UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

A Log Gaussian Cox process for predicting chimney fires at Fire Department Twente

Martine Leonarda School

M.Sc Thesis
August 2018

Graduation committee:
Prof. Dr. R.J. Boucherie
Dr. Ir. M. de Graaf
Prof. Dr. M.N.M. van Lieshout
E.M.A. Sanders
Prof. Dr. A.A. Stoorvogel

Stochastic Operation Research
Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Abstract

The Twente fire department is developing an interest in the use of Business Intelligence for their operations with the data they have available from all emergency calls made in the past twelve years (2004 until 2016). Last year, an applied mathematics student started a collaboration with the fire department by modelling fire-related emergency calls in the region of Twente [21]. He investigated whether an inhomogeneous Poisson process could describe these emergency calls. The answer was unfortunately not satisfying, because too many incidents of different types were considered and the relatively simple inhomogeneous Poisson process did not cover the data well. In this research we focus on one of the largest types of fires, chimney fires, and expand the model to also encounter spatially dependent noise. The inhomogeneous Poisson process is therefore extended with a random field which results in a Log Gaussian Cox process. The research includes finding the spatial and temporal influence covariates of chimney fires and modelling the Log Gaussian Cox process in two steps, first modelling the inhomogeneous Poisson process and then adding the random field corresponding to the spatially dependent noise. The number of residents in an area and the mean daily temperature have the highest influence on the occurrence of chimney fires, with an extension for the month October where people start using their chimneys. The resulting Log Gaussian Cox process is dependent on the above three variables based on residents, temperature and the month October and together with the spatially dependent noise it delivers satisfying results for predicting chimney fires in Twente. Finally, a dashboard is constructed to put the prediction into practice and to make Business Intelligence visible in the organisation.

Keywords Point processes, inhomogeneous Poisson process, chimney fires, Log Gaussian Cox process, spatio-temporal, distance analysis, correlation analysis, minimum contrast method.

Preface

This thesis is the result of my Applied Mathematics graduation assignment that I have been working on for the past seven months. The work was carried out at the Fire Department Twente and the University of Twente which made it a combination of two different worlds: One very practical and not used to a lot of theoretical talk and one focussing on the mathematical analysis of problems. To have both of these aspects in my graduation taught me more mathematical background but also I developed the skill to clarify the process in an unfamiliar way, to make it understandable for everyone. Working at the fire department strengthened my motivation because this organisation is full of heroes who take risk for every one in this society.

There are a few people I want to thank in particular for their help and extensive support during my work. At the University of Twente two supervisors made it possible for me to do this thesis and to expand my mathematical expertise in the field of point processes, Dr. Maurits de Graaf and Prof. Dr. Marie-Colette van Lieshout. Both of them were closely involved in the research and gave me perfect guidance during the past months. My third supervisor, Emiel Sanders, was involved from the Fire Department Twente and during the days I spent there he gave me a lot of opportunities to broaden my mathematical perspective so that the results would be pertinent from a business point of view. I also want to thank him a lot to give me the chance to see also the practical side of the fire department and be fully involved in the team and the organisation.

Beside that, I want to thank the people of the fire department, who showed me their part of the job and explained their tasks and ways to me in the field of fire handling. To join the fire men made the research more concrete for me and strongly strengthened my motivation to work hard. The moments we ran to the fire engine were amazing moments for which I am very thankful. Also the opportunity of writing an article is a wonderful chance which I am grateful for. Thanks for your trust.

Finally, I want to thank my friends and family for their support and belief in me. My boyfriend, my parents and sisters, my sorority, student association and my mathematician friends deserve extra attention because of their support during my whole study. I hope you enjoy reading my thesis.

Tineke School,
August 2018, Enschede

Contents

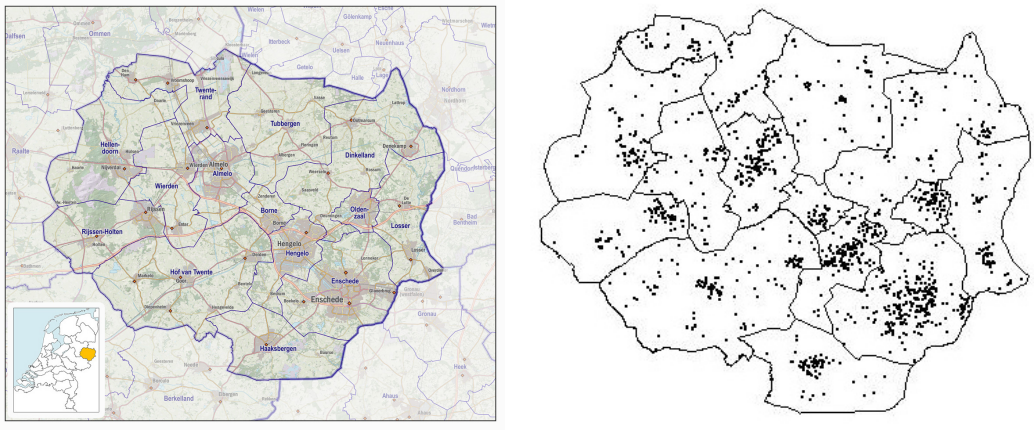
1 Introduction

1.1 Situation

Fire departments all over the world are every minute of every day prepared to provide help to citizens in the neighbourhood. Their tasks can be described as fire suppression and prevention, rescue, basic first aid, and investigations. In Twente, a region in the east of the Netherlands as displayed in Figure 1a, the fire department consists of 29 fire houses and together they handle almost 5000 incidents a year of which around 1400 incidents are actual fires. Other incidents are for example accidents or finding a gas leak. Some days an endless number of incidents arise and the fire fighters are having a full plate, but no incident happening the whole day is a scenario which is also not uncommon. Imagine that we can predict the number of incidents in the upcoming week and also know where these incidents will happen. With this information supplies and cars can be relocated and fire fighters with specific skills can be moved to areas where we expect a specific incident to happen.

The above prediction dream is an element of the new way the fire department wants to enter. In 2010 the fire department constructed a new policy with three key points: optimise their own organisation, share knowledge with partners and to counsel inhabitants 'right on time'. The fire department owns a lot of data and is eager to utilize this data to put their policy into practise. The thirst of using Business Intelligence already resulted in some analysis, for example the statistical analysis done in [12], but these researches do not go into debt yet. Therefore the fire department and the mathematical department from Twente joined forces to let Business Intelligence be involved in the field of fire incidents and to learn about the possibilities in their organisation.

Last year a bachelorthesis was already performed to make a small step to reach this goal [21]. The purpose of that research was to investigate the possibilities of predicting incidents in general based on the incident data the fire department has available, but modelling the types together gave unsatisfying results. In this report, we want to go further by using a focus on chimney fires. In this way we reduce the amount of data and are more likely to find a fitting model. This specific type of fire is chosen because chimney fires are strongly season dependent and it is the most common type of fire in Twente with around 1800 fires over the years 2004 until 2015 (11%). This sounds like a small part, but there are more than 200 types of fires and within these specifics, chimney fires happen the most. The chimney fire incidents in our data are visualized in Figure 1b.



(a) The map of Twente.

(b) All chimney fire incidents.

Figure 1: The left image shows the real map of Twente with all towns and cities in the region and the right image displays all incidents during the years 2004 until 2016 in Twente where a chimney was involved in the cause of the fire.

Because of the strong relationship with the seasons and thus also weather, we expect to create a model with a small amount of variables. When a satisfying prediction model is found, the procedure can be repeated for other types of fires which may end up with a fire prognosis for the upcoming week for all types. Therefore the central question in this work is: Can we find a mathematical model and a procedure to fit this model which can capture the properties of chimney fires? With such a model chimney fires can be predicted and the procedure can be repeated for other types of fires.

Beside focussing on a specific type of fire, the research of [21] also suggested to extend the model they used: the inhomogeneous Poisson process. In this thesis, spatial and temporal covariates (weather conditions, specifics of Twente) are chosen which are expected to have a high impact on the number of incidents happening. For all these covariates, a correlation coefficient was calculated and the six indicators with the highest correlation were used to fit an intensity function for the inhomogeneous Poisson process. With this intensity function, the model can be fully described and predictions can be made. Wendels has proposed such a model for the five general types of incidents: Fire, Service, Accident, Alert and Environmental. All of these models included a different combination of six covariates and therefore also another resulting intensity function. Through analysis of the data, it seemed better to add some spatially dependent noise but this cannot easily be contained in an inhomogeneous Poisson process. An inhomogeneous Poisson process can namely cover influence of for example weather or the number of specific buildings in an area or other variables we have data on, but dependence between covariates and incidents are excluded. We may miss there some variables which have a reasonable impact when other covariates are present or variables that do not follow a pattern, for example human behaviour. These unknown covariates and the human influence can be seen as random noise. To improve the results, another suggestion was therefore to add spatially dependent noise through a random field.

To combine the theoretical influences we know and this random noise, a Log Gaussian Cox process is introduced. Cox processes are often used in similar problems, see [8] and [10], but it hasn't been used yet in the occurrences of chimney fires. The Log Gaussian Cox process uses the inhomogeneous intensity function of a Poisson process together with a random field which can add this random noise based on spatial and temporal characteristics. The random field makes it possible to include for example the possibility that people seem to be more likely to cause a chimney fire when they live in an area with more chimneys, because they come more in contact with chimneys in general. Risky areas and also risky time periods can be extracted from data and be described by the random field. The goal of this work is find a procedure to model the Log Gaussian Cox process and check if the model gives a better description of the behaviour of chimney fires.

1.2 Contribution

The contribution of this work starts already with the analysis. For the specific data of chimney fires in Twente, we investigated which explanatory variables describe these fires. Let's call these explanatory covariates here key variables. These key variables help us in building a good prediction model, but knowing the variables can also help us to direct preventive actions. Building the model is factorised into two steps, first fitting an inhomogeneous Poisson process and after that the characteristics of the random field corresponding to the Log Gaussian Cox process. To give more detail about the accurateness of the model also the matching confidence bounds and residuals are computed.

Additionally, during the process the field of point processes is studied which is a (relatively) new field in mathematics. With broadening my mathematical knowledge, and getting used to the procedures at the fire department, expertise was needed in R, QGIS and Microsoft PowerBI. Beside the theoretical part, also more background about the work of a Twente fire station, including fire suppression, helped the process.

The final result of this work can be used in the daily life of the firemen because the prediction software tool developed in R and PowerBI can be used by every worker of the fire department. With the tool, the fire department can anticipate on the expected number of chimney fires to plan the daily life of the firemen, for example the relocation of firemen and the distribution of the day work a firemen can handle beside providing help to the citizens of the neighbourhood. Beside that, the tool represents by itself the possibilities of Business Intelligence in an organisation as the fire department.

1.3 Outline

The structure of the report is as follows. Chapter 2 involves a background and including the basic properties of the Log Gaussian Cox process. This includes framework about point processes, Poisson processes and the extension to the Log Gaussian Cox process. Earlier work concerning these processes which is used in this research is also elaborated on in this chapter. The motivation for the choice of the used processes is deeply discussed with the help of a distance analysis performed in Chapter 3. The fitting procedure and the results of the inhomogeneous Poisson process are described in Chapters 4 and 5 and this procedure is expanded or the Log Gaussian Cox process in Chapters 6 and 7. The dashboard which is developed for the fire department is elaborated in Chapter 8, thereafter Chapter 9 closes the report with a conclusion and recommendations for future research.

2 Background of the Log Gaussian Cox process

As explained in the introduction, a Log Gaussian Cox process combines the good properties of the inhomogeneous Poisson process with a random field which is included to extend the model to also inherit spatially dependent noise. To improve the predictions, we continue in this thesis with Log Gaussian Cox processes. In this section we present the background and a more extensive reasoning behind applying Log Gaussian Cox processes. Also the inhomogeneous Poisson process will be explained in more detail.

2.1 Background

Let's take a closer look on the Log Gaussian Cox process, i.e. Cox processes where the logarithm of the intensity surface is a Gaussian process, see [16]. As said in the introduction, with this model it is possible to model where clusters of events will appear and where not, which is called the stochastic interaction between events. The inhomogeneous Poisson process is still present as a foundation but we include another variable which models this interaction. Together they involve the connection between covariates and events and, if available, the interaction between events. This last part models whether new events are happening close to these events or definitely not. To achieve that, a random field is introduced which we will define first. Formally, a Gaussian random field is in [6] defined as

Definition 1. *A random field $\epsilon(x) \in \mathbb{R}$ is a Gaussian random field if for every finite number $m \in \mathbb{N}$, $\epsilon(x_1), \cdots, \epsilon(x_m)$ is multivariate normal for any $x_i \in M \subset \mathbb{R}$.*

As mentioned before, the inhomogeneous Poisson process can be used as a basis and a random field can be added to create a Log Gaussian Cox process. The intensity of such a process is

$$\lambda(u) = \exp[Y_u] = \exp[C_u] \exp[W_u] \tag{1}$$

with $u \in \mathbb{R}^d$ and where Y_u is a Gaussian random field with a mean function $m(\cdot)$ and a covariance function $\rho(\cdot, \cdot)$ which can be chosen to fit the data. The $m(\cdot)$ actually represents the basis of the model, and we can say that this part represents the underlying process. The $\rho(\cdot, \cdot)$ represents the stochastic interaction. When we choose $\rho(\cdot, \cdot) = 0$, no interaction is considered and the model reduces to the inhomogeneous Poisson process Wendels used. We include the covariance function to predict the (human) noise which cannot be described by covariates.

The second part of Equation (1) represents the factorisation of the Log Gaussian Cox process into the inhomogeneous Poisson process represented by C_u and the correlation represented by W_u. The deterministic C_u does not involve a correlation function and reduces therefore to only a non constant intensity function. The W_u is a Gaussian random field which includes the correlation function and therefore will describe the noise. The field can be interpreted as a stochastic process taken values according to the multivariate normal distribution. Both C_u and W_u (can) include a part of the mean function from Y_u which means that W_u is a Log Gaussian Cox process on itself.

Summarising, the Log Gaussian Cox process can be factorised into an inhomogeneous Poisson part and a random field part. The Poisson part takes certain covariates into account, just as in the research [21] and the human behaviour/noise will be modelled in the random field.

2.2 Point Processes

In the preceding we have talked about the background of the Log Gaussian Cox process. To understand this process completely, we will elaborate in this section on the C_u from Equation (1). Therefore the (Poisson) point processes will be defined formally.

2.2.1 Definition

Point processes are defined to be processes where events occur on random locations. It creates a set of mathematical points (locations) irregularly distributed within a designated region and generated by a kind of stochastic mechanism, see [16]. In most applications the designated region is the two-dimensional Euclidean plane. One can think of a lot of examples, such as the epicentres of earthquakes, outbreaks of forest fires and also, in our case, the outbreaks of chimney fires.

Before we can define a point process in a formal way, the following definition is needed. Both definitions in this subsection are taken from [15].

Definition 2. *The family $N^{lf}(\mathbb{R}^d)$ of locally finite point configurations in $\mathbb{R}^d$ consists of all subsets $\boldsymbol{x} \subseteq \mathbb{R}^d$ that place finitely many points in every bounded Borel set $A \subseteq \mathbb{R}^d$.*

This formal definition is included to exclude some extreme cases while the possibility that a lot of the variables we use are in such an extreme case is filtered out by only looking at the family given above. The locally finite point configurations of which is spoken in the definition above *can* contain multiple points, so points on the same location. This is not necessary but depends on the considered process. A point process can then be defined according to the following definition.

Definition 3. *A point process $X \in N^{lf}(\mathbb{R}^d)$ on $\mathbb{R}^d$ is a random locally finite configuration of points such that for all bounded Borel sets $A \subseteq \mathbb{R}^d$ the number of points of X that fall in A is a finite random variable which we shall denote by $N_X(A)$.*

Point processes exist in multiple forms. Previously we spoke of locations where events occur but we can also look at a time line and check when events occur. These two cases can also be combined into a spatio-temporal point process which thus checks location and time. These cases will be explained in more depth in the next subsections.

2.2.2 Poisson process

The Poisson process is one of the easiest point process to work with because of its strong independence properties. In this report we consider two types of Poisson processes, the homogeneous and the inhomogeneous Poisson processes. Both of the processes will be tested for the best fit on the data in Chapter 3.

The homogeneous Poisson process is described by an intensity which is constant. In terms of point processes, the number of events occurring will increase with a constant intensity when $|A|$ increases. Formally, as described in [4] and [15],

Definition 4. *A point process X on $\mathbb{R}^d$ is a homogeneous Poisson process with intensity $\lambda > 0$ if*

- *$N_X(A)$ is Poisson distributed with mean $\lambda|A|$ for every bounded Borel set $A \subseteq \mathbb{R}^d$;*
- *for any k disjoint bounded Borel sets $A_1, ..., A_k$, $k \in \mathbb{N}$, the random variables $N_X(A_1), ..., N_X(A_k)$ are independent.*

Recall here that the property discussed at the second bullet point is strict, it implies the independent behaviour of the occurring events.

The most important difference between the inhomogeneous and homogeneous Poisson process is that the intensity is not constant any more. There exist more differences but they are a result from this change in intensity. In our case for predicting chimney fires, we are dealing with an intensity function where the intensity can change over time and space. In the above definition the $\lambda|A|$ mentioned in the first bullet point can be replaced by

$$\int_A \lambda(x)dx$$

for an integrable function $\lambda : \mathbb{R}^d \to \mathbb{R}^+$. With this replacement the inhomogeneous Poisson process can be defined as well:

Definition 5. *A point process X on $\mathbb{R}^d$ is an inhomogeneous Poisson process with intensity function λ if*

- *$N_X(A)$ is Poisson distributed with mean $\int_A \lambda(x)dx$ for every bounded Borel set $A \subseteq \mathbb{R}^d$;*
- *for any k disjoint bounded Borel sets $A_1, .., A_k$, $k \in \mathbb{N}$, the random variables $N_X(A_1), ..., N_X(A_k)$ are independent.*

When we consider a spatio-temporal Poisson process, the Borel set A describes two sets, time T and space S which results in

$$\int_A \lambda(y)dy = \int_T \int_S \lambda(x,t)dxdt$$

with $y = (x, t)$.

2.2.3 Confidence intervals

For a spatio-temporal inhomogeneous Poisson process, the probability of n events occurring in a time period $(a, b]$, is defined as:

$$P\{N(a,b] = n\} = \frac{[\Lambda(a,b)]^n}{n!} e^{-\Lambda(a,b)} \tag{2}$$

where

$$\Lambda(a,b) = \int_a^b \int_S \lambda(x,t)dxdt.$$

With the expected number of events, say E_n, being the center of a confidence interval, the corresponding $p\%$ confidence interval for time period $(a, b]$ has bounds $(c_1 = E_n - c, c_2 = E_n + c)$ where for c the following holds

$$\sum_{i=c_1}^{c_2} P\{N(a,b] = i\} = p\%. \tag{3}$$

2.3 Log Gaussian Cox processes: Elementary properties

In this subsection we will explain the Log Gaussian Cox process and show some important properties.

2.3.1 Principle

As said in Section 2, Cox processes are an extension of the inhomogeneous Poisson process. Because of [21] we have reason to believe that the inhomogeneous Poisson process will not fit the data completely. For the modelling of chimney fires it therefore seems a logical choice to use this model. Before explaining the Log Gaussian Cox process completely, let's dive first into the general Cox process as defined by [1].

Definition 6. ***X*** *is a Cox process driven by the random intensity function Z if, conditional on $Z = z$, **X** is an inhomogeneous Poisson process with intensity function z.*

The random intensity function is here defined as $Z = \{Z(u) : u \in \mathbb{R}^d\}$, which is a locally integrable, non-negative random field (recall Definition 1 in Section 2.1). This Z can be applied to certain random fields to get special properties. Particular properties are captured by summary statistics

such as intensity, product density and pair correlation function. These functions are shown in Equation (4), (5) and (6) respectively.

$$\rho(u) = \mathbb{E}[Z(u)] \tag{4}$$

$$\rho^{(2)}(u,v) = \mathbb{E}[Z(u)Z(v)] \tag{5}$$

$$g(u,v) = \frac{\mathbb{E}[Z(u)Z(v)]}{\mathbb{E}[Z(u)]\mathbb{E}[Z(v)]}. \tag{6}$$

The first two equations are known as the first and second order product densities, coming from the nth order product densities $\rho^{(n)}$. Intuitively, $\rho^{(n)}(u_1,..,u_n)du_1 \cdots du_n$ is the probability that the Cox process has a point in each of n infinitesimally small disjoint regions of volumes $du_1,...,du_n$. For the intensity for example we can translate this to the number of points in our Cox process per unit u.

When we now consider the Log Gaussian Cox process, our random intensity function is constructed as the exponential of a Gaussian random field, so $Z(u) = \exp[W_u]$ where W_u, $u \in T \subseteq \mathbb{R}^d$, is a random field. We define a Borel measure Λ as the integral of the random intensity function Z, which in this case can be described as follows:

$$\Lambda(A) = \int_A \exp[W_u]du$$

where $A \subseteq T$, see [15].

2.3.2 Properties of the Log Gaussian Cox process

Here we shall discuss some properties of the Log Gaussian Cox process, starting with the moments of a lognormal distribution. At first we derive the moments of the origin which we use to calculate the moments about the mean in the second part. When we know the moments corresponding to this distribution, we also know the moments of our Log Gaussian Cox process and with these moments we can derive the intensity, second order product density and the pair correlation function.

Say we have the Cox process X which is log Gaussian with the exponential random intensity function $Z(u)$ as defined in the previous subsection. We thus apply the intensity function to a Gaussian field. The properties discussed above still hold and we end up with $Y = \log X$ is Gaussian distributed with mean μ and variance σ^2 so $X = e^Y$ with Y Gaussian.

