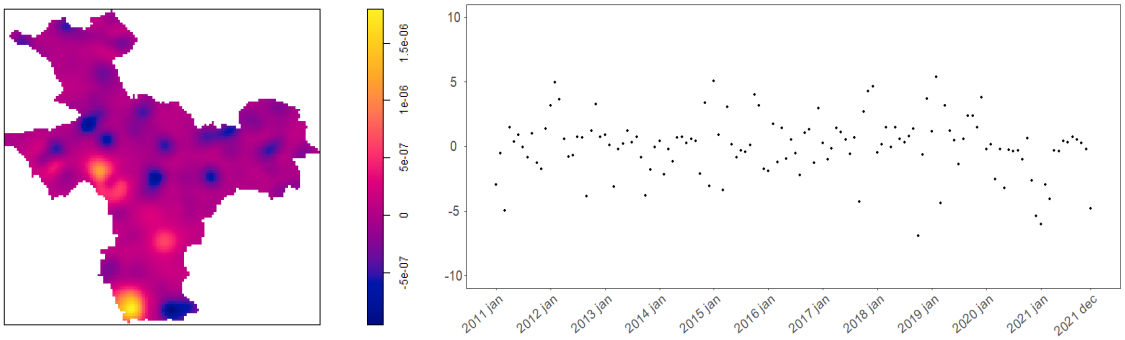
## 5 Discussion

### 5.1 Residual analysis

To validate our choice of the point process model, we perform a residual analysis on the spatial and temporal domain with the model functions defined as in (4) and (11)-(16). Just like in the Twente research, we fit the model on all data from 2011-2021, and calculate the temporal and spatial residuals. The spatial residuals can be seen in Figure 10a. We see that we still overestimate the big cities Deventer and Zwolle, and slightly underestimate the smaller areas. The reasoning for this is the same as for the prediction of 2021 in Section 4.4, where we go to saturated values for higher intensities in Figure 6, which means that the high intensity areas will be overestimated, and low intensity areas slightly underestimated. A possible solution for his could be a piece-wise function instead of the linear function between the chimney fires and the density of the six housetypes.
For the temporal residuals in Figure 10b, there is no special pattern to be seen. We see that most of the outliers are around the winter months, which makes sense, since those months also have more flunctuations. Furthermore, we find that the average residual per month is basically zero, as it obviously should be.

## 6 Conclusion

In this paper, we used the similar approach as in the Twente research in [8]. We first used random forests with conditional permutation importance techniques to find the most important variables out of the list in Table 1. Similar to the Twente region, we see that pre-war detached or semi-detached houses run a higher chimney fire risk. The difference between the important variables was that houses built before 1920 are a lot more important in the IJsselland region. This was explained by the fact that IJsselland has many more houses from this era compared to Twente. For the prevention, the fire services should therefore target their public awareness campaigns mostly to house owners of such houses. For temporal variables, we found windchill dominating, the same as in the Twente research. However, we had temperature as second instead of windchill. Since both are included in windchill, this was explained to just be random because of that, since they both get suppressed by windchill. Secondly, based on the observed relations between chimney fires




(A) spatial residuals

(B) temporal residuals (per month)

FIGURE 10: *The spatial (A) and temporal (B) residuals from the chimney fire data from 2011-2021. The unit in the spatial plot is metre$^{-2}$, the unit of the temporal plot is month$^{-1}$.*

and the variables, we defined a Poisson point process model to learn the observed spatio-temporal point patterns to predict the fire risks. We included confidence intervals for the estimation parameters. Afterwards, we did the predictions on the spatial and temporal domain, and compared those to the actual data for this. We included confidence intervals for the temporal predictions as well. Lastly, we did some model validation with residual analysis, where we look at the difference between the prediction over all years and the actual data.
As explained in Section 4.4, it is possible that our Poisson point process assumption was not correct and that there would be spatial or temporal correlation. To be certain about this, it would need second-order analysis with the pair correlation function and the K-function as in [8]. However, due to errors with the dataset, there is sadly not time left to do this part. This will therefore be left as a suggestion for future research. This also holds for trying a piece wise linear function for the house density influence on the intensity. This could be a great improvement, but would be a suggestion for future work as well.
On top of that, for future research, it could be a great idea to look into the influence on other regions, which are similar to the Twente and IJsselland regions. This could be either in the Netherlands or maybe even for other countries, where similar regions will probably exist as well. It would also be a good idea, like explained in the conclusion in [8] to do a study into data-driven variables selection when the variables are available for every point in space and time.

# References

[1] Bijna vijf schoorsteenbranden per dag in 2020. `https://www.unive.nl/actueel/bijna-vijf-schoorsteenbranden-per-dag-in-2020`.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[3] A. Baddeley, J. Coeurjolly, E. Rubak, and R. Waagepetersen. Logistic regression for spatial gibbs point processes. 101:377–392, 2014.

[4] A. Baddeley and R. Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42, 2005.

[5] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[6] D. Debeer and C. Strobl. Conditional permutation importance revisited. *BMC Bioinformatics*, 21:1–30, 2020.

[7] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2:18–22, 2002.

[8] C. Lu, M. N. M. van Lieshout, M. de Graaf, and P. Visscher. Data-driven chimney fire risk prediction using machine learning and point process tools. 2021.

[9] R. P. Møller, J .and Waagepetersen. *Statistical inference and simulation for spatial point processes*. 2004.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2020.

[11] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:1–11, 2008.

[12] M. N. M. van Lieshout and C. Lu. Infill asymptotics for logistic regression estimators for spatio-temporal point processes. 2022.

[13] J. M. Ver Hoef. Who invented the delta method? *The American Statistician*, 66:124–127, 2012.