Theorem 1. *The Cox process $X = e^Y$ with mean μ and variance σ^2 has a mean number of events corresponding to*

$$\mathbb{E}[X^k] = \exp[\mu k + \frac{1}{2}k^2\sigma^2].$$

Proof. This Cox process has an exponential random intensity function, so

$$\begin{aligned}
\mathbb{E}[X^k] &= \mathbb{E}[e^{kY}] \\
&= \int_{-\infty}^{\infty} \exp[ky] \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] dy \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[ky - \frac{(y-\mu)^2}{2\sigma^2}\right] dy.
\end{aligned}$$

The exponential can then be rewritten as follows:

$$\begin{aligned} ky - \frac{(y-\mu)^2}{2\sigma^2} &= \frac{2\sigma^2 ky - (y^2 - 2\mu y + \mu^2)}{2\sigma^2} \\ &= -\frac{1}{2\sigma^2}(y^2 - 2(\mu + \sigma^2 k)y + \mu^2) \\ &= -\frac{1}{2\sigma^2}(y^2 - 2(\mu + \sigma^2 k)y + \mu^2 + (\mu + k\sigma^2)^2 - (\mu + k\sigma^2)^2) \\ &= -\frac{1}{2\sigma^2}(y^2 - 2(\mu + \sigma^2 k)y + (\mu + k\sigma^2)^2) + \mu k + \frac{1}{2}k^2\sigma^2. \end{aligned}$$

This expression can then be integrated in the above formula which gives the following results.

$$\mathbb{E}[X^k] = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} \exp[ky - \frac{(y-\mu)^2}{2\sigma^2}]dy = \exp[\mu k + \frac{1}{2}k^2\sigma^2]. \tag{7}$$

Here we used that the integrand in

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(y - (\mu + \sigma^2 k))^2]dy$$

is again a normal density function with mean $\mu + \sigma^2 k$ and thus the integral of this function is equal to one. □

Now we know the formula for these moments, the following corollary exist for the moments about the mean. We will only look at the second and third moments, denoted by μ_2 and μ_3.

Corollary. *The second and third moment about he mean of a log Gaussian Cox process are defined as*

$$\begin{aligned} \mu_2 &= \mu_2' - (\mu_1')^2 \\ &= \exp[2\mu + \sigma^2](\exp[\sigma^2] - 1) \qquad (8) \\ \textit{and } \mu_3 &= \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 \\ &= \exp[3\mu + \frac{3}{2}\sigma^2](\exp[\sigma^2] - 1)^2(\exp[\sigma^2 + 2). \qquad (9) \end{aligned}$$

Theorem 2. *The pair correlation function of the Log Gaussian Cox process with mean $m(u)$ and variance $\rho(u,u)$ with $u \in T$ where T is a Borel set is*

$$g(u,v) = \exp[\rho(u,v)].$$

Proof. In our log Gaussian Cox process, we have our Gaussian field $Y = \{Y(u) : u \in \mathbb{R}\}$ which is distributed with mean $m(u)$ and variance $\rho(u,u)$ with $u \in T$ where T is a Borel set. The random intensity function we consider is then $Z(u) = \exp[Y(u)]$. The first moment ($k = 1$) of the distribution can then be derived from Equation (7).

$$\begin{aligned} \rho(u) = \mathbb{E}[Z(u)] &= \mathbb{E}\exp[Y(u)] \\ &= \exp[m(u) + \frac{1}{2}\rho(u,u)]. \end{aligned} \tag{10}$$

The second order product density as shown in equation (5) is then calculated as the expectation of two Gaussian fields, namely $\exp[Y(u) + Y(v)]$ where $u, v \in T$. We know here that $Y(u) + Y(v)$ is again Gaussian distributed with mean $m(u) + m(v)$ and variance $\frac{1}{2}(\rho(u,u) + 2\rho(u,v) + \rho(v,v)$ and we end up at the following second moment:

$$\begin{aligned} \rho^{(2)}(u,v) = \mathbb{E}\exp[Z(u)Z(v)] &= \mathbb{E}\exp[Y(u) + Y(v)] \\ &= \exp[m(u) + m(v) + \frac{1}{2}(\rho(u,u) + 2\rho(u,v) + \rho(v,v))]. \end{aligned} \tag{11}$$

With $\rho(u)$ and $\rho^{(2)}$ defined in Equation (10) and (11) respectively, the pair correlation function as in Equation (6) can then easily be derived:

$$\begin{aligned} g(u,v) &= \frac{\mathbb{E}[Z(u)Z(v)]}{\mathbb{E}Z(u)\mathbb{E}Z(v)} \\ &= \frac{\rho^{(2)}(u,v)}{\rho(u)\rho(v)} = \exp[\rho(u,v)]. \end{aligned} \tag{12}$$

□

The only function that needs to be defined to complete the description of the Log Gaussian Cox model is $\rho(u,v)$. There are different choices for this covariance function available, such as exponential, Matérn etcetera, see [19]. These functions all have different parameters which need to be fitted to the data to complete the model. The specific type of stochastic interaction between events is then defined and the model can be used for prediction. The fitting of this covariance function will be done later on in this report.

2.3.3 Simulations

To gain more insight in the Log Gaussian Cox process, we did some simulations of this process. In the process defined in the previous subsection we did not specify $\rho(u,v)$ yet. In these simulations an exponential covariance function is chosen. This function is one of the most chosen covariance functions and has the form:

$$\rho(u,v) = \sigma^2 \exp\left[-\frac{||u-v||}{\beta}\right].$$

To specify the behaviour of this covariance function, we need both parameters variance (σ^2) and scale (β). Simulations are made for three different values of both parameters to show how these parameters through the covariance function influences the behaviour of the Log Gaussian Cox process. These plots are shown in Figure 2 and were made with the `R`-package `spatstat`. A low β indicates clearly a very random plot, which can be influenced by σ. For fixed σ^2 we see that the higher the β the less and more dense clusters. The β seems therefore to indicate the foundation of the image where σ comes in to increase the density of the clusters. This is clearly visible in the increasing numbers on the ribbon on the right of the image. Remember that the Log Gaussian Cox process is added to include some clustering and dependence between events so this clustering is important behaviour for the rest of this report.

In the fitting of the covariance function for the chimney fire data, we want to extend these simulations, when also taking an intensity function into account, to a plot in which we can recognise the point pattern given in Figure 1b.

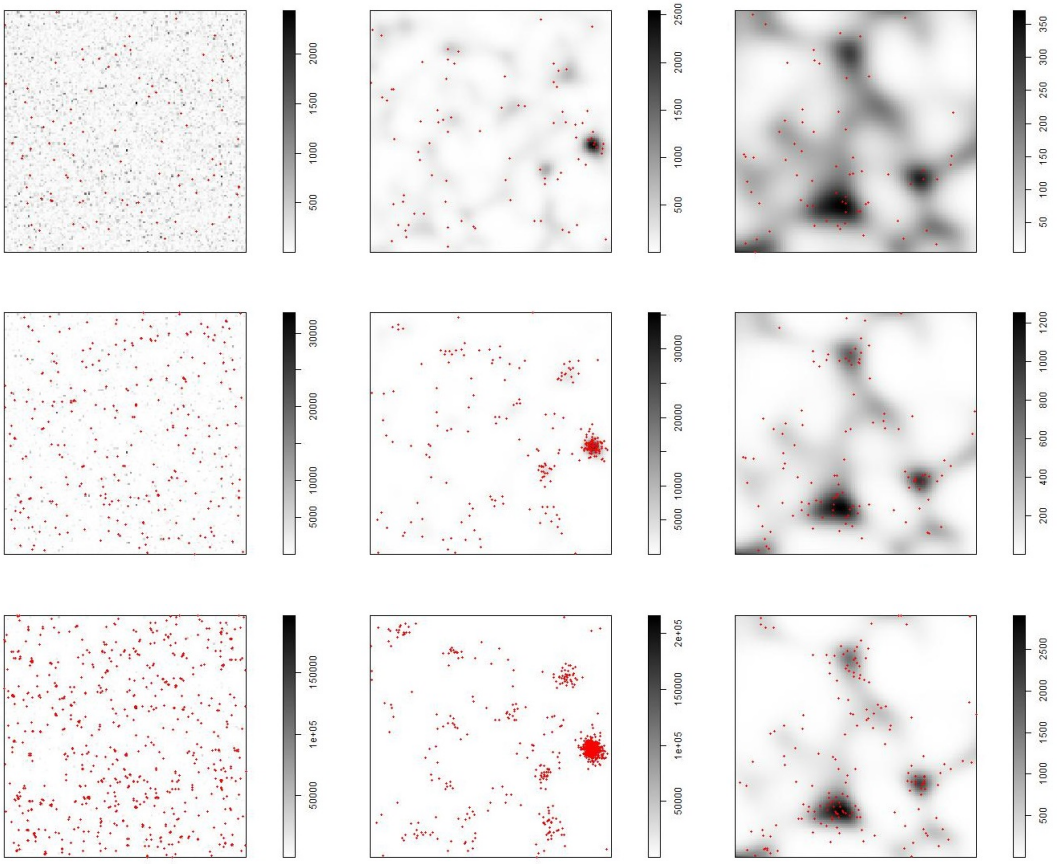


Figure 2: Realizations of the Log Gaussian Cox process with an exponential correlation function with varying variance and scale parameter but fixed mean of 4.25. For every row, the variance differs with σ^2=1, 3, 5 and over the columns, the scale differs with β=0.005, 0.075, 0.145.

3 Distance analysis

Before continuing to the Log Gaussian Cox process, first the (in)homogeneous Poisson process is tested to the data. The suggestion to extend the Poisson process to the Log Gaussian Cox process was based on the data in general and because we are focussing specific on chimney fires, we confirm with a distance analysis if this process is the right choice, see [7] and [21]. The distance analysis is one way of testing the pattern of the data and thus which model can result in a good fit. We will check for two different patterns and will therefore do the distance analysis twice, one to check for a homogenous Poisson process and after that for an inhomogeneous Poisson process because this gives us simply more information about the behaviour of the data. If none of the Poisson processes fits the data according to this method, we will extend the model to a Log Gaussian Cox process. In either case, a Poisson process will be fitted to our data because of the idea behind Equation (1) as described in Section 2. When the distance analysis confirms that one of the models describes the data well, we can stop there. If not, we can fit the Gaussian field W_u from Equation (1) and add this to the Poisson C_u which then results in the Log Gaussian Cox process Y_u.

First the distance analysis will be explained in Section 3.1 and this will then be applied to homogeneous and inhomogeneous empirical functions in Section 3.2 and 3.3 respectively. Finally we will perform a distance analysis based on the temporal data as well in Section 3.4.

3.1 Explanation of the distance analysis

The principle of the distance analysis is to check if the spatial point pattern behaves according to three classifications, as defined by [7]:

- A spatial point pattern with no obvious interaction structure is called *completely spatially random*, often abbreviated as *CSR*;
- A spatial point pattern with a structure in which points tend to cluster together is called *aggregated*;
- A spatial point pattern with a structure in which points tend to be evenly distributed is called *regular*.

With these classifications, the type of model which would describe our data the best can be extracted. When the pattern is considered to be CSR, the data is likely to fit a Poisson process, both homogeneous and inhomogeneous. To check both models, we use different functions to describe the pattern, which will be explained later in this section. An aggregated pattern is a clustered pattern which would indicate a Cox process while a regular pattern indicates other processes which we do not consider here. The goal of our analysis is thus to decide if the spatial point pattern is CSR, aggregated or regular to find the best model to fit on the data.

Consider here a spatial point pattern as defined in [7] and [20] so we have a data set $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, $\mathbf{x}_i \in T, 1 \leq i \leq n$, distributed within the region of interest $T \subset \mathbb{R}^m, m \in \mathbb{N}$. One assumption needed for this analysis is that the process is stationary and isotropic, which implies that $\rho^{(2)}(u, v)$ is only dependent on the distance between u and v, which we call $r = ||u - v||_2 = ||u - v||$. In formula:

$$\rho^{(2)}(u, v) = \rho^{(2)}(||u - v||) = \rho^{(2)}(r). \tag{13}$$

Later it will become clear why we need this assumption.

We want to set boundaries in which the process still fits a CSR pattern. These boundaries are dependent on the distance measure function which we will elaborate on later in this report but they will already be defined here.

Definition 7. *Let $A \subset \mathbb{R}^2$ be the region of interest, $\tilde{S}_1, \tilde{S}_2, \ldots, \tilde{S}_n$, n newly sampled CSR spatial point patterns in A and $\hat{f}_i(r)$ be the empirical function representing the chosen distance measure*

of interest for $\tilde{S}_i, 1 \leq i \leq n$. Then the upper critical (simulation) envelope $U(r)$ and the lower critical (simulation) envelope $L(r)$ for S are defined as

$$U(r) = \max_{i=1,2,\ldots,n} \hat{f}_i(r) \tag{14}$$

$$L(r) = \min_{i=1,2,\ldots,n} \hat{f}_i(r) \tag{15}$$

respectively.

These envelopes give a boundary in which the spatial point pattern can still be considered as a CSR distributed pattern, with a certain tolerance. This tolerance is created by estimating the empirical functions multiple times. In this report we use a significance of 95% which results in 39 simulations, see [20], so $n = 39$ in the above definition.

In the next part of this report four choices of $\hat{f}$ will be explained. As can be seen in the following subsection, all functions are dependent on the difference r between events or locations (depending on the used function) and thus the distance analysis method is also dependent only on this r. To confirm this in our data, we assume that the product density function is stationary and isotropic (as we already did in Equation (13)) because this results in only a dependence on r. The following step is to measure the information contained in S with the empirical functions $\hat{f}(r)$. For the CSR pattern, we also have a theoretical value of these functions, which is contained in the functions $f(r)$. the comparison of the theoretical value and the estimated value gives us an indication of the pattern the data follows. If the analysis results in a significant difference between $\hat{f}(r_0)$ and $f(r_0)$ for some predetermined value r_0 for r and a significance level α, the analysis concludes a point pattern which is not CSR distributed. The significance level is then contained in the calculation of the critical envelopes which we defined in Definition 7.

For checking the homogeneous Poisson process, we consider homogeneous measure functions f, which assume homogeneity. When we check the inhomogeneous Poisson process, inhomogeneous functions need to be used to make sure that the inhomogeneity properties are assumed, such as the possibility of a spatial/time dependent intensity function. The homogeneous functions are explained first and the inhomogeneous functions are defined in a similar way which is specified later in this section. The envelopes can be interpreted as the boundaries of the region in which a spatial point pattern is still CSR and thus follows a homogeneous or inhomogeneous Poisson process, depending on the empirical functions used. Beside that, an envelope plot concerning the homogeneous distance analysis can also show insight in the other two classifications. We explain this also later in this section.

3.1.1 Distance analysis functions

Ripley's reduced second moment function $K(r)$
The first function we want to discuss is *Ripley's reduced second moment function* $K(r)$. This function chooses an arbitrary event and checks the number of events within a distance r from that particular event. This function is a characterization of the second order properties of the process we are considering, so when the differences are in for example the third or fourth order of the function, this function is not reliable. In formula:

$$K(r) = \lambda^{-1}\,\mathbb{E}[\text{number of other events within distance r of an arbitrary event}] \tag{16}$$

and when we assume a CSR pattern, this reduces to:

$$K(r) = \lambda^{-1}\pi r^2 \lambda = \pi r^2 \tag{17}$$

because under CSR, the expected number of events in an area is the intensity multiplied with this area. For a clustered process, around one event there are a lot of other events so then $K(r) > \pi r^2$

while for a regular process the opposite $K(r) < \pi r^2$ is the case. Because of the assumption of stationarity, the selection of the 'arbitrary' event does not matter.

Nearest neighbour distance distribution function $G(r)$
The second function we consider is the *nearest neighbour distance distribution function* $G(r)$, which uses the distance between an event and its nearest neighbouring event. Excess of small nearest neighbour distances tells us that probably clusters exists in this data which is characteristic for an aggregated pattern. Deficiency of these distances points to a lack of clusters so a regular process. With $d(x, X\backslash\{x\})$ the distance between event x and the other points of the point process X, we have

$$\begin{aligned} G(r) &= P[\text{distance from an arbitrary event of } X \text{ to the nearest other event of } X \text{ is at most } r] \\ &= P(d(x, X\backslash\{x\}) \leq r | \forall x \in X) \end{aligned}$$

When we assume a CSR pattern, and thus the probability distribution given above, this reduces to:

$$G(r) = P(N(ball(o, r) > 0)) = 1 - P(N(\pi r^2) = 0) = 1 - e^{-\lambda \pi r^2}$$

where $N(B)$ represents the number of events in set B of area $|B|$ as above explained.

Empty space function $F(r)$
The third analysis we do is based on the *empty space function* $F(r)$, in which the distances between an arbitrary point and its nearest event are analysed. We mean with an arbitrary point a random location in the region of interest, so this is not related to events. When these distances are small, there are always events close to an arbitrary point so probably a regular pattern is considered, while large distances indicate a more aggregated pattern. Let y be an arbitrary point in the region of interest and X here the point pattern itself, then

$$F(r) = P(d(y, X) \leq r)$$

When we assume a CSR pattern, with the same reasoning as function $G(r)$, we end up at

$$F(r) = P(N(ball(o, r)) > 0) = 1 - P(N(ball(o, r)) = 0) = 1 - e^{-\lambda \pi r^2}$$

where $N(B)$ represents the number of events in a set B of area $|B|$ as explained above.

Summary function $J(r)$
The fourth and final function we use for distance analysis is the *summary function* $J(r)$. This function is based on the $G(r)$ and $F(r)$ function described above and reads $J(r) = \frac{1-G(r)}{1-F(r)}$. For a CSR pattern, the J-function is identically equal to 1. Values $J(r) < 1$ or $J(r) > 1$ typically point to an aggregated or regular pattern, respectively.

With the help of the estimator $\hat{\lambda} = |S||A|^{-1}$ where $|S|$ is the number of events in S, the previous four functions can be estimated, see for different estimations [13]. Depending on the confidence interval, a certain amount of simulations are performed. These simulations together form the envelope plot with $U(r)$ and $L(r)$ from Equations (14) and (15) respectively.

Concluding, for the plots with function $\hat{J}(r)$ we check if

$$\begin{aligned} J(r) &= 1 && \rightarrow \text{CSR} \\ J(r) &< 1 && \rightarrow \text{aggregated} \\ J(r) &> 1 && \rightarrow \text{regular} \end{aligned}$$

Envelope plots with the estimated functions $\hat{K}(r)$ and $\hat{G}(r)$ give the following:

$$\begin{aligned} L(r_0) \leq \hat{f}(r_0) \leq U(r_0) &\quad \rightarrow \text{CSR} \\ \hat{f}(r_0) > U(r_0) &\quad \rightarrow \text{aggregated} \\ \hat{f}(r_0) < L(r_0) &\quad \rightarrow \text{regular} \end{aligned}$$

while envelope plots with function $\hat{F}(r)$ give the following:

$$\begin{aligned} L(r_0) \leq \hat{f}(r_0) \leq U(r_0) &\quad \rightarrow \text{CSR} \\ \hat{f}(r_0) > U(r_0) &\quad \rightarrow \text{regular} \\ \hat{f}(r_0) < L(r_0) &\quad \rightarrow \text{aggregated} \end{aligned}$$

For more information on these four functions, see [7].

3.2 Distance analysis for homogeneity

For the estimation of the above functions we need an estimator of the intensity, which we define first. After that, the actual distance analysis for finding the pattern under homogeneity assumptions is executed.

3.2.1 Intensity estimation

For the estimation of the above functions and thus to complete the distance analysis, we need an estimator of the intensity. Let S be the spatial or temporal point pattern of interest, where $|S|_N = n_v$ the number of events, and A the region of interest. Partition A in k polygons B_i, where $i \in \{1, .., k\}$, with the same area $|B|$. Then we have N_i the random variable indicating the number of events in B_i and n_i its realizations in S and we propose the following estimator for intensity function λ:

$$\hat{\lambda}(k) = \frac{\sum_{i=1}^{k} n_i}{k|B|}.$$

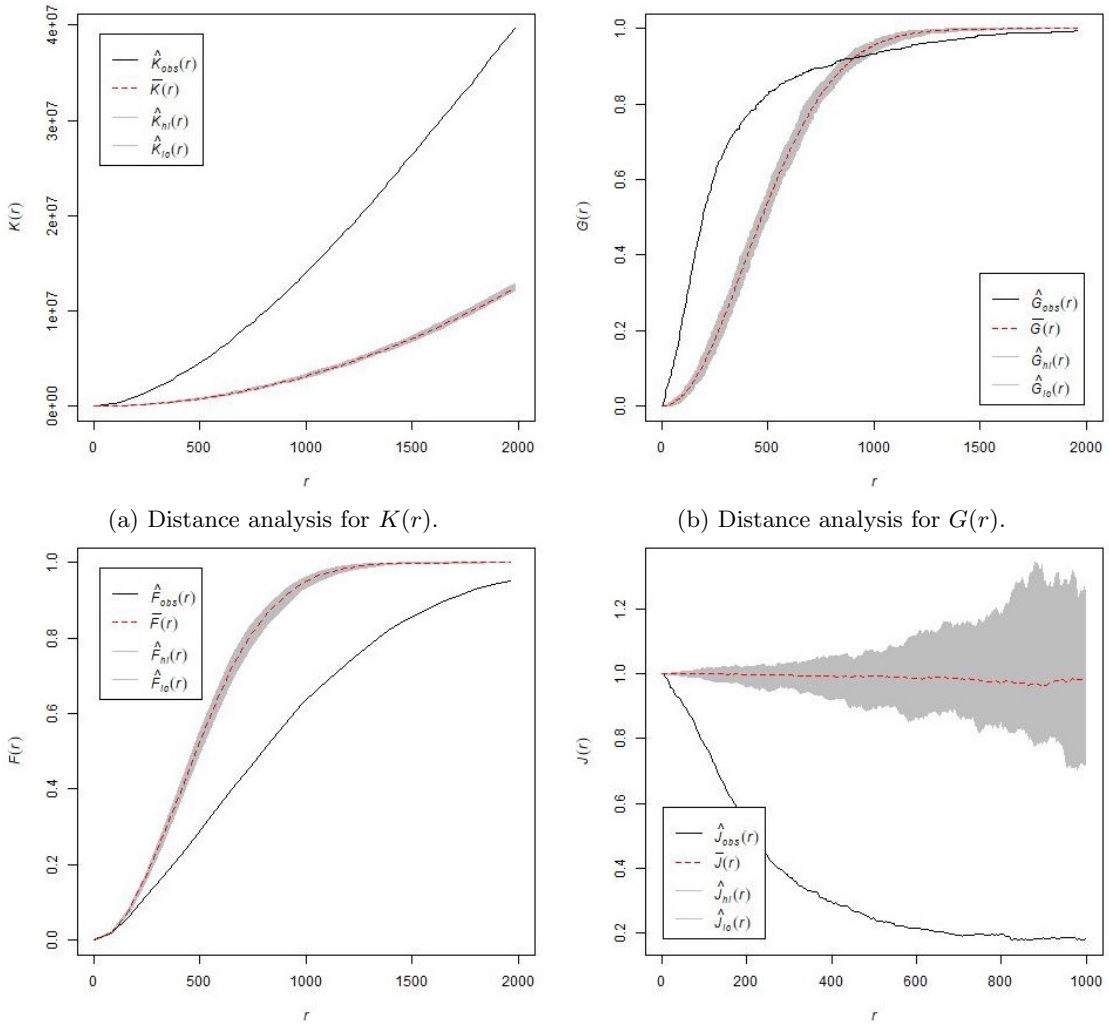
For a CSR pattern, N_i is Poisson distributed with mean $\lambda|B|$ so that $\mathbb{E}[\hat{\lambda}] = \lambda$. More formally, N_i is Poisson distributed with probability mass function:

$$P_n(|B|) = \exp(-\lambda|B|)\{\frac{\lambda|B|^n}{n!}\} \qquad \text{for } n = 0, 1, 2, ...$$

For the homogeneous distance analyses, we use $k = 1$ which results in $\lambda = |S|_N|A|^{-1}$, which will be used in the next functions.

3.2.2 Results

With the help of the package `spatstat` in `R`, the distance analysis including the above four homogeneous functions has been performed. For our chimney data, the spatial point pattern of interest is the region of Twente which we call S_m. The distance analysis is executed with a significance level of $\alpha = 0.05$ and the resulting plots are shown in Figure 3. We focused on the spatial point process of our data.
As follows from the explanation above, all four plots indicate clearly an aggregated pattern. This can be concluded because $\hat{K}(r)$ and $\hat{G}(r)$ lie above the envelopes and $\hat{F}(r)$ and $\hat{J}(r)$ lie below the envelope. We draw therefore the conclusion that the pattern does not follow a CSR pattern under homogeneity and therefore a homogeneous Poisson process does not describe the data well. This conclusion could have been expected because the occurrence of chimney fires is probably strongly dependent on the location. For example, the fires only occur in houses so the intensity should probably differ over the grassland and city centres in Twente. A constant intensity corresponding to a homogeneous Poisson process does not cover this change, while an inhomogeneous Poisson process can take that into account.



(a) Distance analysis for $K(r)$.

(b) Distance analysis for $G(r)$.

(c) Distance analysis for $F(r)$.

(d) Distance analysis for $J(r)$.

Figure 3: Distance analyses with the estimated functions $\hat{K}(r)$ (figure a, left above), $\hat{G}(r)$ (figure b, right above), $\hat{F}(r)$ (figure c, left below) and $\hat{J}(r)$ (figure d, right below) applied on the data of S_m, plotted against the corresponding theoretical functions (red striped line) and critical envelopes. The plots are provided by the package `spatstat` in `R`.

3.3 Distance analysis for inhomogeneity

So far, we have concluded that a homogeneous Poisson process is not a good fit. To continue our analysis, we will also test the point pattern for inhomogeneous Poisson properties. Testing inhomogeneity can be executed with four inhomogeneous empirical functions, namely $\hat{K}_{\text{inhom}}$, $\hat{G}_{\text{inhom}}$, $\hat{F}_{\text{inhom}}$ and $\hat{J}_{\text{inhom}}$ as in [2] and [14]. The difference between a homogeneous and inhomogeneous Poisson process is the different intensity estimator, while the explanation of these inhomogeneous functions stay as their homogeneous counterparts.

3.3.1 Intensity estimation

To generalise the analysis to non-stationary point processes, a new non-constant intensity estimator is used. Time is chosen to be fixed and the analysis will result in a plot displaying a cross-section of the actual spatio-temporal distance analysis. The estimator used here uses improved edge

correction as described by Diggle (1985). The intensity value at point u is defined by

$$\hat{\lambda}(u) = e(u) \sum_i k(x_i - u) w_i$$

where k is the Gaussian smoothing kernel, $e(u)$ is an edge correction factor and w_i are the weights, we will explain these in this order in the following.

The Gaussian kernel function k smooths the values by taking the average of neighbouring points, with a weighting factor according to the Gaussian function. To define the kernel k, a value for σ is chosen, which is taken as the standard deviation of the Gaussian kernel k. The σ can also be interpreted as the smoothing factor of the density. With a small σ the density differs a lot between two close locations. The higher the σ the smoother the dense areas become, where a really large σ results in one big clustered area.

The edge correction makes sure to eliminate the bias that is caused by edge effects. These edge effects can exists because we are checking a bounded window (Twente) but a small disc around an event close to the boundary can extend outside this window. This event is then not observable. The edge correction used in this estimation is the reciprocal of the kernel mass inside window W:

$$\frac{1}{e(u)} = \int_W k(v - u) dv$$

The weights corresponding to the data are calculated as follows: When a longitude latitude combination occurs more than once in our data we remove the second (and third, fourth etc) from our data and add a weight to the location corresponding to the number of incidents that took place on this particular point. The location of all other events occurring in the data are assigned a weight equal to one.

The intensity function is estimated by the function `density.ppp` with edge correction, a corresponding $\sigma = 1520$ and with a weight vector including the duplicated points. This particular σ value is chosen because it gives the clearest result in the sense that not only the hotspots are highlighted but the smaller towns as well. The corresponding image is shown in Figure 4.

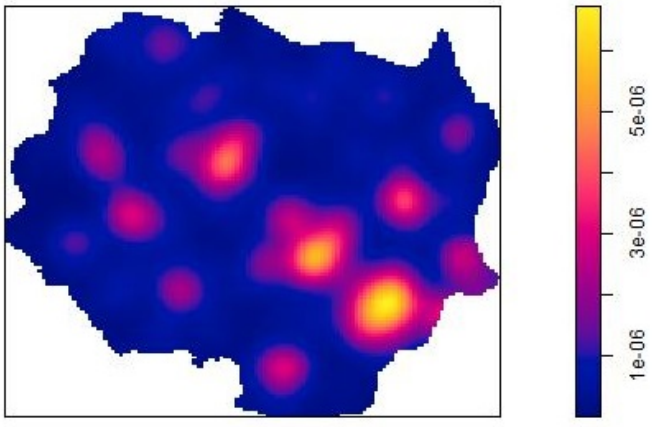


Figure 4: The fitted intensity function for our data with $\sigma = 1520$, taking weights and edge correction into account.

This image shows clearly the cities and towns in Twente, as follows from a comparison with Figure 1a. What particularly characterises this plot is the diagonal from the bottom right to the upper left, on which the larger cities Enschede, Hengelo and Almelo are placed.

3.3.2 Results

With this new intensity function, the empirical functions can again be estimated. The plots concerning the empirical functions with a non constant intensity function are shown in Figure 5.

At first sight the absence of a CSR pattern is clearly visible, namely the estimated inhomogeneous functions do not lie inside the envelopes. Figure 5a shows that for $r < 1700$ clustering is still present that is missing in the model. Figure 5b suggests the same for $r < 1000$ while Figure 5c indicates clustering for the whole interval, except maybe at the beginning. The summary function displayed in Figure 5d suggests a clustered pattern for $r < 1000$. At some points the estimated J-function hits the envelope but only enters the envelope after $r \approx 1000$. This gives us the information that for small distances, there is extra noise present and that chimney fires have a higher chance of appearing within 1 kilometre of a previous chimney fire. We do not have a researched explanation for this, but we can guess for example that in this area the people are maybe poorer and cannot afford to let someone clean their chimney or various other reasons.

An inhomogeneous Poisson process is thus also not the perfect fit for the data, and mostly because we miss a certain clustering for small distances. Adding the random field to inherit spatially dependent noise seems therefore a good extension of the model.

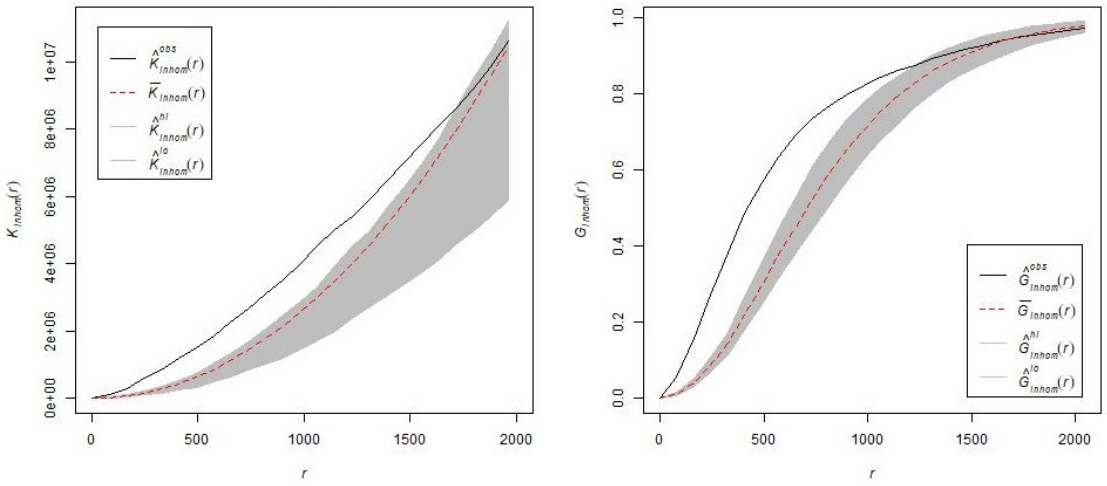
3.4 Application on temporal point pattern

Until now, we checked whether the data contains spatial clustering, but to complete the analysis also the temporal clustering must be checked which we will do in this subsection. We make use of the same K, F, G and J-functions. In the previous subsections we tested if events happen closely to each other in a spatial sense. With temporal clustering, we want to check if events happen close to each other in a temporal sense, so on the same day or a few days apart.

In the spatial analysis, the point pattern consisted of the x and y-coordinate of every event in the data and to make the distance analysis work, the events need to have two coordinates as well. The first coordinate is chosen to be a decimal number between 1 and 4381 which indicates the day and time in a year the event happened while, because for temporal clustering we can only check one simple difference (days in this case), the second coordinate is always set to zero. The time range is chosen like that because we are considering data from 1-1-2004 until 31-12-2015 which translates to 4380 days. For example an event which happened 20 February 2011 at 11:30:56, the first coordinate is calculated as follows. The 20th February is the 51st day of the year and the 2606th day of all the data we analyse (from 2004 until February 2011). The time 11:30:56 translates to the 42456th second of the day. With 86400 seconds in a day, the first coordinate is $2606\frac{42456}{86400}$. For all events we repeat this calculation and transform the events into a temporal point pattern, with which the temporal distance analysis can be performed. The distance r from the previous section is thus transformed to the distance in days between events. As in the spatial distance analysis, first the homogeneous functions will be analysed and after that also inhomogeneous analysis will be done.

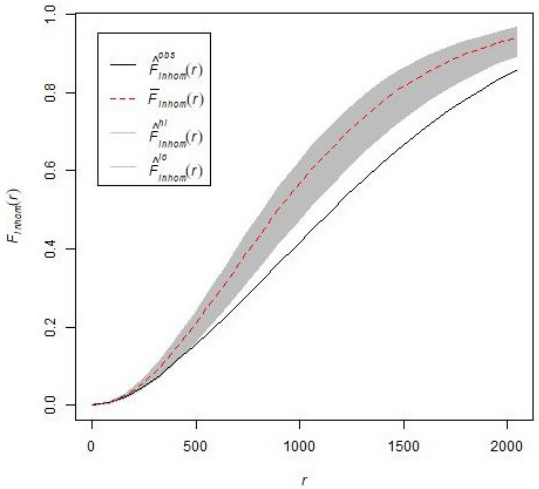
3.4.1 Homogeneous analysis

For the homogeneous functions, the analysis is done on the above explained temporal point pattern and the results are given in Figure 6. The estimated functions only exists for small r, with a maximum of 10 days apart. In all figures clearly a clustered pattern is recognised. In Figure 6c the distribution function bends and comes close to the envelopes. From these figures we can conclude that, when we assume homogeneity, also temporal clustering is still included in the data and thus a homogeneous Poisson process is also in temporal sense not a good fit.

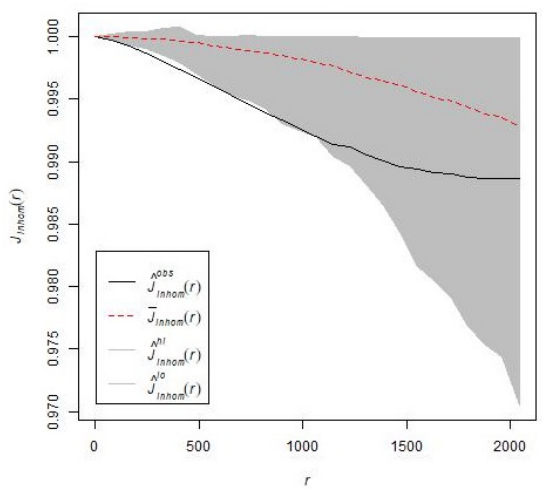


(a) Distance analysis for $K_{inhom}(r)$.

(b) Distance analysis for $G_{inhom}(r)$.

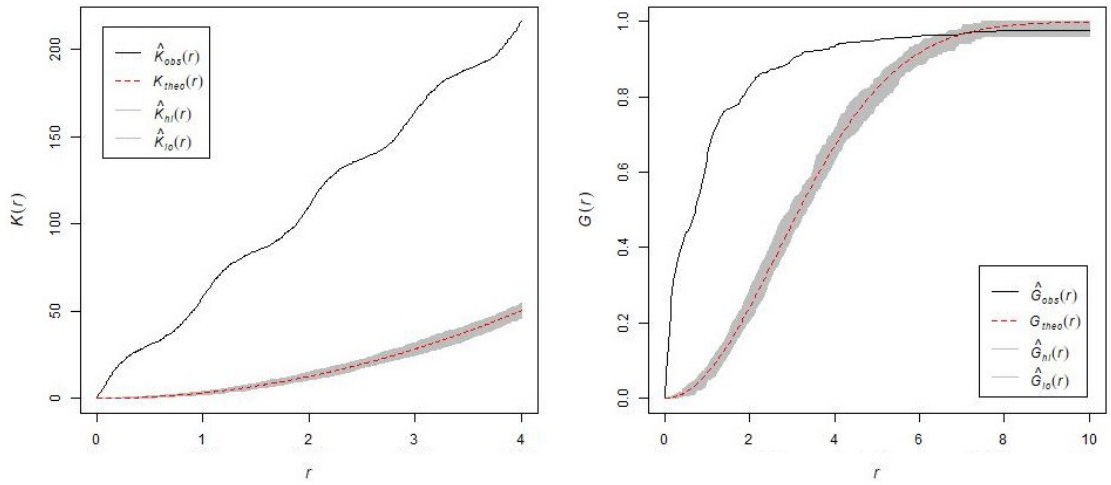


(c) Distance analysis for $F_{inhom}(r)$.



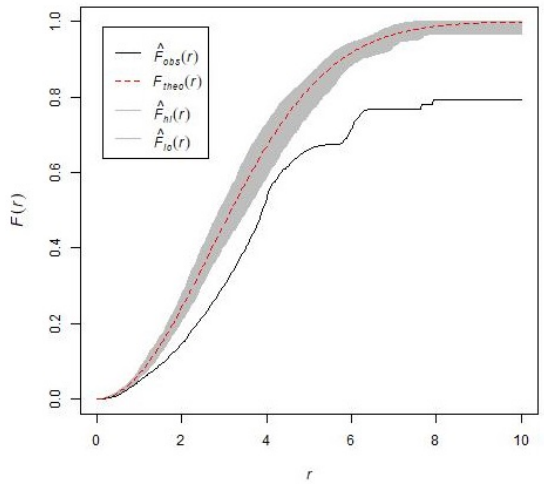
(d) Distance analysis for $J_{inhom}(r)$.

Figure 5: Distance analyses with the estimated functions $\hat{K_{inhom}}(r)$ (figure a, left above), $\hat{G_{inhom}}(r)$ (figure b, right above), $\hat{F_{inhom}}(r)$ (figure c, left below) and $\hat{J_{inhom}}(r)$ (figure d, right below) applied on the data of S_m with an intensity function fitted with `density.ppp`, plotted against the corresponding theoretical functions and critical envelopes. The plots are provided by the package `spatstat` in `R`.

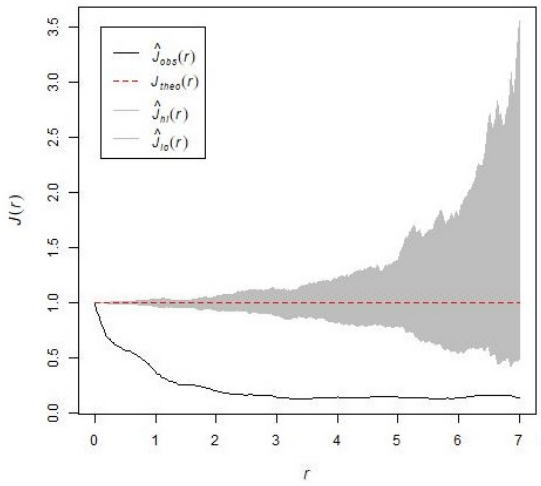


(a) Temporal distance analysis for $K(r)$.

(b) Temporal distance analysis for $G(r)$.



(c) Temporal distance analysis for $F(r)$.

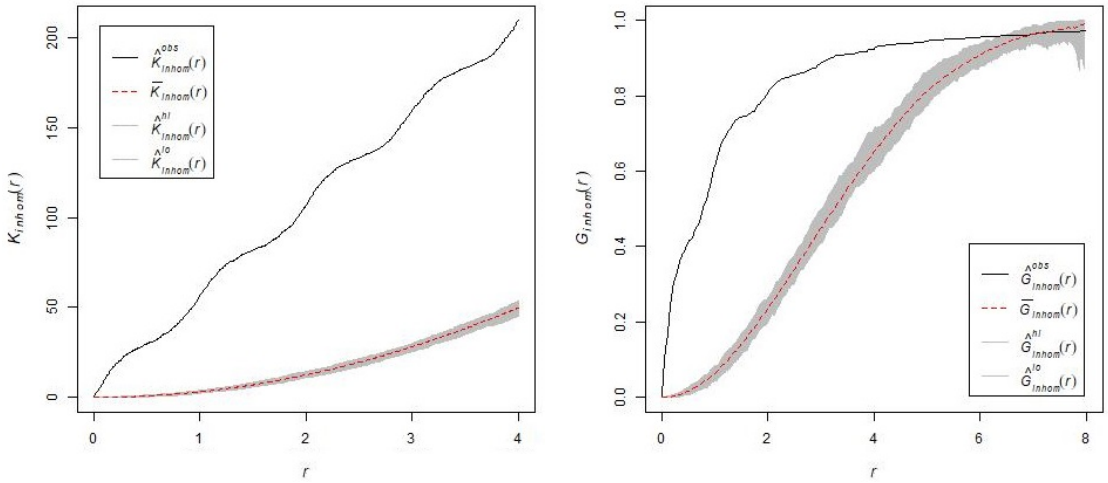


(d) Temporal distance analysis for $J(r)$.

Figure 6: Temporal distance analyses with the estimated functions $\hat{K}(r)$ (figure a, left above), $\hat{G}(r)$ (figure b, right above), $\hat{F}(r)$ (figure c, left below) and $\hat{J}(r)$ (figure d, right below) applied on the data of W_T, plotted against the corresponding theoretical functions (red striped line) and critical envelopes. The plots are provided by the package `spatstat` in `R`.

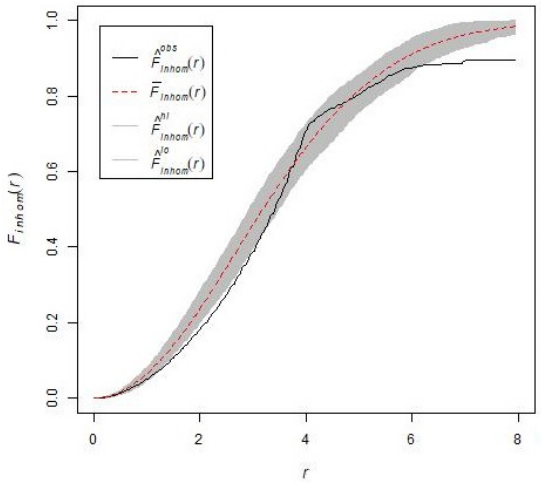
3.4.2 Inhomogeneous analysis

For the inhomogeneous analysis, the inhomogeneous functions explained earlier are used. The results of this analysis is shown in Figure 7. Figures 7a, 7b and 7d still imply clearly a clustered

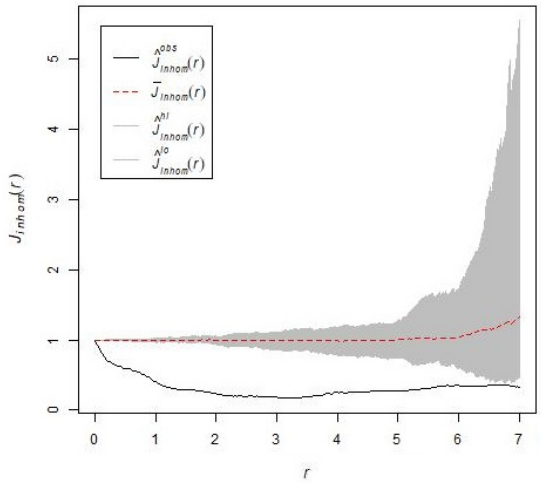


(a) Temporal distance analysis for $K_{inhom}(r)$.

(b) Temporal distance analysis for $G_{inhom}(r)$.



(c) Temporal distance analysis for $F_{inhom}(r)$.



(d) Temporal distance analysis for $J_{inhom}(r)$.

Figure 7: Temporal distance analyses with the estimated functions $\hat{K}_{inhom}(r)$ (figure a, left above), $\hat{G}_{inhom}(r)$ (figure b, right above), $\hat{F}_{inhom}(r)$ (figure c, left below) and $\hat{J}_{inhom}(r)$ (figure d, right below) applied on the data of W_T, plotted against the corresponding theoretical functions (red striped line) and critical envelopes. The plots are provided by the package `spatstat` in `R`.

pattern. Figure 7c indicates a weak clustering for small distances and regularity for larger ones, while the data bends inside the envelopes. In the previous subsection, all statistics indicate strong clustering and here, because three of the statistics out of four indicate clustering, we then still conclude a clustered pattern. When assuming inhomogeneity, we still conclude that clustering is included in the data and an inhomogeneous Poisson process is also in temporal sense not a good fit.

With this conclusion, we continue with fitting the models. Because of the factorisation we saw in Equation (1) of the Log Gaussian Cox process in a random field W_u and an inhomogeneous Poisson intensity function included in C_u, we will first fit this intensity function.

4 Fitting of the inhomogeneous Poisson process

To fit the inhomogeneous Poisson process, we use the procedure of [21] as a guideline. The Poisson process is only dependent on its non constant intensity function so our goal is to find the best fitting intensity function. In Section 4.1 the different covariates are lighted out and thereafter a correlation analysis is carried out in Section 4.2 to calculate the correlation coefficients which indicate the covariates that should be included in the intensity function. To complete the definition of the intensity function, a regression analysis is performed in Section 4.3. During this analysis we search for the function which describes the connection between the covariate and the data the best.

4.1 Spatial and temporal covariates

In cooperation with the fire department Twente we found a list of 34 covariates which could have a high influence on chimney fires, see Table 12 in the Appendix. We selected the covariates based on the government and weather information that we have available from Statistics Netherlands (CBS) and the Royal Netherlands Meteorological Institute (KNMI). In this information we made choices based on the list of covariates from [21] and the extensive experience of the fire department.

Ideally, one would like to have the number of chimneys involved as a covariate, but unfortunately this information is only accessible for us in an indirect way. To cover this problem, we include the building information we have available to estimate the number of chimneys. Because chimneys are mostly included in older houses and/or in stand-alone and town houses, we inherit these covariates in the correlation analysis. The stand-alone and town houses are included in two separate ways: the number of the stand-alone and town houses and the number of residents living in these houses. To make the influence more clear, we also included the same kind of covariate but then for all other types of houses, such as apartments. In terms of weather conditions, we also included two covariates which we will elaborate on: the presence of mist and the wind chill. The presence of mist can cause a decrease in movement inside chimneys and therefore can keep the smoke of a chimney inside the house, which can cause a chimney fire. Secondly the wind chill is also included as a covariate because people probably only light their chimney when they feel the cold. Unfortunately the wind chill itself is not saved by the KNMI so we calculate the value from the other weather conditions by Randall Osczevski en Maurice Bluestein [11]. We separated the mean temperature and the wind chill because the actual temperature can be lower, but when the people do not feel the cold, the chimneys will probably not be lighted.

In Table 12 in the Appendix, first the spatial and then the temporal covariates are listed. Formally, 24 spatial covariates $C_{\sigma,k}, 1 \leq k \leq 24$ are involved in the analysis by partitioning the region of interest A, here Twente, into square boxes with a side length of 500 metres. Let $P_\sigma = \{P_{\sigma,1}, P_{\sigma,2}, \ldots, P_{\sigma,6291}\}$ be the partition of region A where $|P_{\sigma,1}| = |P_{\sigma,2}| = |P_{\sigma,6291}| = 2.5 \cdot 10^5$ squared meter. We used this procedure because the spatial covariates are available per 500 meter box and because it also helps us to have a homogeneous deviation of the region of interest and thus to compare the covariates equally.

For the ten temporal covariates $C_{\tau,l}, 1 \leq l \leq 10$ the definition follows a similar path, but we partitioned the time period of interest T_m. Formally, let $P_\tau = \{P_{\tau,1}, P_{\tau,2}, \ldots, P_{\tau,4380}\}$ be the partition of the period T, where $|P_{\tau,1}| = |P_{\tau,2}| = |P_{\tau,4380}| = 1$ day. As one can see we have removed the leap days to make the dataset manageable, which also means that the predictions do not take the leap days into account and here we will predict the 29th of February as the 28th of February.

4.2 Correlation analysis

Of course, the values of the covariates change over the time and to compare the number of chimney fires against the height of the covariate, we need to summarise the data. For every covariate and

for every year a vector is created which shows for every box the value of the covariate in that year. To compare all chimney fires over the years available these vectors are combined. Consider a spatial covariate $C_{\sigma,k} = C^y_{\sigma,k}(x), 1 \leq k \leq 24$, the value of the covariate in box x and year y. This combined vector has length 6291×12 for the data 2004-2015 and boxes 1-6291, this vector looks as follows:

$$[C^{2004}_{\sigma,k}(1), \cdots, C^{2004}_{\sigma,k}(6291), C^{2005}_{\sigma,k}(1), \cdots, C^{2014}_{\sigma,k}(6291), C^{2015}_{\sigma,k}(1), \cdots, C^{2015}_{\sigma,k}(6291)]^T.$$

These vectors are created for every spatial covariate. For the number of chimney fires, call this $N_x = N^y(x)$ in year y and box x, a similar vector is computed:

$$N_x = [N^{2004}(1), \cdots, N^{2004}(6291), N^{2005}(1), \cdots, N^{2014}(6291), N^{2015}(1), \cdots, N^{2015}(6291)]^T.$$

For the temporal covariates $C_{\tau,l}, 1 \leq l \leq 10$ the analysis follows a similar path. We can now construct a vector based on the days of the year instead of the boxes. The corresponding length is, because we have data on twelve years, 365×12 and the vector for covariate $C_{\tau,l} = C^y_{\tau,l}(t), 1 \leq l \leq 10$, corresponding to year y and day t is computed as:

$$[C^{2004}_{\tau,l}(1), \cdots, C^{2004}_{\tau,l}(365), C^{2005}_{\tau,l}(1), \cdots, C^{2014}_{\tau,l}(365), C^{2015}_{\tau,l}(1), \cdots, C^{2015}_{\tau,l}(365)]^T.$$

These vectors are created for every temporal covariate. For the total number of chimney fires in year y and day t, call this $M_t = M^y(t)$, a similar vector is computed:

$$M_t = [M^{2004}(1), \cdots, M^{2004}(365), M^{2005}(1), \cdots, M^{2014}(365), M^{2015}(1), \cdots, M^{2015}(365)]^T.$$

The correlation coefficient between the number of chimney fires and the covariates can then easily be calculated through these vectors. This coefficient tells us if there is a positive or negative relation between the covariate and the data, where the higher the coefficient in absolute value, the higher the impact on the data. We calculate the correlation using the Pearson's correlation coefficient: Let X and Y be two random variables, then Pearson's correlation coefficient $\rho_{X,Y}$ is defined as follows:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}. \tag{18}$$

Because the covariances and variances between all the covariates considered are unknown, we need to make use of estimations. We will therefore calculate the sample covariance and variance from the dataset and use this in the calculation of the correlation coefficients. This analysis can easily be done with the `stats` package in `R`. With $X = N_x, M_t$ and $Y = C_{\sigma,k}, C_{\tau,l}$ as explained above, Equation 18 reduces in our case to:

$$\rho_{N_x,C_{\sigma,k}} = \frac{\text{cov}(N_x, C_{\sigma,k})}{\sqrt{\text{var}(N_x)}\sqrt{\text{var}(C_{\sigma,k})}}, \qquad \rho_{M_t,C_{\tau,l}} = \frac{\text{cov}(M_t, C_{\tau,l})}{\sqrt{\text{var}(M_t)}\sqrt{\text{var}(C_{\tau,l})}}. \tag{19}$$

The results are shown in Table 1, where $\rho_{X,Y}$ implies the value of one of equations described in Equation 19. According to the analysis, $C_{\tau,2}$ has by far the highest influence and it is a negative one, which indicates that the lower the temperature, the more often chimney fires happen. We will continue to fit the model with a small amount of covariates which have the highest influence and are also having a different influence. The latter will be explained in the following.

The ten covariates with the biggest influence according to our analysis are shown in Table 2. The first thing that comes to mind is that the values of the correlation coefficients lie really close to each other, with the exception of the temperature covariates, $C_{\tau,2}$ and $C_{\tau,3}$. Because of the independence porperties of the inhomogeneous Poisson process we will not include both temperature covariates in the model thus because the mean temperature has a slightly higher correlation, we will include this parameter in the model.

	$\rho_{X,Y}$		$\rho_{X,Y}$		$\rho_{X,Y}$		$\rho_{X,Y}$		$\rho_{X,Y}$
$C_{\sigma,1}$	0.2637	$C_{\sigma,8}$	0.1451	$C_{\sigma,15}$	0.2494	$C_{\sigma,22}$	0.2490	$C_{\tau,5}$	0.0317
$C_{\sigma,2}$	0.1571	$C_{\sigma,9}$	0.1446	$C_{\sigma,16}$	0.2178	$C_{\sigma,23}$	0.0327	$C_{\tau,6}$	-0.2274
$C_{\sigma,3}$	0.0065	$C_{\sigma,10}$	0.1705	$C_{\sigma,17}$	0.2539	$C_{\sigma,24}$	0.1788	$C_{\tau,7}$	-0.0524
$C_{\sigma,4}$	0.2669	$C_{\sigma,11}$	0.2664	$C_{\sigma,18}$	0.2706	$C_{\tau,1}$	0.0632	$C_{\tau,8}$	0.2509
$C_{\sigma,5}$	0.1423	$C_{\sigma,12}$	0.2277	$C_{\sigma,19}$	0.2215	$C_{\tau,2}$	-0.3769	$C_{\tau,9}$	0.1408
$C_{\sigma,6}$	0.1972	$C_{\sigma,13}$	0.2727	$C_{\sigma,20}$	0.2685	$C_{\tau,3}$	-0.3758	$C_{\tau,10}$	-0.0184
$C_{\sigma,7}$	0.1867	$C_{\sigma,14}$	0.0608	$C_{\sigma,21}$	-0.0052	$C_{\tau,4}$	-0.1648		

Table 1: Correlation coefficients between the number of chimney related fires and the covariates as given in Table 12, where a positive and negative value relates to a positive and negative influence respectively.

	$C_{\tau,2}$	$C_{\tau,3}$	$C_{\sigma,13}$	$C_{\sigma,18}$	$C_{\sigma,20}$	$C_{\sigma,4}$	$C_{\sigma,11}$	$C_{\sigma,1}$	$C_{\sigma,17}$	$C_{\tau,8}$
$\rho_{X,Y}$	-0.3769	-0.3758	0.2727	0.2706	0.2685	0.2669	0.2664	0.2637	0.2539	0.2509

Table 2: The ten covariates with the biggest influence following our covariate analysis in Table 1

Half the other covariates included in this table are the number of residents in a certain age group and another two are related to the number of buildings in the area. These covariates are strongly related which is again not desired in an inhomogeneous Poisson process, so we choose again to include the covariate with the highest correlation: the total number of residents per box. Adding another strongly correlated covariate does not make a difference because most of this covariate is already captured by $C_{\sigma,13}$.

When we check again the hypotheses we briefly made in Section 4.1, Table 2 actually is a surprising outcome. Mist does not have a high influence according to this analysis and the age or a specific type of house does not influence the chimney fires more than the number of residents. We think that this is mostly due to the choice of the sub division of Twente in the 500 meter boxes, it may be that an analysis based on neighbourhoods with the same building style will have a higher correlation coefficient than the number of residents. We did not use that procedure in this case because deviation in unequal areas makes the analysis much more complex. Apparently, when we base the analysis on the 500 by 500 meter boxes, the number of residents seems like a good indication of the number of chimneys and therefore we include it in the model.

We continue thus to fit the inhomogeneous Poisson process on $C_{\tau,2}$, the temperature on a day and $C_{\sigma,13}$, the number of residents per box.

4.3 Regression analysis

As concluded in the previous subsection, the two covariates $C_{\tau,2}$ and $C_{\sigma,13}$ are the covariates with the biggest influence and to finish the covariate analysis, a regression analysis is performed. The regression analysis finds the best way to include the two covariates in the intensity function of the inhomogeneous Poisson process. The intensity function corresponding to this process, as explained earlier is not constant and because one spatial and one temporal covariate is considered, the intensity $\lambda(x,t)$ dependent on location x and day t is assumed to be separable:

$$\lambda(x,t) = \lambda(x)\lambda(t). \tag{20}$$

This assumption is needed for the upcoming analysis, but has some implications. Because of separability, it is possible to fit the two λ's separately, but it also causes that the correlation between time and space is not captured. We see the spatial and temporal intensity function as two separate functions and these functions will describe different relations. Together the relations can capture

all chimney fires, but it is more likely that some chimney fires can be captured by both (or none) relations. The goal of the regression analysis is to find this relation, say a function $f(C_{\tau,2})$ and $f(C_{\sigma,13})$ between the covariate and the data. Because we have only one spatial and one temporal covariate, these functions represent the $\lambda(x)$ and $\lambda(t)$ in Equation 20. This analysis finds therefore the best subintensity function per covariate.

Of course, the exact relation described by f is unknown, and therefore an estimation is necessary. Three kinds of functions are therefore fitted to the data and the best fit is used in further calculations.

$$y = \phi_1 + \phi_2 x + \phi_3 x^2 \qquad \phi_i \in \mathbb{R} \text{ for } i \in \{1, .., 3\} \tag{21}$$

$$y = e^{\psi_1 + \psi_2 x + \psi_3 x^2} \qquad \psi_i \in \mathbb{R} \text{ for } i \in \{1, .., 3\} \tag{22}$$

$$y = x^{\theta_1 + \theta_2 x} e^{\theta_3 + \theta_4 x} \qquad \theta_i \in \mathbb{R} \text{ for } i \in \{1, .., 4\} \tag{23}$$

The y represents the number of fire emergency calls where a chimney is involved while x represents one of the two covariates.

The fitting is done in `R` and the resulting plots for both covariates are shown in Figure 8. In the left plot in this figure the y-axis corresponds to the mean number of chimney fires per number of residents. We have chosen to display the mean number of chimney fires because the number of residents is given in groups with steps of five residents (so a box contains 5, 10, 15, etc. residents). To reduce the amount of points, all boxes with the same number of residents are grouped and per group and per resident group the mean number of chimney fires is calculated. This preprocessing makes it easier to fit the best function. A point in the left plot thus indicates the mean number of chimney fires calculated over all boxes and all years with this number of residents, for example (500,0.83) indicates that in the boxes with 500 residents over twelve years on average 0.83 events happened. A point in the right plot shows the mean number of chimney fires per day where the temperature had the corresponding value, for example (-50,1) indicates that on the days where the mean temperature was equal to -5C° in twelve years on average 1 event happened.

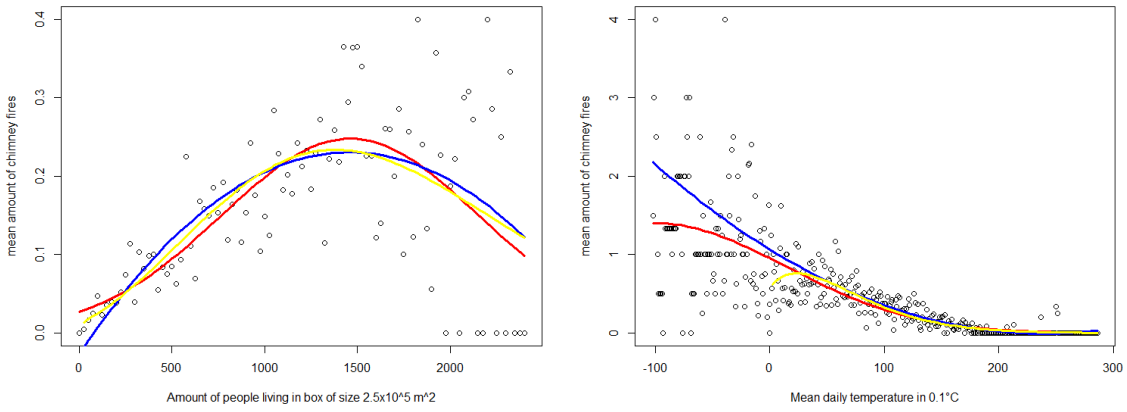


Figure 8: The data of the examined covariates $C_{\sigma,13}$ (left) and $C_{\tau,2}$ (right) plotted against the chimney fire emergency calls with the corresponding regression analysis, involving Equations (21) (blue), (22) (red) and (23) (yellow). The x axis displays the value of the covariate and the y axis the number of fires in the same box/day.

When checking the residuals corresponding to these fits, we found which function fits the data best; for the $C_{\sigma,13}$ data this is Equation (22) and for the $C_{\tau,2}$ data this is Equation (21). The

corresponding coefficients are given below.

$$C_{\sigma,13} = \begin{cases} \psi_1 & = -3.623 \\ \psi_2 & = 3.050 \cdot 10^{-3} \\ \psi_3 & = -1.044 \cdot 10^{-6} \end{cases} \qquad C_{\tau,2} = \begin{cases} \phi_1 & = 1.067 \\ \phi_2 & = -8.983 - \cdot 10^{-3} \\ \phi_3 & = 1.878 \cdot 10^{-5} \end{cases} \quad (24)$$

Concluding, the inhomogeneous Poisson process fitted to the data is defined with intensity function:

$$\lambda(\boldsymbol{x}, t) = \lambda_R(\boldsymbol{x}) \cdot \lambda_T(t)$$

where

$$\lambda_R(\boldsymbol{x}) = e^{\psi_1 + \psi_2 C_{\sigma,13}(\boldsymbol{x}) + \psi_3 C^2_{\sigma,13}(\boldsymbol{x})}$$
$$\lambda_T(t) = \phi_1 + \phi_2 C_{\tau,2}(t) + \phi_3 C^2_{\tau,2}(t)$$

The first intensity function is, as an intensity function should be, always larger than zero and the maximum of the function lies around 1470 residents. As we see in Figure 8 the probability of a chimney fire lowers again when the number of residents exceeds this number. Probably this is because in a box with more residents, more apartments are built, instead of for example free-standing or semi-detached houses. Apartments mostly do not have a chimney, while free-standing and semi-detached houses have, so an area with more apartments also has less probability of chimney fires.

The second fitted intensity function implies an intensity slightly below zero between a mean temperature of 22.1 and 25.6 degrees. For temperatures higher than 25.6 degrees, the intensity increases again. In our data of twelve years, there are 85 days present with a temperature higher than 22.1 degrees, from which 70 days are assigned a negative intensity. This appears to happen very rarely and therefore we continue with these two intensity functions because it is still the best fit for the other temperatures happened in the past.

5 Model 1: Inhomogeneous Poisson process

With the intensity function from the previous section the model is complete and in this section the model is tested. First predictions are made in Section 5.1, after that the model is extended with a new variable concerning the weather tipping point in Section 5.2. The final model definition is given in Section 5.3 and we create confidence intervals and validate the model in Section 5.4.

5.1 Predictions

Because the model is based on the data from 2004 until 2015, we were able to test the model for the year 2016. Because the mean temperature per day and the number of residents per 500 by 500 metre box are known, we can calculate for every day and every box the intensity. All these intensities are saved in a matrix with 365 columns and 6291 rows. The expected number of points in a region A and time period T is equal to:

$$\int_A \int_T \lambda(\boldsymbol{x}, t) dt d\boldsymbol{x}. \tag{25}$$

We chose here to display the expected number of fires per month and compare this with the actual number of chimney fires that occurred in the same month. The chosen A in Equation (25) is then the whole region of interest, Twente, and T is equal to one month. In Table 3 and 4, we show the expected number of events coming from the inhomogeneous Poisson process per month and the actual number of events per month. According to these tables, our predictions are in general

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	**Sum**
Simple	21	19	18	12	6	3	2	2	6	9	11	21	**130**
IPP	18	15	15	10	4	2	2	2	2	9	15	16	**110**
Real	**17**	**13**	**16**	**14**	**6**	**1**	**4**	**1**	**5**	**18**	**14**	**16**	**125**

Table 3: The predicted and real amount of chimney fires for 2016. The extra row 'Simple' is the current practice. These predictions are equal to the mean number of chimney fires in that month over the years 2004 until 2015 and the blue colour shows which model predicts the closest to reality.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	**Sum**
Simple	21	19	18	12	6	3	2	2	6	9	11	21	**130**
IPP	23	14	10	10	4	2	1	2	4	5	12	16	**103**
Real	**22**	**13**	**9**	**17**	**5**	**2**	**3**	**1**	**6**	**11**	**10**	**16**	**115**

Table 4: The predicted and real amount of chimney fires for 2017. The extra row 'Simple' is the current practice. These predictions are equal to the mean number of chimney fires in that month over the years 2004 until 2015 and the blue colour shows which model predicts the closest to reality.

a bit lower than the reality. The 'Simple' model actually predicts the data pretty well, it beats our model for some months, but still, most of the times our model predicts better. Also when our model beats the 'Simple' model, the difference is mostly higher then the other way around.

But our model is not perfect yet, especially in the months April and October, our predictions are definitely off track. These months were actually the tipping points between cold and warm weather. We made this visible in Figure 9. In this figure we see indeed that April is the end of five cold months and the beginning of five warm months while October behaves as a tipping point the other way around. So in October people in general use their chimneys for the first time while

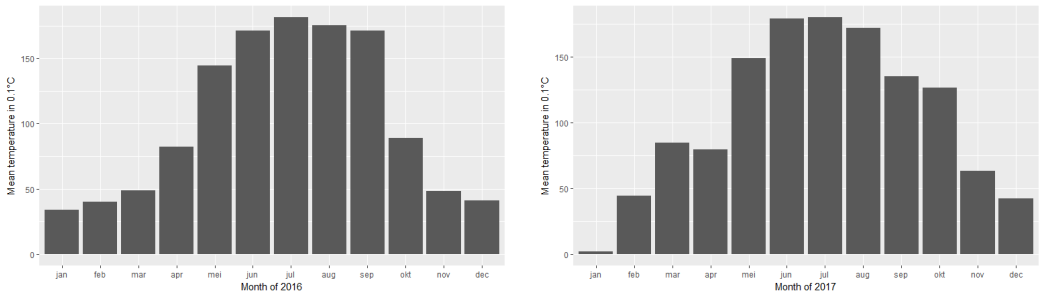


Figure 9: The mean temperature in 0.1°C against the months of 2016 (left) and 2017.

in April people stop using their chimneys. This change of weather conditions is therefore a new covariate which we add to our model.

5.2 Weather tipping point

In the previous subsection, an extra covariate has come up, namely the weather tipping point. We see in Tables 3 and 4 in combination with Figure 9 that especially in October, when people start using there chimneys for the first time, there is a higher change of chimney fires than modelled by temperature and residents so far. In April we see a deviation, also a higher chance of chimney fires seems to appear here when people stop using their chimneys. In this section we investigate if October always contains the tipping point of weather. When this is the case, a simple covariate can be added, namely a boolean variable indicating if the current month is October. The mean temperatures of the years 2004 until 2015 are shown in Figure 10. As one can see in Figure 10, it

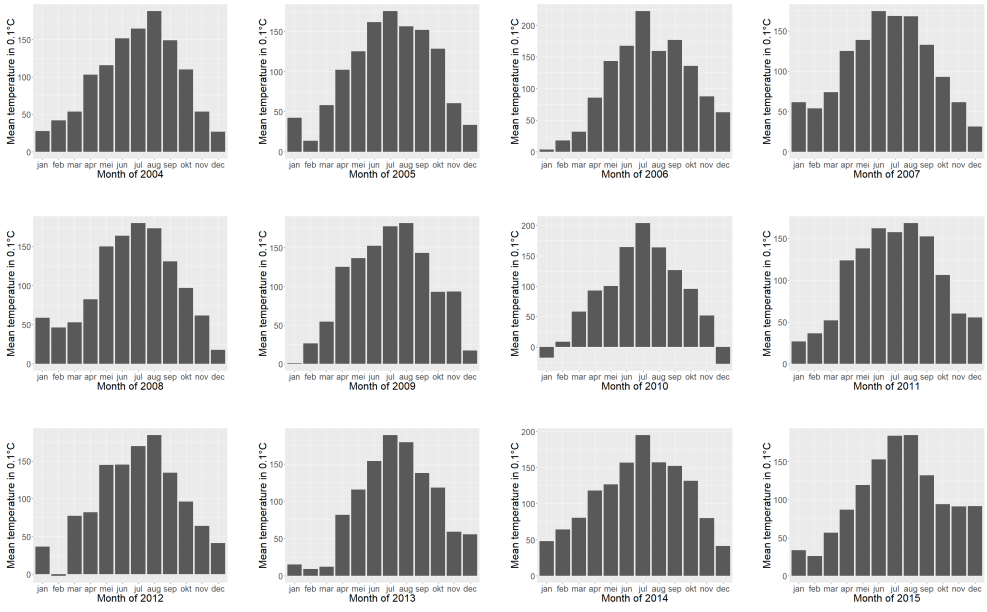


Figure 10: The mean temperatures per month shown in plots per year between 2004 (upper left plot) and 2015 (lower right plot).

was not always the case that the weather tipping point occurs in October. We may therefore also think of another extra covariate, based on the first N days where the temperature drops below T. To find the best value for N and T, we calculate again the correlation coefficients between the data of the years 2004 until 2015 and a boolean variable equalling one for the first N days with

a temperature below T and zero otherwise. We tested $N \in \{2, 3, .., 20\}$ and $T \in \{0, 1, 2, .., 120\}$ where N indicates the number of days and T the temperature in units of 0.1°C. The five combinations with the highest correlation coefficient are shown in Table 5. We directly see that the

T	N	corr
6.8	2	0.144991
6.8	6	0.144830
6.8	3	0.144451
6.8	4	0.144451
6.8	7	0.144294

Table 5: The five highest correlation coefficients between the data and a boolean variable indicating the first N days with a temperature below T in °C.

covariate which checks for temperatures below 6.8°C has the highest correlation coefficients. We choose to include the boolean covariate which indicates the first six days with a temperature below 6.8°C. We chose the first six days instead of two because a large number of events is not predicted in October (for 2016 nine and for 2017 six events). In our data we see that it never happened that more than three chimney fires occurred on the same day (and this happened twice in twelve years) so we need more days with this specific number of events to find a good fit.

We would like to see the impact of three Boolean covariates, one only looking at the month we are in, and one based on the actual temperature:

$$C_{\tau,11} := \begin{cases} 1 & \text{if } \tau \text{ lies in October} \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

$$C_{\tau,12} := \begin{cases} 1 & \text{if } \tau \text{ is one of the first } N \text{ days with a temperature below } T \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

$$C_{\tau,13} := \begin{cases} 1 & \text{if if } \tau \text{ lies in April} \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

We model these covariates by adding them to the already existing intensity function from the first model, $\lambda(\boldsymbol{x}, t)$, in the following way.

$$\lambda_k = \lambda(\boldsymbol{x}, t) \cdot (1 + \gamma_k C_{\tau,k}) \text{ with } k \in \{11, 12, 13\}$$

The γ_k are fitted in the same way as in the previous model and are given the values 0.771589, 1.440355 and 1.154091 for $k = 11$, 12 and 13 respectively. With this new intensity function fitted also to the data from 2004 until 2015, the expected number of events per month for the years 2016 and 2017, is shown in Tables 6 and 7 respectively. Because the predictions using the boolean

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	**Sum**
Simple	21	19	18	12	6	3	2	2	6	9	11	21	**130**
IPP	18	15	15	10	4	2	2	2	2	9	15	16	**110**
IPP-Oct	18	15	15	10	4	2	1	2	2	16	15	16	**116**
IPP-NT	17	15	14	9	4	2	2	2	2	13	15	16	**111**
IPP-Apr	17	16	15	22	4	2	1	2	2	9	15	16	**121**
Real	**17**	**13**	**16**	**14**	**6**	**1**	**4**	**1**	**5**	**18**	**14**	**16**	**125**

Table 6: Predictions of the five models for 2016, where the blue colour shows which IPP model (basis or with extension) predicts the closest to reality

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	**Sum**
Simple	21	19	18	12	6	3	2	2	6	9	11	21	**130**
IPP	23	14	10	10	4	1	1	2	4	5	12	16	**102**
IPP-Oct	23	14	10	10	4	2	2	2	4	9	12	16	**108**
IPP-NT	23	14	10	10	4	2	1	2	4	6	15	16	**107**
IPP-Apr	23	14	9	23	4	2	2	2	4	5	12	16	**120**
Real	**22**	**13**	**9**	**17**	**5**	**2**	**3**	**1**	**6**	**11**	**10**	**16**	**115**

Table 7: Predictions of the five models for 2017, where the blue colour shows which IPP model (basis or with extension) predicts the closest to reality

variable indicating October (Equation (26)) for both years give the best results, we include it in our model. The boolean variable indicating April gives really high results which seems not reliable. Therefore we will continue with only the October boolean and because the April boolean does not make the model better, we will leave it out. Now all obvious failures of the model are investigated, so we end the fitting with the final intensity function including the October boolean covariate.

In Figure 11 the actual occurred events of 2016 and 2017 are compared with the marginal spatial intensity plot made from the fitted intensity function. Most of the events in the left figures occur

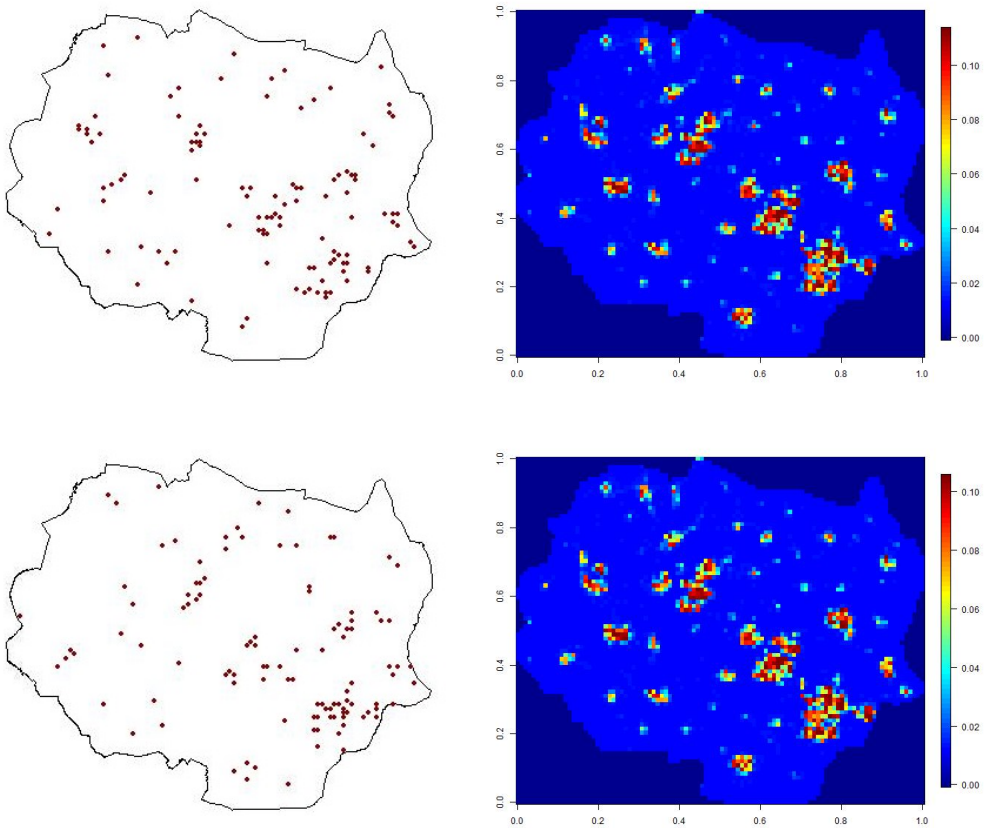


Figure 11: Actual occurrences of chimney fires (left) versus the summed intensities for the whole year according the the model (right) for year 2016 (top) and 2017 (bottom).

in the hotspots of the intensities on the right and a high similarity can be seen between the figures. One must bear in mind that discrepancies will happen due to random fluctuations over the years. To make this more clear, we explain this with an example: In 2017 no event has occurred in the hotspot around the point (0.18,0.65) in the figure, even while this has a higher intensity than the smaller hotspot next to it on the left, (0.1,0.4). We conclude that the hotspots of occurring events are highlighted out well in the intensity figures.

5.3 Final model definition

Our resulting model is an inhomogeneous Poisson process based on three variables, namely the number of residents, the temperature and a boolean variable indicating the month October. The intensity function is presented below:

$$\lambda(\boldsymbol{x}, t) = \lambda_R(\boldsymbol{x}) \cdot \lambda_T(t) \cdot \lambda_O(t) \tag{29}$$

where

$$\lambda_R(\boldsymbol{x}) = e^{\psi_1 + \psi_2 C_{\sigma,13}(\boldsymbol{x}) + \psi_3 C^2_{\sigma,13}(\boldsymbol{x})} \tag{30}$$

$$\lambda_T(t) = \phi_1 + \phi_2 C_{\tau,2}(t) + \phi_3 C^2_{\tau,2}(t) \tag{31}$$

$$\lambda_O(t) = 1 + \gamma_k C_{\tau,11}(t) \tag{32}$$

with the fitted parameters as defined in Equations (24) and (26).

5.4 Validation

The above model is easy to work with, because it has only three covariates whose value is easy to find out. The intensity functions corresponding to these three covariates are also easy to work with because they do not use high order polynomials or exponentials. In this subsection we continue to check if this easy-to-handle model also predicts the chimney fires well, and we do this by checking the confidence intervals and the residuals of the model.

5.4.1 Confidence Intervals

In the preceding section the expected number of events are calculated for the years 2016 and 2017. Because we are still dealing with a Poisson process, the actual predicted number of events will differ. Therefore we compute the confidence intervals of the inhomogeneous Poisson process and check if the events of both years lie in these intervals.

To compute the intervals, subsection 2.2.3 is used. Because the predictions in the previous subsection are given per month, we first produce monthly confidence intervals. For every month and the region of interest Twente the probability described in Equation (2) is computed which resulted in Figure 12.

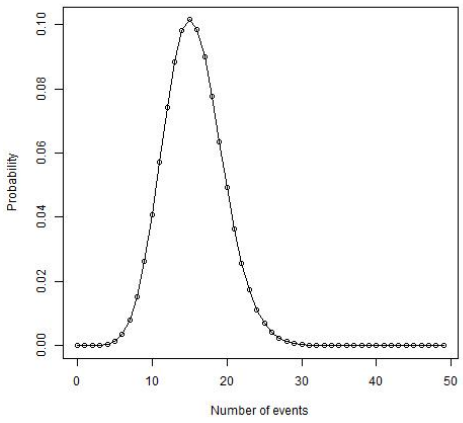


Figure 12: Probability on a number of events in December 2017.

From Equation (3), the actual confidence interval can be computed. Because we are dealing with small probabilities, we choose to compute the 80% confidence interval. The computed intervals for both years including the expected number of events and the actual occurred events, are displayed in Figure 13. The less likely an event is going to happen, the smaller the confidence interval.

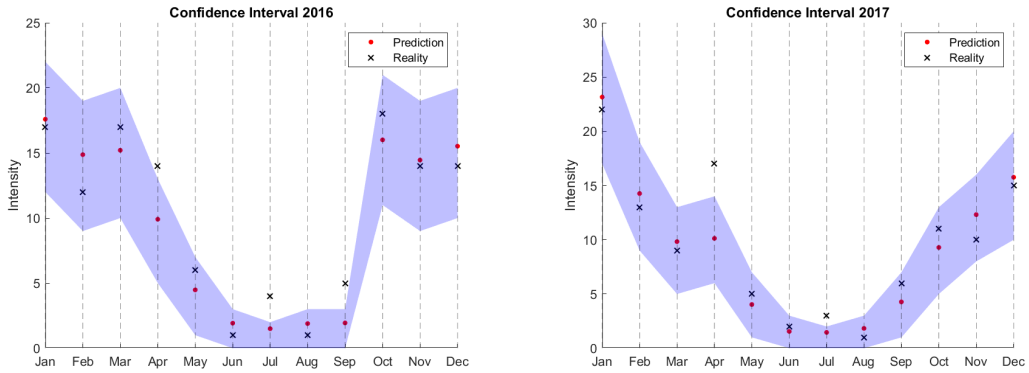


Figure 13: The confidence intervals for the years 2016 and 2017 displayed as a blue area. The red dots indicate the prediction as displayed in Tables 6 and 7 and the crosses are the number of actual occurred events.

For 2016, three events are outside the confidence interval and for 2017 this reduces to two events. This is promising while mostly also reality lies close to the prediction. Sometimes, the confidence interval is quite large, which makes it easier to fit to the actual occurring events. On a monthly basis, we can say that the confidence intervals overall capture the actual occurring events, but the intervals itself are a bit large.

The firemen from Twente want to use this model to predict the number of chimney fires for the coming week. Therefore we follow the same procedure but now focus on a weekly basis. The corresponding confidence intervals are given in Figure 14. Still, most of the weeks during winter, the actual occurred number of events are included in the confidence intervals. For the first and last twelve weeks in both years, only two crosses lie outside the confidence interval. These two crosses happen both in 2016 in week 5 and week 47. During summer, we see that the actual number is mostly directly at the corner or outside the interval. Because in the summer we are mostly dealing with events happening coincidentally, this result is expected. From these figures

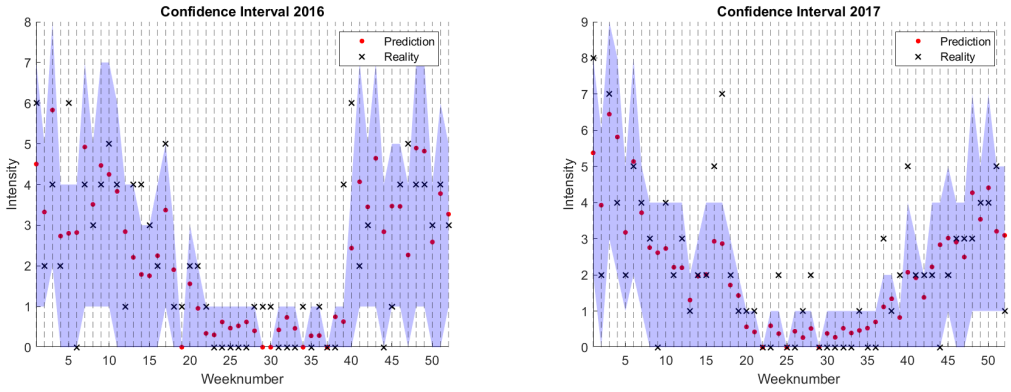


Figure 14: The confidence week intervals for the years 2016 and 2017 displayed as a blue area. The red dots indicate the prediction according to the inhomogeneous Poisson process and the crosses are the number of actual occurred events.

we see clearly that the predictions on a weekly basis is good in the winter and becomes more and more guesswork in Summer.

5.4.2 Residuals

At first, for the covariates $C_{\sigma,13}$ and $C_{\tau,2}$ we found the corresponding residual plot. These plots are given in Figure 15. The left plot in Figure 15 corresponding to the number of residents ($C_{\sigma,13}$)

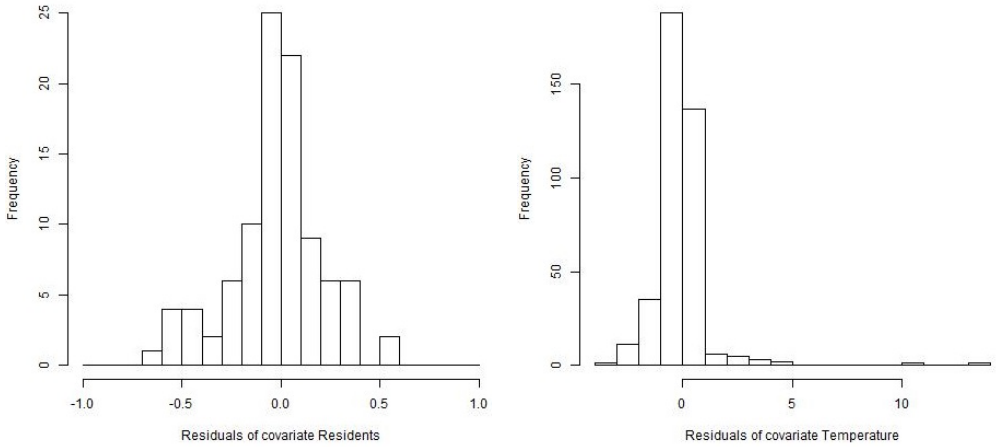


Figure 15: The residual plot of the fitted intensity function of (top to bottom, left to right) $C_{\sigma,13}$ and $C_{\tau,2}$ against the emergency calls of chimney fires.

shows us a histogram which is centred around zero and has highest absolute values around 0.6-0.7. That are small residuals so we can conclude that function $\lambda_R(\boldsymbol{x})$ fits well to the data. The right plot show that the residuals of $C_{\tau,2}$ are also centred around zero, but also contains higher differences between the data and the fitted function. We did not include the residuals of $C_{\tau,11}$ because of its boolean character.

But of course, we consider a model based the three variables together. Therefore, we also wanted to find the residuals for the total model and we calculated the *raw residuals* [3]. For the estimated marginal temporal intensity function $\hat{\lambda}(t)$ with fitted parameters summarised in $\hat{\theta}$ and N_t the

number of occurrences in time period $[0, t]$ and for the whole region Twente, the *raw residuals* for $\hat{\lambda}(t) = \lambda_{\hat{\theta}}$ are given by:

$$R(t) = N_t - \int_0^t \hat{\lambda}(s)ds.$$

Because we deal with a spatio-temporal intensity function $\lambda(\boldsymbol{x}, t)$, the corresponding residuals include not only time but space as well. First we define the residuals in time as the residuals for the whole region of interest Twente, W_S, for the time period $[0, t_0]$:

$$R(t_0) = N_{t_0} - \int_0^{t_0} \int_{W_S} \hat{\lambda}(\boldsymbol{x}, t)d\boldsymbol{x}dt \tag{33}$$

The corresponding plot for years 2016 and 2017 are given in Figure 16. In this plot the residuals

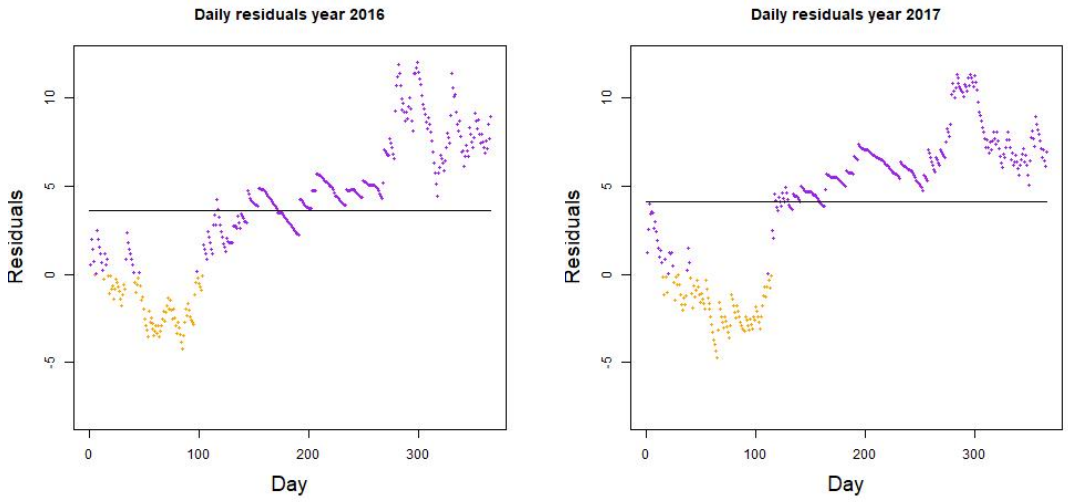


Figure 16: Residual time plot corresponding to Equation (33), where the black solid line indicates the mean of the residuals.

with a negative value are given the colour orange while residuals with a positive value are displayed in purple. Here we see that in general we predict less events than actually occur, while most dots lie above zero. Also, the same trend can be seen in both years: In the first third of the year there is a period with too many predicted events and in the last third a period occurs where too few events were predicted, in terms of intensity of course. The mean of the residuals of both years lies around four events.

To check the spatial marginal intensity, a similar procedure is followed. In the marginal temporal residuals, $[0, t]$ was interpreted as the time period over which we could integrate. A similar approach for locations is not intuitive so the spatial residuals are calculated for every box separately (so not summed up). We therefore change N_x into $\bar{N}_x$ which is a result from a Gaussian blur applied to the actual events. A blurring kernel with standard deviation σ is used. The standard deviation is chosen visually to be $\sigma = 2.14$, the blurred image still shows hotspots but is a little bit smoother. This smoothing helps us to extract better results. In formula, the marginal spatial residual for box $\boldsymbol{x}$ is defined as:

$$R(\boldsymbol{x}) = \bar{N}_x - \int_0^{365} \hat{\lambda}(\boldsymbol{x}, t)dt.$$

The three figures resulting from this procedure are displayed in Figure 17. The computed intensity

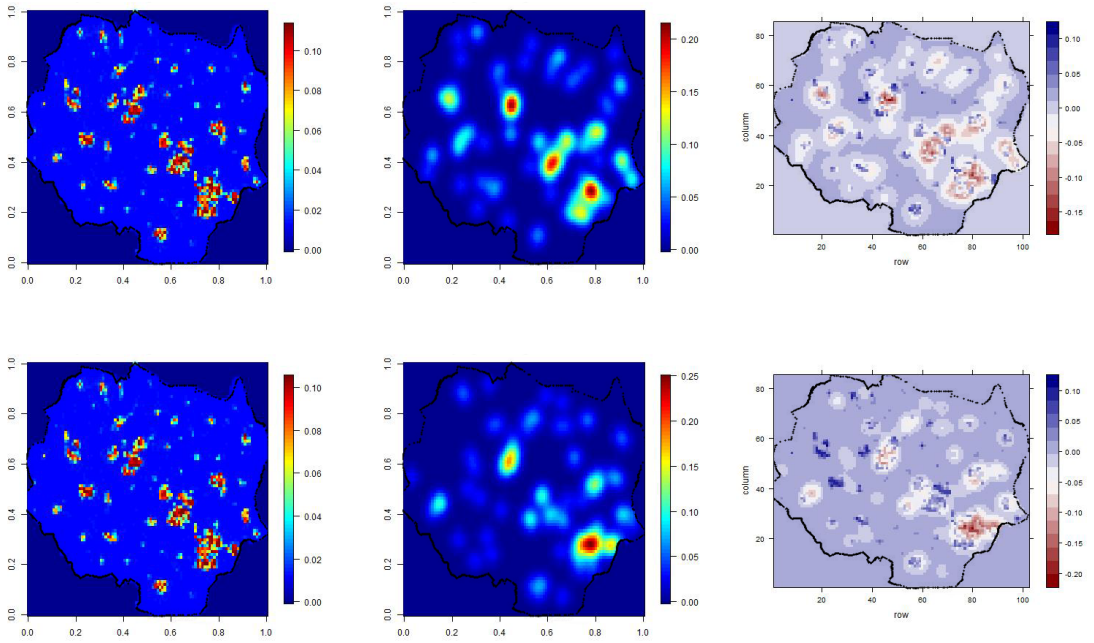


Figure 17: From left to right: The computed intensity plot according to the fitted inhomogeneous Poisson process, the Gaussian kernel applied to the actual occurred events and the residual plot. The first and second row corresponds to the data of year 2016 and 2017 respectively.

plots are very similar. Because here the integrals of the fitted intensities over a full year are taken and in both years a similar weather behaviour has occurred this is completely logical. The middle image shows us the Gaussian blurred events where clearly the three cities are recognized. The difference between the two then results in the right image. There are a lot of areas where the computed intensity comes close to the real image. For 2016, the residual image has 'high' values $R(\boldsymbol{x})$ in the bigger cities but also in some towns around. In some parts of the cities the computed intensity is higher (so blue) and other parts close by this intensity is lower (red) than reality. For 2017 the red parts are more centered around Enschede (right bottom city) and the blue parts around Hengelo (middle city) which is a very logical result when checking the Gaussian blurred 'real' image.

In the residual plots of both years, no structural differences are visible, but we mostly see lower values which indicates that less points are expected than occurred which also was a conclusion from the time residuals. For both years we calculated the sum of the residuals which is -6.20 and -2.91 for 2016 and 2017 respectively. The residuals for 2016 are thus quite high but both values are due to the fact that the inhomogeneous Poisson process predicts less events than actually happening.

Concluding, the inhomogeneous Poisson process is easy to work with, but however, it predicts too few events. Also as we have seen in the distance analysis, the interaction between events is missing. These extra events can be modelled with the random field in the Log Gaussian Cox process which will be done in the next section.

6 Fitting of the Log Gaussian Cox process

In the last section we derived an inhomogeneous Poisson process which most of the time provides an underestimation of the number of chimney fires. Therefore, and to include the spatially dependent noise, we will extend the model to a Log Gaussian Cox process. As explained earlier, this process adds a noise to the process which covers all small influences which we cannot predict well, such as human behaviour. In this section we try to add the noise in the form of this new random variable.

As discussed in Section 2.3.2, we want to add a random field which transforms the model into a Log Gaussian Cox process. We defined the process in general in Section 2, hence we only need to specify an expression for the covariance function $\rho(u, v)$ and a value for the mean μ. When the covariance function and the mean are defined, the model is complete and can be added to the inhomogeneous Poisson model we derived so far.

Since the covariance function is free to pick, we chose the exponential covariance function. This was also done in earlier papers with similar problems [5], [8], [9] and [17]. Based on these papers and the fact that we consider a spatio-temporal model, we choose the following covariance function:

$$\rho(r, v) = \sigma^2 \exp\left[-\frac{r}{\beta_S}\right] \exp\left[-\frac{v}{\beta_T}\right] \tag{34}$$

Here we consider two events $x_i = (u_i, t_i)$ and $x_j = (u_j, t_j)$, where $u_i, u_j \in W_S$ and $t_i, t_j \in W_T$. The $r = ||u_i - u_j||$ corresponds to the distance difference in meters while $v = |t_i - t_j|$ corresponds to the time difference in days between the events.

Equation (34) has three unknown parameters, β_S, β_T and σ^2. Together with the mean μ, these parameters need to be fitted to the data. We call β_S and β_T the spatial and temporal scale, while σ^2 corresponds to the variance as introduced in Section 2.3.1. These parameters are fitted using a minimum contrast method in Section 6.1 and μ is chosen based on the number of events we miss in the model which will be explained in Section 6.2. The actual calculation and results will be discussed in Section 6.3.

6.1 Minimum contrast method

A minimum contrast method is a method to fit unknown parameters by minimising the difference between a theoretical expression for a function and an estimation of the same function. We know the theoretical expression for this function but we do not now the exact values of different parameters. This theoretical expression can be estimated based on the data and we want to find the corresponding parameters for which the theoretical function is closest to the estimated version. Different functions can be chosen, provided that they can be estimated, for example the K-function [9] and the covariance or pair correlation function [17], [5] and [8] (recall from Equation (12) the relation between these functions). We choose to base the method on the pair correlation function, because this was the easiest to implement in the programming language `R`.

The goal is to find the three parameters σ^2, β_S and β_T which minimize the difference between the estimated value of the correlation function and the actual value based on the two unknown parameters. We thus want to minimize the following criterion

$$\int_{\delta}^{v_0} \int_{\epsilon}^{r_0} [\hat{g}(r, vt) - g(r, v)]^2 dr dt \tag{35}$$

To estimate $\hat{g}(r, v)$, we define a new formula [5] and we need to introduce some parameters. Say x_i is again an event with corresponding time $t_i \in W_T$ and location $u_i \in W_S$. In this research

$W_T = [0, 365]$ and W_S is the spatial polygon describing Twente. With the use of the estimated inhomogeneous λ from Section 3.3.1, based on data giving the locations of events $x_i = (u_i, t_i)$, $i = 1..., n$, the pair correlation function can be estimated:

$$\hat{g}(r, v) = \frac{1}{|W_S \times W_T|} \sum_i \sum_{j \neq i} \frac{1}{w_{ij}^{(s)} w_{ij}^{(t)}} \frac{k_s(r - ||u_i - u_j||) k_t(v - |t_i - t_j|)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j)} \tag{36}$$

where k_s and k_t are two kernels corresponding to the spatial and temporal part respectively and following the reasoning in [13], for both functions box kernels are used. The spatial edge correction $w_{ij}^{(s)}$ is chosen to be Ripley's edge correction as explained in [5] while the temporal edge correction $w_{ij}^{(t)}$ follows the following formula:

$$w_{ij}^{(t)} := \begin{cases} 1 & \text{if both ends of the interval of length } 2|t_i - t_j| \text{ centred at } t_i \text{ lie within } W_T \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

The 'real' pair correlation function resulting from Equation (12) and Equation (34) is then:

$$g(r, v) = \exp\left[\sigma^2 \exp\left[-\frac{r}{\beta_S}\right] \exp\left[-\frac{v}{\beta_T}\right]\right]. \tag{38}$$

We thus want to find estimated values $\hat{\sigma^2}$, $\hat{\beta_S}$ and $\hat{\beta_T}$ which minimize the criterion as specified in Equation (35) with $\hat{g}(r, v)$ coming from Equation (36) and $g(r, v)$ coming from Equation (38) The parameters $\epsilon \geq 0$ and $\delta \geq 0$ are user specified and are included to avoid numerical instabilities around zero. The r_0 (v_0) is the upper limit of the values of r (v) which are of interest [5].

6.2 Mean

The Log Gaussian Cox process we are defining here is almost complete, the missing parameter is the mean μ. To fit this parameter, we use again the fitted intensity function from the inhomogeneous Poisson process as summarised in Equation (29) and the fitted intensity function in the distance analysis from Figure 4. The mean μ can namely be chosen to set an expected value of predicted points, so μ will be chosen to fill in the amount of points we are missing in the inhomogeneous Poisson model.

The fitted intensity function from Figure 4 defines the spatial intensity function based on twelve years of data, we call this function $\bar{\lambda}(\boldsymbol{x})$. The integral of $\bar{\lambda}$ over space is equal to the expected number of events in the twelve years we have data on. We choose that the Log Gaussian Cox process should predict the mean number of events per year coming from this intensity function. The expected number of events coming from $\bar{\lambda}(\boldsymbol{x})$ is

$$\int_A \bar{\lambda}(\boldsymbol{x}) d\boldsymbol{x} \tag{39}$$

with A the region of interest. The expected number of events we predict with the inhomogeneous Poisson model defined in Section 5 is equal to:

$$\int_A \int_T \lambda(\boldsymbol{x}, t) dt d\boldsymbol{x} = \int_A \int_T \lambda_R(\boldsymbol{x}) \cdot \lambda_T(t) \cdot \lambda_O(t) dt d\boldsymbol{x} \tag{40}$$

with A the region of interest and T the time period of interest and $\lambda_R(\boldsymbol{x})$, $\lambda_T(t)$ and $\lambda_O(t)$ as defined in Equations (30), (31) and (32).

The difference between the values of Equations (39) and (40) is the under or overshoot number of points. This difference is called R_m and is equal to the mean of the total residuals. R_m is

calculated on a yearly basis so we chose T to be one year. With the estimation $\bar{\lambda}(\boldsymbol{x})$ is based on twelve years of data and T is one year, R_m can be calculated as follows:

$$R_m = \frac{1}{12}\int_A \bar{\lambda}(\boldsymbol{x})d\boldsymbol{x} - \int_A\int_T \lambda_R(\boldsymbol{x})\cdot\lambda_T(t)\cdot\lambda_O(t)dtd\boldsymbol{x}. \tag{41}$$

When we derived σ^2 from the minimum contrast method, then μ can easily derived from the following which combines Equation (7) and Equation (41).

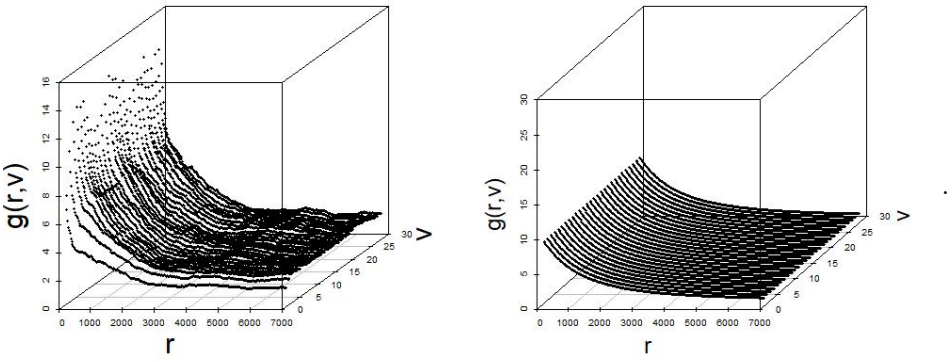
$$R_m = \mathbb{E}[X] = \exp[\mu + \frac{1}{2}\sigma^2]. \tag{42}$$

6.3 Results

The previously described method is applied to the data and in this subsection we display the fitted parameters. The Log Gaussian Cox model is then described and new simulations are made.

6.3.1 Minimum contrast method: σ^2, β_S and β_T

When we applied the above described method to our data, we found the estimated pair correlation function first. This function is due to capacity reasons estimated for $r \in \{25, 50, 75, .., 10975, 11000\}$ and $v \in \{1, .., 28\}$. The first range has been extracted from the estimation of the K-function in Section 3. In this estimation, this range has also been used and from these results, no covariance hence correlation is deducted after 11000 meters. The range for v is chosen to cover four weeks because, in consultation with the fire department, we assume that two events are uncorrelated when these happen four weeks apart from each other. With the help of the R-package `stpp` the pair correlation function $\hat{g}(r, v)$ is estimated, which is displayed in Figure 18a. In the data we



(a) $\hat{g}(r, v)$ (b) $g(r, v)$

Figure 18: The estimated pair correlation function $\hat{g}(r, v)$ for data in Figure 1b for the range $r \in \{25, 50, 75, .., 10975, 11000\}$ and $v \in \{1, .., 28\}$ (left) and the pair correlation function $g(r, v)$ with the fitted parameters $\sigma^2 = 2.336$, $\beta_S = 3001$ and $\beta_T = 2200$.

found that the estimated values for the pair correlation function decrease when r increases, which is logical. When r exceeds the 7 kilometres, the estimated values start to fluctuate. Because of this we think that for larger r the estimation is no longer stable and in the minimum contrast method we will therefore only use the estimated values corresponding to $r \leq 7000$. This change will prevent that the unstable part of the estimated pair correlation function has a lot of influence on the parameter fitting. In the above minimization we chose thus $r_0 = 7000$. Because the estimation of the pair correlation function is equal to infinity for $r = 0$ and other large values for

small r we use the third r value for which an estimation exists, $\epsilon = 50$. The estimation in time seems stable so we use all values of v, which results in $\delta = 1$ and $v_0 = 28$. The results of the minimization equation as defined in Equation (35) become $\sigma^2 = 2.336$, $\beta_S = 3001$ and $\beta_T = 2200$. The corresponding pair correlation function is plotted in Figure 18b.

For different values of r or v the fit can be investigated further. Three plots which display the estimated and fitted pair correlation function for a fixed v are given in Figure 19. All 28 plots can be found in Figures 28 and 29 in the Appendix. As also can be seen in the Appendix, most

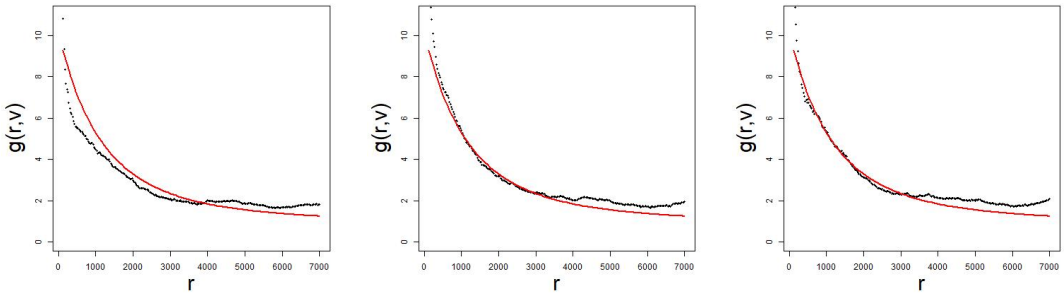


Figure 19: The pair correlation function as in Equation (38) with the fitted parameter values (solid red line) against the estimate of the pair correlation correlation function as in Equation (36) (black dots) for the time differences 2, 14 and 24 days respectively.

plots fit well. The first few plots, so the plots concerning a small time difference, show the highest deviation especially for small distance differences. Probably this is due to the lack of data for small time and distance differences between events. We do not have a lot of data for events which occur only a few days and a few kilometres apart. Of course Twente has hotspots where a lot of chimney fires occur, but within a year around 10-20 events happen in a city. The data includes therefore a few to none events which happen within 3 days and 2 kilometres apart. The estimation is thus based on only a small amount of data and therefore not reliable.

6.3.2 Mean: μ

Now we fitted the σ^2 corresponding to the model, Equation (42) can be filled in. The total number of chimney fires we expect to happen each year is easily derived from Figure 4 and Equation (39), which results in an average amount of 121 chimney fires per year.

To calculate Equation (40), all information about residents, temperature and the month October is used. Because the three intensities in this equation are dependent on the number of residents per box, the temperature per day and the month this day is in, we combine this information and reduce the integral to a sum of all intensities corresponding to a mean value of all variables per day and per box. So first we create three vectors: one of length 6291 (the amount of boxes) and two of length 365 (the amount of days per year).

The first vector gives us for every box the mean number of residents over the years 2004 until 2015. The second vector gives us for every day the mean temperature calculated from the same years and the last vector is just the boolean vector indicating which day is an October day and which day is not. We then calculate all intensities (a number of 6291×365 in total) and sum them.

The expected amount of points predicted with only the inhomogeneous Poisson process is 110, so we miss $121 - 110 = 11$ chimney fires. These will be included in the random field defined in the

previous section, and from Equation (42) we deduce

$$\mu = \ln[R_m] - \frac{1}{2}\sigma^2 = \ln[11] - \frac{1}{2} \cdot 2.336 \approx 1.230 \tag{43}$$

6.3.3 Simulations

Now we know the missing parameters, the Log Gaussian Cox model is complete and simulations can be made. The same procedure is followed as in Section 2.3.3 and the resulting simulations are shown in Figure 20. To show the variability in simulations we display here twelve simulations with the same parameters. The randomness is clearly visible in these plots which indicates a good

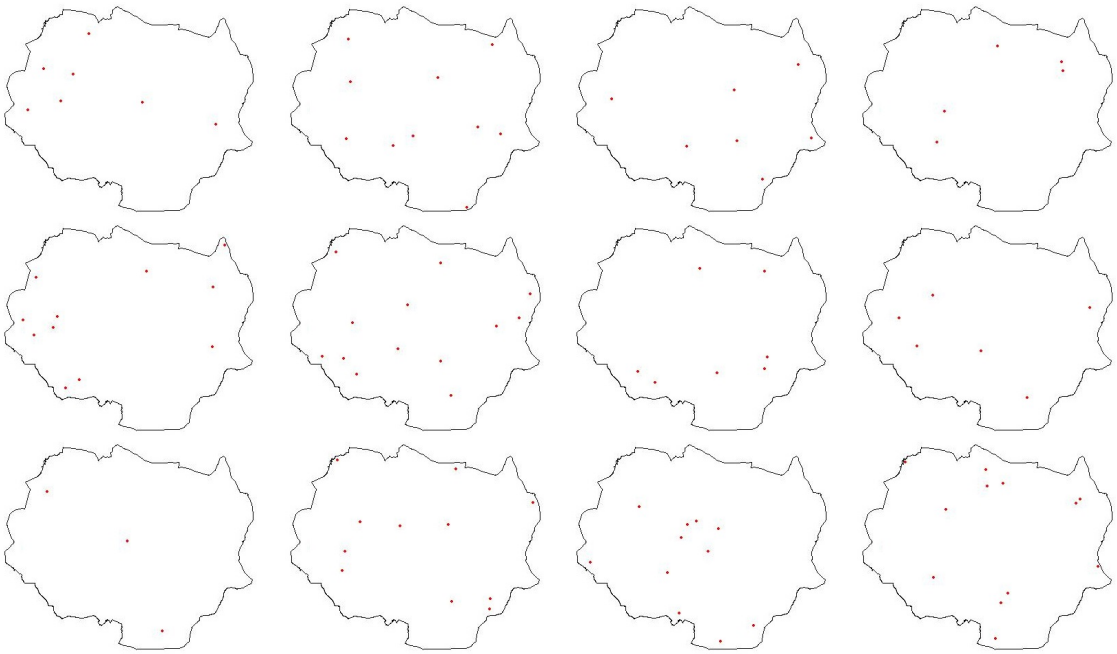


Figure 20: Realizations of the LGCP with an exponential covariance function with $\mu = 1.230$, $\sigma^2 = 2.336$, $\beta_S = 3001$ and $\beta_T = 2200$.

Gaussian field. This model is in the next section added to the inhomogeneous Poisson process to create simulations of our final model.

7 Model 2: Log Gaussian Cox process

The Log Gaussian Cox process is fitted to the data and the resulting model is complete. With this model we can also check the predictions and validate the model which we do in this section.

7.1 Log Gaussian Cox model definition

The two processes can be combined into a new prediction model. Following the same correlations as in Section 2.3.1, the resulting combined model is defined by its intensity function and the definitions corresponding to the random field W. For a complete description of the model we need to define the total intensity function $\lambda(\boldsymbol{x}, t)$, the correlation function $g(r, t)$ and the mean μ of the random field:

$$\lambda^{LGCP}(\boldsymbol{x}, t) = \lambda(\boldsymbol{x}, t) \exp[W_{\boldsymbol{x},t}] \tag{44}$$

where

$$\lambda(\boldsymbol{x}, t) = \lambda_R(\boldsymbol{x}) \cdot \lambda_T(t) \cdot \lambda_O(t) \tag{45}$$

with $\lambda_R(\boldsymbol{x})$, $\lambda_T(t)$ and $\lambda_O(t)$ defined by Equations (30), (31) and (32). The random field W is defined by its correlation function $g(r, v)$ with r and v the spatial and temporal distance and mean μ:

$$g(r, v) = \exp\left[2.336 \cdot \exp\left[-\frac{r}{3001}\right] \exp\left[-\frac{v}{2200}\right]\right] \tag{46}$$

$$\mu = 1.230. \tag{47}$$

7.2 Predictions

With this model, the years 2016 and 2017 can again be predicted. The same procedure as in Section 5 is used so per model the expected number of events per month is calculated. Results are shown in Tables 8 and 9.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	**Total**
PoisP	18	15	15	10	4	2	1	2	2	16	15	16	**116**
LGCP	18	15	16	11	5	2	2	3	2	17	15	17	**123**
Real	17	13	16	14	6	1	4	1	5	18	14	16	**125**

Table 8: Predictions of the inhomogeneous Poisson process against the prediction corresponding to the Log Gaussian Cox process for 2016.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	**Total**
IPP	23	14	10	10	4	2	1	2	4	9	12	16	**107**
LGCP	24	15	11	11	5	2	2	3	5	10	13	17	**118**
Real	22	13	9	17	5	2	3	1	6	11	10	16	**115**

Table 9: Predictions of the inhomogeneous Poisson process against the prediction corresponding to the Log Gaussian Cox process for 2017.

The tables show that the missing number of events in the inhomogeneous Poisson process are present in the combined model. The total number of events comes closer to reality.

Also the intensities can be checked again for this model. The results are shown in Figure 21.

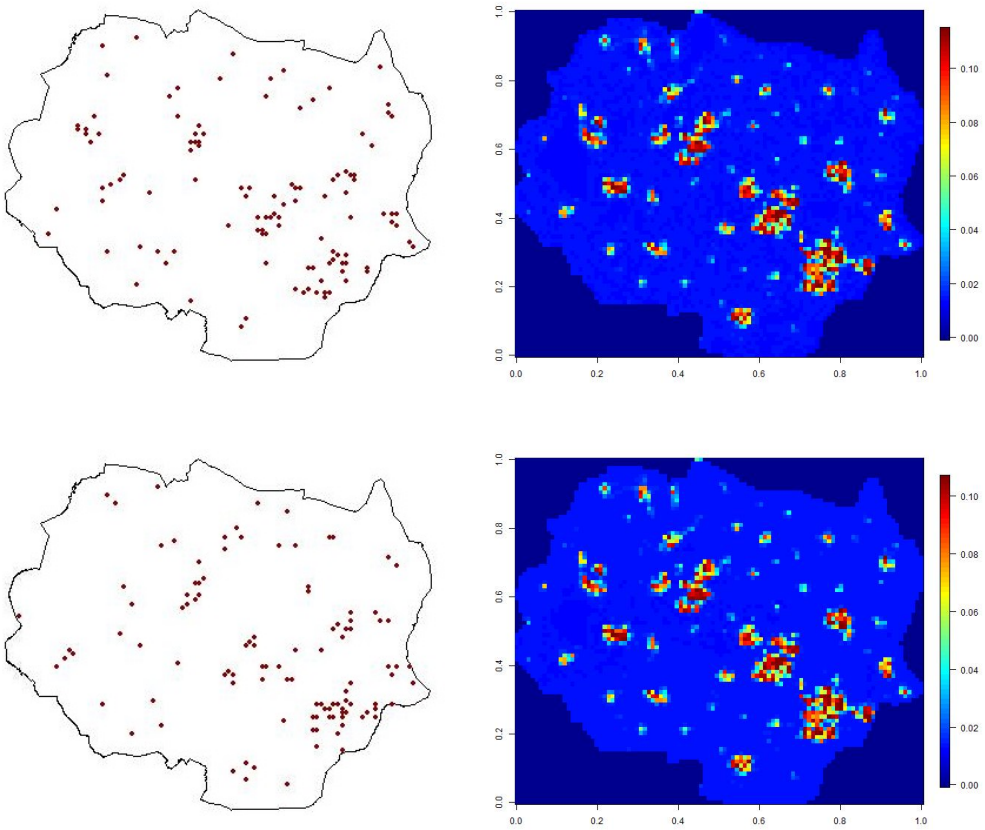


Figure 21: Actual occurrences of chimney fires (left) vs the summed intensities for the whole year according the Log Gaussian Cox model (right) for year 2016 (top) and 2017 (bottom).

The intensities do not differ much with the inhomogeneous Poisson model which is due to the small number of events we predict with only the random field W_r. This small number of events only adds small intensities spread out over the whole of Twente which returns an intensity field with really small differences. Still the hotspots are clearly visible and therefore the intensity plot looks very good.

7.3 Validation

In this subsection, the Log Gaussian Cox model we defined is validated. Just as we did with the inhomogeneous Poisson process, we first elaborate on the confidence intervals of this model in Section 7.3.1 and then continue to the residuals in Section 7.3.2.

7.3.1 Confidence Intervals

For the Log Gaussian Cox process, the confidence intervals are also a good information source about the model. In this paragraph we calculate confidence intervals for Log Gaussian Cox processes. There seem to be no specific papers on this subject, so we combine here different methods to get the best result. At first we will derive the monthly confidence intervals and later continue to the weekly confidence intervals. To compute now a relevant confidence interval, we want to check how many simulated events occur in the same week. When this is known, a confidence interval for the number of events is made and added to the confidence interval corresponding to the inhomogeneous Poisson process. So for example week 1 has a confidence interval for the Poisson process of (3,8) and the in this section described method results in an interval of (0,3), then the confidence interval for week 1 for the combined model is (3,11).

So first the confidence interval of the Log Gaussian Cox process is calculated. In the above, four parameters are fitted to the data, μ, σ^2, β_S and β_T. To compute the confidence intervals, the sensitiveness of these parameters is examined. This examination is based on *leave k out cross validation* [18], or *k-fold cross validation*. This method divides the data into k subsets and fits the model parameters on the data minus one of these subsets. It is possible to choose these k subsets differently to check the difference in the fitting of the parameters. When we know the range of these parameters, new simulations can be made and based on a lot of these simulations, also the range of the model can be computed. In the following this method and the choices made are described in more detail.

First the range of the four parameters are examined. For the mean μ, we fix the number of events coming from the inhomogeneous Poisson process and focus on the 'left over' events predicted by the Log Gaussian Cox process. We chose to remove twelve random months of data and fit μ on the rest of the data. The twelve months are chosen because then still there is enough data to fit the parameters on, but also a significant amount of data is removed (comparable with the loss of a year of data). Say thus D is the set of twelve years (2004-2015) of occurring chimney fires, and from this set twelve random months are chosen which we call M_{12}. We fit μ on the set $D \backslash M_{12}$ by finding the new mean of occurring intervals per year, $\bar{R}_m$:

$$\bar{R}_m = \frac{|D \backslash M_{12}|}{11}$$

where 11 is the number of years left. We performed 6000 simulations of computing the mean and find the following distribution for the lack of points as displayed in Figure 22. In these figures, a normal distribution can be recognized. All the 6000 simulation values of μ need to be combined with the fitting values for the other three parameters to create a confidence interval for the Log Gaussian Cox process.

The other three parameters are fitted together. Here also twelve random months are removed from the data and the minimum contrast method is repeated for $D \backslash M_{12}$. Because of the long

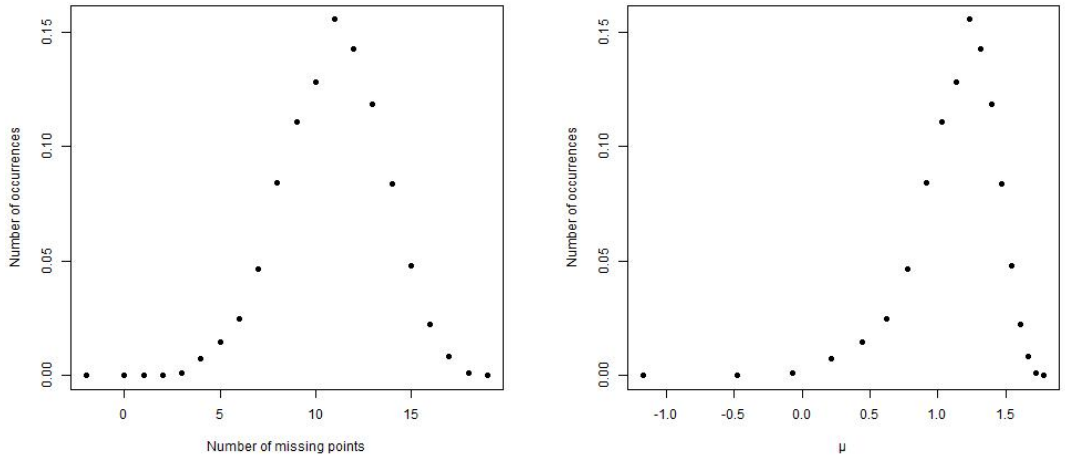


Figure 22: On the left the number of missing points, or $\bar{R}_m$, against the probability of occurring in 6000 simulations. On the right $\bar{R}_m$ is transformed into μ with the help of Equation (41).

computation time of this method, estimating the pair correlation function takes a lot of time, so 6000 simulations are unfortunately not possible. Till now, 13 simulations are made and their results are given in Table 10. Because we only can make this low number of simulations, the exact

Variance σ^2	Spatial Scale β_S	Time Scale β_T
2.322	**2995**	**1.016+07**
2.297	2847	0.591+07
2.337	2934	1.393+07
2.315	2812	0.641+07
2.282	2680	3.920+07
2.230	2908	3.733+07
2.259	2700	3.748+07
2.314	2940	0.232+07
2.306	2964	3.098+07
2.254	2595	1.149+07
2.321	2803	1.852+07
2.317	2925	3.762+07

Table 10: Results of the minimum contrast method with twelve random months of data removed from the actual data. The second row displays the value of the parameters which is used now in the model.

distributions are not available. The confidence intervals coming from this analysis are therefore 'empirical confidence intervals' and not strict ones. We thus approach the confidence intervals by doing this analysis and not derive the actual confidence intervals.

For each combination of a μ value and the three other parameter values, 1000 Log Gaussian Cox simulations are made. For every simulation the number of events per week is counted and saved. After these simulations we thus have for example Table 11. From here the confidence interval of number of occurrences per week is calculated. Directly from the table we conclude that the 80% confidence interval is just (0,0). When repeating the process for all combinations of the four parameter values, we always end up with a confidence interval of (0,0). This is not strange,

Number of occurences	0	1	2	3	4
Frequency	42043	8986	900	69	2
Probability	0.8085	0.1728	0.0173	0.0013	3.85e-05

Table 11: After 1000 simulations of the Log Gaussian Cox process the corresponding frequency of the number of events per week. The probability is calculated as the frequency value divided by the sum of all frequencies.

because the simulations mostly predict around 11 points, and with a number of 52 weeks, there are a lot of weeks with no event at all. The confidence interval plot for the Log Gaussian Cox process is displayed in Figure 23. The difference between this plot and Figure 14 are the red dots. The

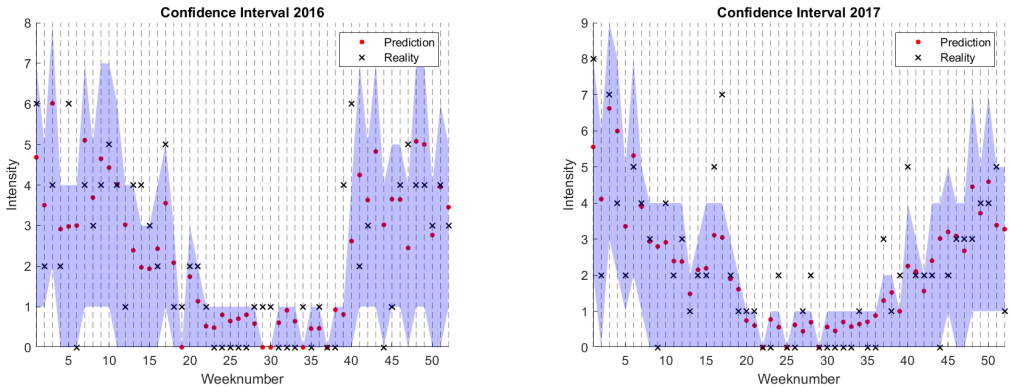


Figure 23: The confidence week intervals for the years 2016 and 2017 displayed as a blue area. The red dots indicate the prediction according to the Log Gaussian Cox process and the crosses are the number of actual occurred events.

predictions per week are different for the Log Gaussian Cox process, but the confidence intervals are the same as explained previously. We see especially for year 2017 that the new predictions are coming a bit closer to the actual occurrences, but in terms of confidence intervals the conclusion is the same as in Section 5.4.1. Recall that these confidence intervals are not strict ones and that we approach the reality because of the low number of simulations. The small difference between the confidence plots is then because of the low number of extra events the Log Gaussian Cox process predicts. The difference between the models is better visible in terms of residuals in the next subsection.

7.3.2 Residuals

To check also the residuals for this new model, the same way of calculation as in Section 5.4 is used. The marginal temporal results are shown in Figure 24. In comparison to the residuals in Section 5.4, this model gives better results. The residuals are closer to zero, the peaks are lower and as a consequence also the mean is closer to zero. Based on these plots, the new model gives better results in terms of residuals.

For this combination model also the marginal spatial residuals are calculated and displayed in the same way as in Section 5.4, see Figure 25. The computed intensity plots on the left side look more smooth than those in the previous validation section. The Gaussian field thus has an extra smoothing property so that the two images look more alike. The highest intensity has a much higher value in the real plot in comparison with the computed intensity plot. The residuals in this case are very similar because of the really small adjustment in intensities from the Gaussian field. Recall that this random field adds around 10-12 points to the plot spread out over 6291

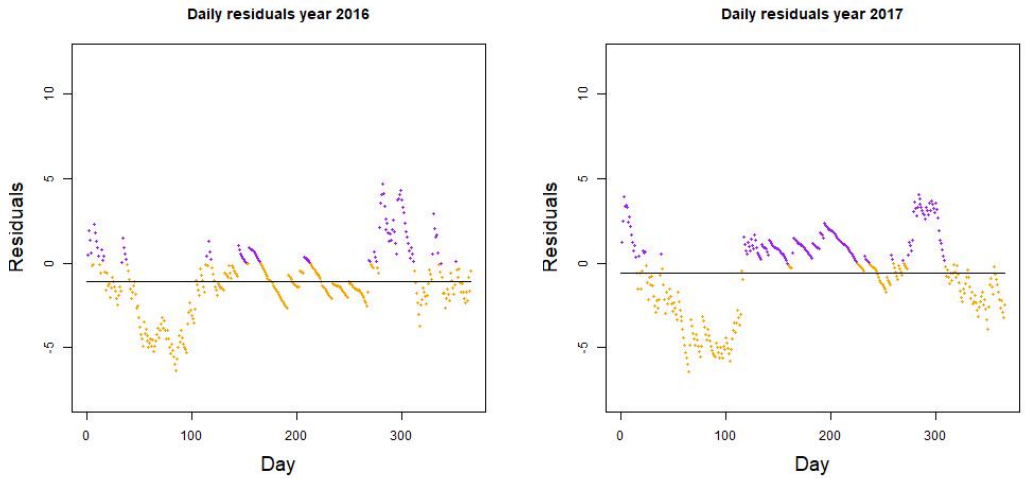


Figure 24: Residual time plot corresponding to Equation (33), where the black solid line indicates the mean of the residuals.

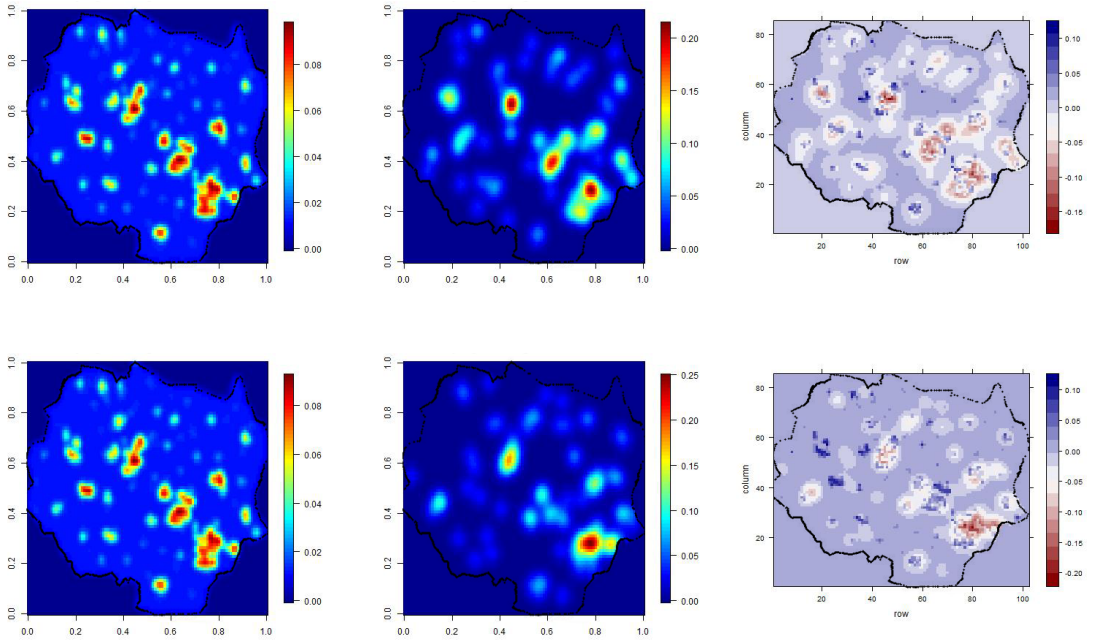


Figure 25: From left to right: The computed intensity plot according to the fitted Log Gaussian Cox process, the blurred image of the actual occurred events and the residual plot. The first and second row corresponds to the data of year 2016 and 2017 respectively.

boxes. The sum of the residuals in for the Log Gaussian Cox process are 2.52 and 5.82 for years 2016 and 2017 respectively. We see that the absolute value of the residuals has decreased for 2016 and increased for 2017. Because the Log Gaussian Cox process introduces a random field, the intensity field is different for every simulation. Therefore the highest intensities will every time be divided over Twente differently, so we can say that for this simulation, the intensities for 2016 are coincidentally better divided over Twente than for 2017.

8 Practical result: Prediction dashboard

The research we did in this thesis is the result of the thirst of using Business Intelligence in the fire department. As said in the introduction, this work is meant to take a first step in using data and see where the right way of using it can lead us. Because of its practical character the fire department desired to develop a concrete result for all its employees and its users: the inhabitants of Twente. Therefore the Log Gaussian Cox model we developed has grown into a dashboard, which makes it possible for everyone to see the results of our research by putting it on the website. The ambition was to create a concrete prediction product where the weather forecast and day of the year could be filled in after which the chimney forecast in time and space would be displayed.

Currently the website of the fire department shows a map of Twente where per 500 by 500 meter box the number of chimney fires actually occurred from 2004 until now. The dashboard is going to replace the current map and is a pioneering accessory for fire department websites, no other fire department region in the Netherlands uses a similar dashboard on their website. Visitors of the website are able to see the expected number of chimney fires for the coming week and also if it is likely that a fire will occur in their own borough.

Currently, the fire department uses intern already the visualization tool Microsoft Power BI, and therefore we chose to build the dashboard in this program as well. Because the employees of the fire department are already used working with this program, implementing seems then like a smaller step. Beside the amazing visualisations this program is capable of, also a connection with our programming language `R` is available which makes it easier to implement the already existing scripts. This section explains first the implementation of the tool, after that the design and the visualisations we used and concludes with a description on updating the tool. The last part is important to make the tool adaptive for many years.

8.1 Implementation

The basis of the dashboard is the intensity function defined in Equation (44). This function consists of two parts, the concrete inputs temperature, day of the year and the number of residents and the random field which is defined by its pair correlation function and the mean. Beside that, the dashboard will also show the confidence intervals corresponding to the expectation we show.

Unfortunately, the `R` connection of Microsoft Power BI cannot be used to calculate directly with the inputs and therefore the calculations need to be done beforehand. To overcome this problem, we give options to the user for the time related inputs and for these options the subintensity function is already calculated. The information about the number of residents per box in Twente is not considered as an input because it does not change over the year and therefore the data and the corresponding subintensity functions are already loaded in the program by using the `R` connection in the program. In the Netherlands the highest temperature ever measured is 38.6°C while the lowest temperature ever measured is -27.4°C, the possible temperature options are thus all numeric values between -28 and 39. For the October variable, the user can choose the date of today and the program knows the subintensity function value for October of the six upcoming days. So in the end, we consider two inputs; the temperature for the upcoming six days and the date of today. If these values are filled in, the subintensity values of the inputs and the number of residents per box are all looked up by a formula in Power BI and can be combined, which results in the final intensity function of our process, see Equation (29).

The random field with its pair correlation function and mean needs to be included by using the `R` connection in the program. In this case we also want to know the noise intensity of the random field for every box and every year. When generating a random field in `R`, the field can be transformed into a matrix of 6291 by 365, which gives the intensity per box per day of the year. Because of memory reasons in the program, the random field matrix is saved in two different

tables, one for the noise intensity per day and one for the noise intensity. Then the noise and the intensity values can be combined to calculate the number of chimney fires we expect in the upcoming week by Equation (44).

Also the confidence interval is included into the program in a similar way. Because in Equation (2) we see that only the expected number of chimney fires, the region and the time period is needed to calculate the confidence interval, we can calculate the confidence interval per possible expected number of chimney fires. The confidence parameter for the interval is chosen to be always higher than 80% but can convert into bigger values as well, because we chose to round the confidence interval to integers.

8.2 Design

As said in the previous subsection, only the temperature (six times) and the date of today are user inputs of this dashboard, while our information about the number of residents per 500 by 500 meter box does not change over the year. The day of the week can set to be fixed by the program itself, which means that we reduce the number of inputs from seven to six. The input page is currently looking as in Figure 26.

Figure 26: The first page, the interface of the Power Bi dashboard.

In the minimum and maximum boxes, the expected minimum and maximum temperature of the upcoming six days can be filled in. In the strip at the bottom of the page, the link to this thesis is given. The temperature inputs show us the expected number of chimney fires corresponding to these inputs, which is shown in Figure 27. Here the intensities are translated into a few visualisations and a summarising sentence. The visualisations correspond to the expected number of chimney fires per location, per day, in total and the confidence of the model. In the figure of Twente on the left, the intensity per 500 by 50 meter box is shown, while on the smaller figure on the right the intensity per neighbourhood is shown. The intensities per box are integrated per neighbourhood and per borough, the small table on the left of the figure shows the top 3 of boroughs where we expect the most chimney fires. The graph on the right displays the behaviour of the intensity per day, so that the days with the highest expectation can be filtered by the user.

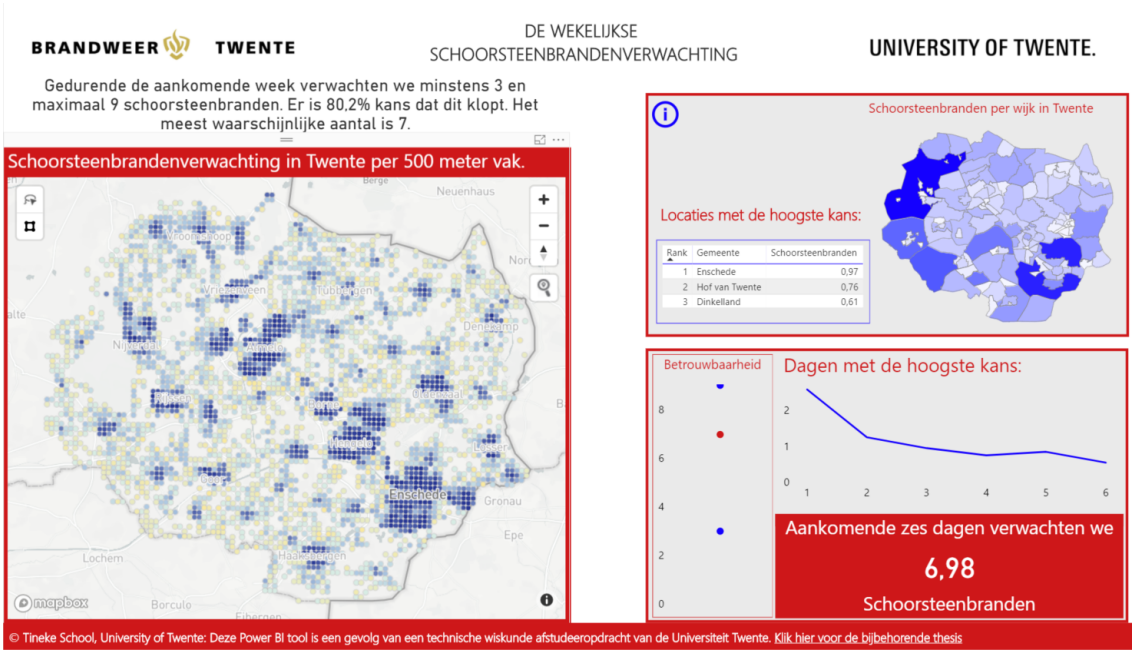


Figure 27: The expected number of chimney fires according to the model, displayed in several visualisations.

Right under the graph the total expected number of chimney fires of the upcoming six weeks is display and on the left the confidence bounds are shown. The red dot represents the predictions while the blue dots give the bounds. The sentence on top of the page summarised the information and adds the confidence parameter to the visualisation. In the middle also an information button is added, which opens a pop up with additional information about the visuals.

The official fire department colour codes are used in the dashboard and other layout choices are made in cooperation with the fire department.

8.3 Updating

Because we use the dates of 2018, the number of residents of 2018 and the fitted parameters on data until 2017, the tool should be updated for usage after 2018. In loading the previous explained tables, the parameters are listed out, so that they can easily be changed. To make this adjustment as easy as possible, a manual is written for the fire department, where they can follow a step by step report to newborn the tool in 2019 (and every upcoming year). The procedure mainly comes down to adding the new information about the chimney fires, the weather and the residents of Twente per box. The program fits then again the subintensity functions and the pair correlation and mean function. These new functions can easily be added to the Power BI tool which makes it ready to use in the next years.

9 Conclusions and Further Research

9.1 Conclusion and Discussion

The fire department owns a lot of information and has the thirst to find the possibilities of Business Intelligence to improve their organisation and create a safer society. In this work the first step of exploring these possibilities is taken by modelling a spatio-temporal point process for chimney fires in Twente. We focussed on a specific type of fire because an earlier study on fires in general did not have satisfying results and chimney fires are chosen because of their season dependence and it is the most occuring type of fire. This thesis was based on the work of [21] which helped a lot to understand the mathematical field of point processes; the procedures were namely well explained which made it easier to get started with the data. The study we did in this paper, resulted in fitting two different models, namely the inhomogeneous Poisson process and the Log Gaussian Cox process which was proposed by [21], where the Log Gaussian Cox process is an extension of the inhomogeneous Poisson process which can include spatially dependent noise.

In Section 3 we tested by a distance analysis if the data involves an obvious interaction structure, because when it does not involve interaction, an (in)homogeneous Poisson process would probably fit the data well. The analysis clearly indicated that chimney fire data did not fit a Poisson process well because we saw strong signs of a clustered pattern and thus we extended the inhomogeneous Poisson process to a Log Gaussian Cox process to also include the spatially dependent noise. Because of the factorisation of the Log Gaussian Cox process, the inhomogeneous intensity function will be fitted first and after that the random field to complete the Log Gaussian Cox process will be fitted as well.

To find the intensity function of the inhomogeneous Poisson process, a correlation analysis is performed. We tested several spatial and temporal covariates such as number of houses from a certain building year and weather conditions. We did not have the number of chimneys per area and in the correlation analysis we found that this number can be approximated best with the number of residents per area. Covariates as the number of free standing houses or older buildings did not have the same correlation which can be explained by the chosen 500 meter boxes. These covariates can have a higher influence when we for example test the neighbourhoods with the same building year or same kind of houses. In a 500 meter box this can be too mixed and there are just too many of these boxes to see this correlation well.

The correlation analysis we performed in Section 4.2 resulted then in the number of residents and the mean temperature of a day which had the highest impact on chimney fires. Both of these covariates were included in an intensity function and because we have one spatial and one temporal covariate, the inhomogeneous Poisson process has a spatio-temporal intensity function. In the predictions corresponding to this model, we saw that the weather tipping point also has an influence on the data. After a correlation analysis on two possible weather tipping point covariates, another covariate dependent on the month October was included in the model. The inhomogeneous Poisson process is completed with an intensity function dependent on three subintensity functions, dependent on the number of residents per 500 by 500 meter box, the daily mean temperature and the month October.

This final Poisson model gave good predictions but still missed some events and, as concluded before, the spatially dependent noise. These two things are combined in adding a random field to the process, which resulted in the Log Gaussian Cox model. In Section 6 this model was fitted to the data and new predictions were made. The intensity fields corresponding to both models were similar but the predictions per month came with the second model much closer to reality. Beside predictions per month, we also extended the predictions to confidence intervals per week for both models. We chose to include the weekly intervals because the fire department has the desire to predict chimney fires per week. Because the number of events predicted by only the random

field is low, the confidence intervals for both models are equal. In Figures 14 and 23 the weekly intervals contain mostly the number of actual occurred events and often, the predicted number of events per week is with the Log Gaussian Cox process a bit closer to reality. For weeks closer to or in the summer, the confidence intervals became less accurate which is due to the few number of events happening in summer. From this we can conclude that the Log Gaussian Cox process predicts the data very well and the predictions are certainly reliable in winter while in summer the predictions become more guesswork.

Comparing the models based on their residuals we saw clearly that the time residuals became symmetric, which indicates a better model. With the space residuals being similar, the conclusion from the two models is that the Log Gaussian Cox process describes the data the best. The covariates used, mean temperature, number of residents and indicating the month October, are easy ones in the sense that their value is known. The process is therefore tractable and easily understandable. Thus to answer the central question as defined in the Introduction: The Log Gaussian Cox process is a good model for predicting chimney fires and the procedure we used can be easily translated to model other safety themes.

9.2 Further Research

For further research on the resulting chimney fire model, we have a few suggestions based on the data we used, the procedure we used and possible extensions of this procedure. First of all we suggest to reinvestigate the usage of the 500 by 500 meter boxes on which we based our analysis. As we saw in the correlation analysis, the results were not those expected by the fire department. We namely found that the number of residents had a higher correlation than the number of free standing and town houses and the number of older houses. We know that chimneys are more present in these kind of houses so the number of residents being a better approximation for the number of chimneys in Twente was a strange outcome. The use of 500 by 500 meter boxes is pointed to as the central reason of this result because there are many of these and the structure is often mixed. By doing the analysis with whole neighbourhoods with a certain building style could have other and more intuitive results. Also because of the small size of 500 meter boxes, the contrast between the values of the covariates are small, so a correlation is more difficult to find. When a larger area is considered, the differences of the values are higher and therefore changes are easier to detect.

Furthermore in Section 3 it is assumed that the data can be compared with an isotropic point process because of the convenience in the remainder of the research. There is yet no reason to believe that the data considered is anisotropic, but this assumption can be investigated further.

Third, we investigated the influence of the weather tipping point and included the subintensity function for the month October. Because the subintensity function for April did not improve the predictions, we did not include the boolean variable, but to include the month April also other subintensity functions can be investigated. According to the confidence plots in Figure 13, some improvement is possible in this month.

Also in this work, the chosen procedure for computing confidence intervals is a self developed one. We chose this procedure because there is no theoretical foundation on confidence intervals for the Log Gaussian Cox process yet. In the process we do not consider a likelihood so we developed a resampling method with a relatively small number of simulations. Beside that, the resulting confidence interval is added to the Poisson confidence interval which is also doubtable. To improve the procedure, first more mathematical research on the confidence interval of the Log Gaussian Cox process is necessary.

Beside that, it could be interested to extend the current model with more covariates. As we saw in the correlation analysis, the covariates concerning residents and buildings had a correlation

coefficient which lie close to each other. We did not include covariates of the same type because they strongly relate to each other but it could be interesting to build a model with several covariates of the same type and to include the interactions between these covariates. The resulting model will be much more complex but probably the amount of noise will decrease because simply more covariates are considered.

As a last suggestion we name the procedure of [8] because they inherit dynamical hierarchic extensions of the Log Gaussian Cox processes, for example the re-estimation of the parameters every day and the use of Monte Carlo simulations. For our research, to include these extension was not possible in the time period available and would probably go beyond the purpose for the fire department, but in the future it could be interesting to investigate these extensions.

As a final remark about further research, we mention here the article we are currently writing for the Magazine of Safety (Tijdschrift voor Veiligheid), because our research is relatively new in the fire department field. The writing is still ongoing and hopefully the article will be submitted in September.

References

[1] A. Baddeley, Rubak, E., and Turner, R. *Spatial point patterns: Methodology and applications with R.* Chapman and Hall/CRC, 2015.

[2] Baddeley, A., Møller, J., and Waagepetersen, R. Non- and semiparametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica 54* (2000), 329–350.

[3] Baddeley, A., Turner, R., Møller, J., and Hazelton, M. Residual analysis for spatial point processes. *Royal Statistical Society: Series B (Statistical Methodology) 67* (2005), 617–768.

[4] Boxma, O., and Yechiali, U. *Poisson processes, ordinary and compound.* Tel Aviv University, 2006.

[5] Brix, A., and Diggle, P. Spatiotemporal prediction for Log-Gaussian Cox processes. *Royal Statistics Society: Series B (Statistical Methodology) 63*, 4 (2001), 823–841.

[6] Chung, M. K. *Introduction to random fields.* Department of Statistics, Biostatistics and Medical Informatics of the University of Wisconsin-Madison, 2007.

[7] Diggle, P. *Statistical analysis of spatial point processes. Mathematics in biology*, vol. 2. Academic- Press, London-New York, 1983.

[8] Diggle, P., Rowlingson, B., and Su, T. Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics 16* (2005), 423–434.

[9] Diggle, P., Moraga, P., Rowlingson, B., and Taylor, B. Spatial and spatio-temporal Log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statistical Sience 28*, 4 (2013), 542–563.

[10] Gabriel, E., and Diggle, P. Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica 63* (2009), 43–51.

[11] Groen, G. Wind chill equivalente temperatuur (wcet), knmi-implementatie jag/ti-methode voor de gevoelstemperatuur in de winter. *KNMI* (2009).

[12] Holmes, M., Wang, Y., and Ziedins, I. *The application of data mining and statistical techniques to identify patterns and changes in fire events.* University of Auckland, 2009.

[13] Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. *Statistical analysis and modelling of spatial point patterns.* John Wiley and Sons Ltd,, 2008.

[14] Lieshout, M. v. A J-function for inhomogeneous point patterns. *Statistica Neerlandica 65* (2010), 183–201.

[15] Lieshout, M. v. *Spatial point patterns, Chapter 3: Point processes.* Lecture notes Mastermath, 2018.

[16] Møller, J., and Waagepetersen, R. P. *Statistical inference and simulation for spatial point processes*, vol. 3. Chapman and Hall/CRC, 2003.

[17] Møller, J., Syversveen, A., and Waagepetersen, R. Log gaussian cox processes. *Scandinavian Journal of Statistics 25* (1998), 451–482.

[18] Proietti, T. *Leave k out diagnostics in state-space models.* University of Udine, 2000.

[19] Rasmussen, C. E., and Williams, C. *Gaussian processes for machine learning.* the MIT Press & Massachusetts Institute of Technology, 2005.

[20] Turner, R. Point patterns of forest fire locations. *Environmental and Ecological Statistics 16* (2009), 197–223.

[21] Wendels, M. *A spatio-temporal point process model for firemen demand in Twente.* B.Sc thesis, University of Twente, 2017.

10 Appendix

10.1 Covariates used in inhomogeneous Poisson process

$C_{\sigma,1}$	The total number of buildings [1]
$C_{\sigma,2}$	The number of buildings with an industrial or agricultural function [1]
$C_{\sigma,3}$	The number of buildings with an hotel function [1]
$C_{\sigma,4}$	The number of buildings with a residential function [4]
$C_{\sigma,5}$	The number of buildings build before 1920
$C_{\sigma,6}$	The number of buildings build between 1920 and 1945
$C_{\sigma,7}$	The number of buildings build between 1945 and 1970
$C_{\sigma,8}$	The number of buildings build between 1970 and 1980
$C_{\sigma,9}$	The number of buildings build between 1980 and 1990
$C_{\sigma,10}$	The number of buildings build after 1990
$C_{\sigma,11}$	The number of stand alone and town houses
$C_{\sigma,12}$	The number of other houses than $C_{\sigma,11}$
$C_{\sigma,13}$	The number of residents [4]
$C_{\sigma,14}$	The number of residents with an age in the range of 0 till 14 [4]
$C_{\sigma,15}$	The number of residents with an age in the range of 15 till 24 [4]
$C_{\sigma,16}$	The number of residents with an age in the range of 25 till 44 [4]
$C_{\sigma,17}$	The number of residents with an age in the range of 45 till 64 [4]
$C_{\sigma,18}$	The number of residents with an age of 65 or higher [4]
$C_{\sigma,19}$	The number of male residents [4]
$C_{\sigma,20}$	The number of female residents [4]
$C_{\sigma,21}$	The number of residents who live in a stand alone or town house
$C_{\sigma,22}$	The density of addresses in the neighbourhood [4]
$C_{\sigma,23}$	The urbanity of the neighbourhood [4]
$C_{\sigma,24}$	Boolean variable indicating the presence of a town [5]
$C_{\tau,1}$	Daily mean wind speed (in 0.1 meter per second) [8]
$C_{\tau,2}$	Daily mean temperature (in 0.1 degrees Celsius) [8]
$C_{\tau,3}$	Daily mean wind chill [8]
$C_{\tau,4}$	Sunshine duration calculated from global radiation (in 0.1 hour) [8]
$C_{\tau,5}$	Boolean variable indicating if it is spring (1 March till 31 May)
$C_{\tau,6}$	Boolean variable indicating if it is summer (1 June till 31 August)
$C_{\tau,7}$	Boolean variable indicating if it is autumn (1 September till 31 November)
$C_{\tau,8}$	Boolean variable indicating if it is winter (1 December till 28 February)
$C_{\tau,9}$	Boolean variable indicating if fog appeared this day
$C_{\tau,10}$	The day d_m of period T_m, $1 \leq d_m \leq 4380$

Table 12: The spatial covariates $C_{\sigma,k}, 1 \leq k \leq 28$ involved in the covariate analysis. The number in the square brackets indicates the source covariate data set of that covariate.

10.2 Pair correlation function plots time 1-14

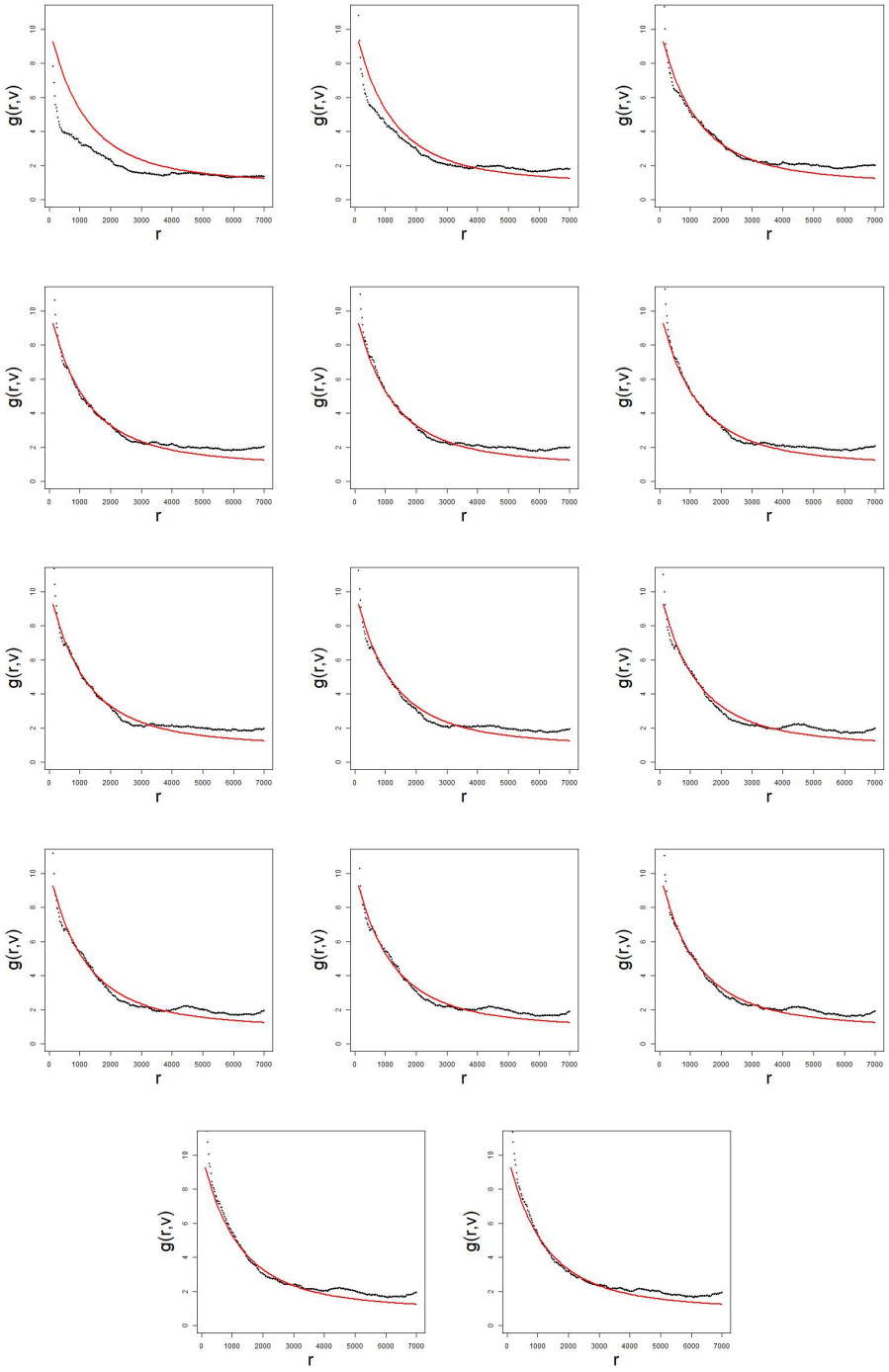


Figure 28: The pair correlation function as in Equation (38) with the fitted parameter values (solid line) against the estimate of the pair correlation correlation function as in Equation (36) (dots) for time difference in the range from 1 day until 14 days.

10.3 Pair correlation function plots time 15-28

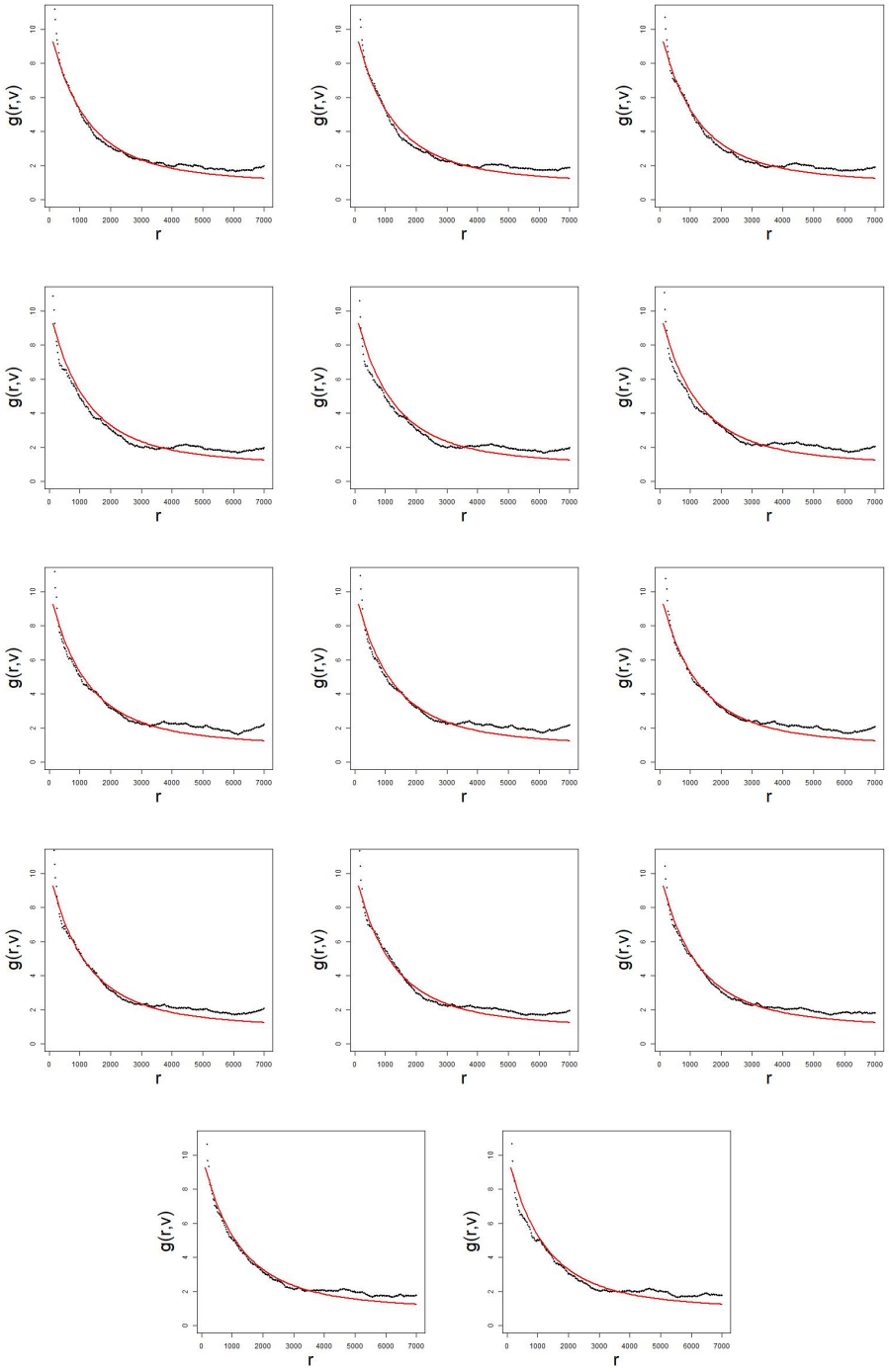


Figure 29: The pair correlation function as in Equation (38) with the fitted parameter values (solid line) against the estimate of the pair correlation correlation function as in Equation (36) (dots) for time difference in the range from 15 until 28 days.